Modélisation des expositions médicamenteuses et de leurs effets variant dans le temps : comparaison de méthodes d'analyse statistique

Liliane Manitchoko

université PARIS-SACLAY

Modélisation des expositions médicamenteuses et de leurs effets variant dans le temps : comparaison de méthodes d'analyse statistique

Modeling drug exposures and their time-varying effects : comparison of statistical analysis methods

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 570 : santé publique (EDSP)
Spécialité de doctorat : santé publique - biostatistiques
Graduate School : Santé publique
Référent : Université de Versailles-Saint-Quentin-en-Yvelines

Thèse préparée dans l'unité de recherche : CESP (Université Paris-Saclay, UVSQ, Inserm),
sous la direction d'Anne THIEBAUT, Chargée de recherche, la co-direction de Jacques BENICHOU, Professeur des universités

Thèse soutenue à Paris-Saclay, le 01 juillet 2022, par

Liliane MANITCHOKO

Composition du jury

Delphine MAUCORT-BOULCH PU-PH, Université Claude Bernard Lyon 1	Présidente et Rapporteure
Karen LEFFONDRE PU, Université de Bordeaux	Rapporteure et Examinatrice
Mounia HOCINE MCF, Chargée d'expertise, Santé Publique France	Examinatrice
Jean-Christophe THALABARD PU-PH, Université Paris Cité	Examinateur
Anne THIEBAUT CR, Université Paris-Saclay	Directrice de thèse

THESE DE DOCTORAT

NNT : 2022UPASR008

université PARIS-SACLAY
ÉCOLE DOCTORALE
Santé Publique
(EDSP)

Titre : Modélisation des expositions médicamenteuses et de leurs effets variant dans le temps : comparaison de méthodes d'analyse statistique

Mots clés : Exposition variable dans le temps, Cohorte, Cas-témoins niché, Simulation, Temps d'évènement ex-æquos, Cancer du sein

Résumé : L'évaluation des effets des expositions médicamenteuses sur la survenue des évènements de santé constitue un réel défi. En effet, l'exposition médicamenteuse des individus étant susceptible de varier dans le temps, elle entraîne des risques associés qui dépendent à la fois de la dose, de la durée et du moment des traitements. Ainsi, différents schémas d'étude et méthodes d'analyse statistique sont utilisés pour estimer le risque associé à une exposition médicamenteuse. Cependant, peu de travaux ont évalué et comparé systématiquement les performances respectives des schémas d'étude de cohorte et cas-témoins niché (CTN) pour estimer l'effet d'une exposition médicamenteuse fixe ou variable au cours du temps. Dans ce travail de thèse, nous avons utilisé des simulations pour examiner et comparer les performances des estimations issues des analyses CTN par rapport à celles issues des analyses de la cohorte entière pour évaluer les associations entre une exposition fixe ou variable dans le temps et le risque de survenue d'un évènement de santé. Pour cette comparaison, nous avons également utilisé les données de la cohorte E3N permettant d'évaluer l'association entre l'utilisation de traitements hormonaux de la ménopause et le risque de cancer du sein. Les résultats de l'étude de simulation que nous avons réalisée ont montré que les estimations obtenues de l'analyse de la cohorte entière étaient non biaisées dans tous les scénarios considérés. Cependant, les estimations issues des analyses CTN étaient substantiellement biaisées, en particulier lorsqu'un seul témoin était apparié à chaque cas. Ce biais des estimations cas-témoins nichées augmentait avec la proportion d'évènements. Une amélioration sensible des estimations CTN a été observée après le recours à une méthode de réduction de biais, semblant suggérer que les biais observés pouvaient résulter de données éparses. Cependant, cette explication ne nous a pas satisfaits car les biais persistaient quel que soit le nombre d'évènements. Nous avons donc poursuivi nos investigations en nous intéressant à la gestion des temps d'évènement ex-æquos en évaluant différentes méthodes pour les prendre en compte dans l'analyse CTN. Notre étude de simulation et l'application aux données de la cohorte E3N a montré que les analyses CTN avec les approximations de Breslow ou d'Efron pouvaient entraîner un biais important lorsqu'il y avait un grand nombre de temps d'évènement ex-æquos dans les données. Cependant, une fois les ex-æquos correctement pris en compte en utilisant la méthode exacte ou une approche permettant d'avoir un seul cas dans chaque strate, les estimations CTN étaient presque exemptes de biais et proches de celles de l'analyse de la cohorte entière. Nous recommandons fortement d'être particulièrement vigilant aux temps d'évènement ex-æquos dans les analyses CTN, en particulier à la façon dont ils sont gérés aussi bien lors de la formation des strates appariées que lors de l'analyse par régression logistique conditionnelle.

Title : Modeling drug exposures and their time-varying effects : comparison of statistical analysis methods

Keywords : Time-varying exposure, Cohort design, Nested case-control design, Simulation study, Tied events, Breast cancer

Abstract : Assessing the effects of drug exposure on the occurrence of health events is a real challenge. As individuals' drug exposures are likely to vary over time, they carry associated risks that depend on the dose, duration, and timing of treatment. Thus, different study designs and statistical analysis methods are used to estimate the risk associated with drug exposure. However, few studies have systematically evaluated and compared the respective performances of cohort and nested case-control (NCC) designs to estimate the effect of fixed or time-varying drug exposure. In this thesis work, we used simulations to examine and compare the performance of estimates from NCC versus whole cohort analyses to assess associations between a fixed or time-varying exposure and the risk of a health event. For this comparison, we also used data from the E3N cohort to assess the association between the use of menopausal hormone therapy and breast cancer risk. The results of the simulation study we conducted showed that the estimates obtained from the analysis of the whole cohort were unbiased in all scenarios considered. However, the estimates from the NCC analyses were substantially biased, especially when only one control was matched to each case. This bias in the nested case-control estimates increased with the proportion of events. A significant improvement in the NCC estimates was observed after the use of a bias reduction method, suggesting that the observed biases could be the result of sparse data. However, we were not satisfied with this explanation as the biases persisted regardless of the number of events. We, therefore, pursued our investigations by looking at the handling of tied event times by evaluating different methods to take them into account in the NCC analysis. Our simulation study and application to the E3N cohort data showed that NCC analyses with Breslow or Efron approximations could lead to a significant bias when there were a large number of tied event times in the data. However, once the tied event times were properly accounted for using the exact method or an approach that allowed for a single case in each stratum, the NCC estimates were almost unbiased and close to those of the whole cohort analysis. We strongly recommend that particular attention be paid to tied events in NCC analyses, in particular how they are handled both when forming matched strata and in the analysis by conditional logistic regression.

Remerciements

Je tiens tout d'abord à remercier grandement ma directrice et mon directeur de thèse, Anne Thiébaut et Jacques Bénichou qui m'ont permis de réaliser cette thèse. Merci à vous de m'avoir accordé votre confiance, de m'avoir guidée et encouragée tout au long de ces années. Je vous remercie pour votre rigueur, votre disponibilité sans faille et pour votre bienveillance.

Je remercie vivement Pascale Tubert-Bitter, directrice de l'équipe pour son accueil depuis mon stage de master 2, ses remarques sur mon travail, ses conseils et d'avoir mis à ma disposition le financement sans quoi cette thèse n'aurait été possible.

Je remercie également les membres de mon jury de thèse pour l'intérêt et le temps consacré à la lecture attentive de mon travail. Merci à Karen Leffondré et Delphine Maucort-Boulch d'avoir accepté d'être les rapporteures de cette thèse ainsi qu'à Jean-Christophe Thalabard et Mounia Hocine pour avoir accepté d'en être les examinateurs.

Je remercie l'Agence Nationale de Sécurité du Médicament et des produits de santé (ANSM) et l'Institut pour la Recherche en Santé Publique, GIS-IReSP pour les subventions des différents projets qui ont permis le financement de toutes ces années de thèse.

Je remercie les investigateurs de la cohorte E3N, en particulier Agnès Fournier et Gianluca Severi qui m'ont permis d'accéder à ces données.

Merci à mes collègues (anciens et actuels) de l'équipe : Sylvie, Ismaïl, Émeline, Étienne, Matthieu, Mélanie, Stéphanie, Hong, Lucas pour vos retours après mes présentations qui me permettaient de les améliorer et les échanges pendant le déjeuner et diverses occasions.

Je remercie ma famille pour leurs encouragements et leur soutien multiforme malgré la distance.

Merci infiniment à ma belle-mère Clarisse et ma belle-soeur Audrey pour votre présence, votre affection et votre soutien indéfectible.

Merci aux hommes de ma vie :
A Rodrigue, mon partenaire, mon ami, mon confident, merci d'avoir toujours été présent depuis la planification de ce projet il y a plusieurs années, de m'avoir soutenue, encouragée à sortir de ma "zone de confort" et à donner le meilleur de moi. Rien n'a été facile (en particulier pendant ces deux

dernières années), mais nous avons pu, ensemble, surmonter les difficultés et prendre soin au mieux des personnes qui comptent pour nous, et ce, malgré nos diverses occupations.
A Ethan et Enzo, mes super "glues", mes p'tits tout, merci de m'avoir permis de m'évader (de toute la pression et du stress que je pouvais parfois ressentir), de relativiser et de toujours voir "le verre à moitié plein", à travers votre innocence, votre sourire, votre joie de vivre et tous les moments passés ensemble. Votre présence m'a donné la force de me surpasser pour achever ce projet et j'espère qu'à travers ce travail accompli vous verrez un exemple de persévérance malgré les difficultés rencontrées.

Valorisations scientifiques

Articles

Liliane Manitchoko, Michal Abrahamowicz, Pascale Tubert-Bitter, Jacques Bénichou, Anne C.M. Thiébaut. Comparison of cohort and nested case-control designs for estimating the effect of time-varying drug exposure on the risk of adverse event in the presence of ties. Soumis le 30 novembre 2021 à *Biometrical Journal* (Annexe B).

Communications orales

Liliane Manitchoko, Jacques Bénichou, Anne Thiébaut. Comparaison de méthodes d'estimation de l'effet d'une exposition médicamenteuse sur le risque d'évènement indésirable. Journées du GDR « statistiques et santé », Paris, 10-11 octobre 2019.

Liliane Manitchoko, Jacques Bénichou, Anne Thiébaut. Comparison of statistical methods for estimating time-varying treatment effect on adverse event risk. 42nd annual conference of the International Society for Clinical Biostatistics, Lyon, 18-22 juillet 2021.

Liliane Manitchoko, Michal Abrahamowicz, Pascale Tubert-Bitter, Jacques Bénichou, Anne Thiébaut. Comparison of cohort and nested case-control designs for estimating the effect of time-varying drug exposures. 31st International Biometric Conference, Riga, 10-15 juillet 2022.

Communications affichées

Liliane Manitchoko, Jacques Bénichou, Anne Thiébaut. Comparison of cohort and nested case-control analysis for estimating the effect of time-varying treatment on the risk of adverse event. Channel Network Conference, Paris, 7-9 avril 2021.

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

CNIL	Commission Nationale de l'Informatique et des Libertés
CTN	Cas-témoins niché
E3N	Étude Épidémiologique auprès de femmes de la Mutuelle Générale de l'Éducation Nationale
HR	*Hazard Ratio*
IC	Intervalle de confiance
MGEN	Mutuelle Générale de l'Education Nationale
RMSE	Erreur quadratique moyenne (pour *Root Mean Squared Error*)
SEE	Moyenne des écarts-types estimés (pour *average standard error estimator*)
THM	Traitements hormonaux de la ménopause

1 - Introduction

Cette thèse se situe dans le champ de la pharmaco-épidémiologie définie comme étant la science qui étudie l'utilisation et évalue les effets bénéfiques ou délétères des produits de santé chez un grand nombre d'individus (Bégaud, 1998 ; Strom et al., 2013). Plus précisément, nous nous intéressons ici à la modélisation de l'exposition médicamenteuse et de ses effets bénéfiques ou délétères pour la santé. Cette modélisation soulève des difficultés méthodologiques spécifiques du fait de la nature dynamique ou dimension temporelle inhérente à cette exposition. Dans une première partie, nous décrivons les particularités de l'exposition médicamenteuse et les modèles statistiques servant à la décrire. Une deuxième partie concerne l'analyse de survie qui est le cadre d'analyse statistique pour la modélisation de l'effet d'une exposition sur la survenue d'un évènement de santé dans le temps. Dans une troisième partie, nous présentons les algorithmes de génération des données de survie. Les quatrième et cinquième parties concernent les schémas des études observationnelles utilisés en pharmaco-épidémiologie et l'état de l'art sur les travaux comparant ces différents schémas d'étude, respectivement.

1.1 Modélisation de l'exposition médicamenteuse

L'exposition médicamenteuse pouvant varier (changement du statut d'exposition ou de l'exposition elle-même en termes de molécule, durée, dosage et fréquence) considérablement d'un sujet à l'autre et au sein d'un même sujet au cours du temps, l'évaluation de ses effets sur l'occurrence des évènements de santé constitue un réel défi. En effet, une exposition médicamenteuse variant au cours du temps doit être représentée ou quantifiée de manière appropriée pour éviter de sous-estimer/sur-estimer son effet (Abrahamowicz et al., 2006 ; Pazzagli et al., 2018 ; Vacek, 1997).

Il existe plusieurs modèles de mesure de l'exposition médicamenteuse pouvant donner des estimations différentes de son effet pour répondre à une même question de l'étude. Ces approches de mesure de l'exposition peuvent être nombreuses et leur choix dans les études sur la sécurité et l'efficacité des médicaments est largement guidé par la question de l'étude et peut être affiné par la connaissance de la fréquence et l'intermittence de la consommation des médicaments, de la durée, du moment et de l'intensité (dose) de l'exposition, des périodes d'induction et de latence. Cependant, la disponibilité des éléments dans un ensemble de données peut dicter le choix d'un modèle de mesure de l'exposition. Nous présentons dans la suite différents modèles permettant de quantifier une exposition médicamenteuse.

1.1.1 Modèles simples

Soit $X(u)$ le vecteur des doses aux temps $u = t_0, ..., t$ d'une exposition médicamenteuse variable dans le temps sur une période d'observation $[t_0, t]$, t_0 et t étant les dates de début et de fin. Sur cette période, les valeurs de cette exposition peuvent :

- être fixes si elles ne sont pas actualisées à compter du début de la période d'observation (ou d'une certaine date définie), le statut d'exposition restant "non exposé" ou "exposé" : nous parlerons d'*exposition fixe dans le temps* ;
- varier une seule fois, le statut d'exposition passant par exemple de "non exposé" à "exposé" et représenté de manière dichotomique : $X(u) = 0$ pour $t_0 \leq u < \tau$ et $X(u) = 1$ pour $u \geq \tau$, τ étant la date de début d'exposition : nous parlerons d'*exposition variable dans le temps à changement unique* ;
- varier plusieurs fois, le statut d'exposition passant par exemple de "non exposé" à "exposé", puis de "exposé" à "non exposé", etc. : nous parlerons d'*exposition variable dans le temps à changements multiples* (Desquilbet & Meyer, 2005).

Plusieurs approches simples existent pour représenter ces deux derniers types d'exposition médicamenteuse variable dans le temps dans les modèles d'analyse de leurs effets sur le risque de survenue d'un évènement. Dans cette partie, nous présentons quelques-unes de ces approches utilisées dans les études pharmaco-épidémiologiques.

1.1.1.1 Exposition actuelle

Soit $Z(t)$ une représentation de l'exposition variable dans le temps à la date t d'évaluation de son effet sur le risque de survenue de l'évènement d'intérêt. Une façon simple de considérer l'exposition est de prendre sa valeur à la date la plus proche précédant l'évènement d'intérêt : il s'agit de l'utilisation ou de la dose actuelle.

L'utilisation actuelle est définie par le statut d'exposition à un médicament à la date t :

$$Z(t) = I[X(t) > s] = \begin{cases} 1 & \text{si la dose est supérieure à un seuil } s \text{ fixé (sujet exposé)} \\ 0 & \text{sinon (sujet non exposé)} \end{cases}$$

On peut de même définir la dose actuelle correspondant à l'intensité de l'exposition à la date t : $Z(t) = X(t)$. Ces deux définitions correspondent donc à des expositions variables dans le temps à changements multiples.

1.1.1.2 Exposition passée

L'utilisation passée peut être définie par l'occurrence de l'exposition médicamenteuse au moins une fois dans la période de temps précédant t :

$$Z(t) = I[X(u) > s, u \leq t] = \begin{cases} 1 & \text{si la dose est supérieure au moins une fois à s (sujet déjà exposé)} \\ 0 & \text{sinon (sujet jamais exposé)} \end{cases}$$

Cette définition d'une exposition variable dans le temps à changement unique est couramment utilisée en pharmaco-épidémiologie (*"ever/never exposed"* en anglais).

1.1.1.3 Exposition récente

L'utilisation récente peut être définie par l'occurrence de l'exposition médicamenteuse au moins une fois sur une période de temps relativement courte (précédant t), par exemple sur les k derniers jours ($k < t - t_0$) :

$$Z(t) = I[X(u) > s, (t-k) < u \leq t] = \begin{cases} 1 & \text{si la dose est supérieure au moins une fois à s (sujet exposé)} \\ 0 & \text{sinon (sujet non exposé)} \end{cases}$$

Si les doses de l'exposition sont enregistrées dans des intervalles de temps discrets (par exemple, les doses quotidiennes), alors la dose moyenne récente peut être représentée par la moyenne des doses passées sur les k derniers jours :

$$Z(t) = \frac{1}{k} \sum_{u=t-k+1}^{t} X(u)$$

1.1.1.4 Exposition cumulée

La durée d'utilisation en jours est représentée par la somme des indicatrices d'utilisation journalières :

$$Z(t) = \sum_{u=t_0}^{t} I[X(u) > s]$$

La dose cumulée est la somme de toutes les doses journalières passées :

$$Z(t) = \sum_{u=t_0}^{t} X(u)$$

Cette mesure de l'exposition repose sur l'hypothèse selon laquelle l'effet de chaque unité de dose est constant au cours du temps. Pourtant, l'exposition peut avoir des effets différents à différents moments.

Ces modèles d'exposition simples ne prennent ainsi en compte qu'un seul élément, en particulier soit la durée, soit l'intensité de l'exposition. Pourtant, le risque observé au temps t associé à une exposition médicamenteuse variable au cours du temps peut dépendre à la fois de l'intensité, de la durée et du délai écoulé depuis l'instant u de mesure des expositions passées (Edwards & Aronson, 2000). La prise en compte de tous ces éléments induit une modélisation complexe de l'exposition médicamenteuse.

1.1.2 Modèles complexes

L'importance relative des intensités des expositions passées sur le risque actuel peut être quantifiée par une fonction $w(t-u)$ du temps écoulé entre u et t (appelée fonction de pondération) qui détermine comment les poids varient en fonction du temps écoulé depuis l'exposition. La forme et les valeurs de cette fonction varient selon le mécanisme biologique qui lie l'exposition médicamenteuse passée au risque actuel au temps t (Abrahamowicz et al., 2006 ; Sylvestre & Abrahamowicz, 2009).

Par exemple, si les expositions récentes ont un impact élevé, alors la fonction de pondération sera une fonction décroissante du temps écoulé depuis l'exposition, avec $w(\Delta(t)) = 0$ pour $\Delta(t) > L$, indiquant que l'intensité d'une exposition médicamenteuse passée il y a plus de L jours n'a plus d'impact et peut être ignorée pendant l'évaluation du risque actuel au temps t (Abrahamowicz et al., 2012).
En effet, pour $L =$ longueur maximale de la fenêtre temporelle d'exposition pertinente du point de vue étiologique, les doses passées x_u au temps $u < t - L$ sont à priori considérées comme non pertinentes pour le risque au temps t, donc on leur assigne un poids nul.

Cette dimension temporelle qui s'ajoute à la relation habituelle "exposition-réponse" induit une modélisation potentiellement complexe de l'impact d'une exposition médicamenteuse dépendante du temps sur le risque d'évènement indésirable. Il s'agit alors de modéliser les relations "exposition-temps-réponse".

1.1.2.1 Exposition cumulée pondérée (*weighted cumulative exposure*, WCE)

Pour modéliser les relations "exposition-temps-réponse", Sylvestre et Abrahamowicz (2009) ont proposé le modèle WCE (*weighted cumulative exposure*) qui représente l'exposition médicamenteuse dépendante du temps comme une somme cumulée pondérée de toutes les expositions passées, les pondérations reflétant l'importance relative des expositions à différents moments (Sylvestre & Abrahamowicz, 2009). La forme de l'exposition est donnée par :

$$WCE(t) = \sum_{u}^{t} w(t-u)X(u) \tag{1.1}$$

où $u < t$ indexe les temps d'exposition précédant t et $X(u)$ représente l'intensité ou la dose de l'exposition individuelle variant dans le temps.

La forme de la fonction w étant inconnue, Sylvestre et Abrahamowicz (2009) proposent de l'estimer à partir des données en utilisant des méthodes flexibles non ou quasi-paramétriques telles que celle basée sur les régressions splines déjà utilisée antérieurement dans une étude analysant des relations "exposition-temps-réponse" (Hauptmann et al., 2000).

La mise en oeuvre de cette modélisation est facilitée par le package *WCE* du logiciel R dans lequel sont implémentées les différentes étapes (Sylvestre et al., 2018).

Une étude comparant différents modèles (exposition/dose actuelle, utilisation récente, exposition cumulée, etc.) pour représenter/quantifier de manière adéquate une exposition médicamenteuse variant au cours du temps à montré que le modèle WCE est meilleur et peut être plus "étiologiquement correct" pour l'association d'intérêt (Abrahamowicz et al., 2012 ; Pazzagli et al., 2018). Dans ce modèle WCE, la pondération des expositions suppose une relation exposition-réponse linéaire. Cependant, il peut arriver que cette relation soit non linéaire. Pour prendre en compte cette non linéarité, Gasparrini a modélisé, à l'aide des *Distributed Lag Non-linear Models* (DLNM) (Gasparrini et al., 2010), l'effet cumulé d'une histoire d'exposition comme une somme pondérée des effets des expositions passées (Gasparrini, 2014).

1.1.2.2 *Distributed Lag Non-linear Models* (DLNM)

Cette approche développée par Gasparrini (2014) généralise la modélisation des associations "exposition-temps-réponse" à l'aide des DLNM (Gasparrini et al., 2010). Elle consiste à évaluer le risque observé à un temps t actuel par une fonction $S(x_{t-l_0}, ..., x_{t-L})$ qui dépend de l'intensité x et du moment u des expositions passées, exprimés à travers :

— une fonction exposition-réponse $f(x)$ pour l'intensité de l'exposition x ;

— une fonction temps-réponse $w(l)$ pour le temps écoulé depuis l'exposition, $l = t - u$ (décalage).

Ces fonctions $f(x)$ et $w(l)$ décrivent respectivement la linéarité/non-linéarité de l'effet de l'exposition et la non-linéarité de l'effet du temps écoulé sur le risque d'évènement indésirable.

Sous les hypothèses d'effets identiques et d'indépendance entre $f(x)$ et $w(l)$ (la forme de la relation exposition-réponse est la même à chaque temps l et celle de la relation temps-réponse est la même pour chaque valeur de x), les fonctions $f(x)$ et $w(l)$ permettent d'avoir :

$$S(x_{t-l_0}, ..., x_{t-L}) = \int_{l_0}^{L} f(x_{t-l}).w(l)dl \approx \sum_{l=l_0}^{L} f(x_{t-l}).w(l) \tag{1.2}$$

avec des pas constants.

Si la fonction $f(x)$ est linéaire, alors on aura à une constante près :

$$S(x_{t-l_0}, ..., x_{t-L}) = \int_{l_0}^{L} x_{t-l}.w(l)dl \approx \sum_{l=l_0}^{L} x_{t-l}.w(l) \tag{1.3}$$

qui est équivalent au modèle WCE (Sylvestre & Abrahamowicz, 2009).

Les formes des fonctions $f(x)$ et $w(l)$ étant inconnues, l'estimation de la fonction $f(x).w(l)$ n'est pas immédiate car il est difficile de la représenter comme une combinaison linéaire de fonctions. Gasparrini propose donc de générer une fonction bi-dimensionnelle exposition-temps-réponse $f.w(x, l)$ telle que :

$$S(x_{t-l_0}, ..., x_{t-L}) = \int_{l_0}^{L} f.w(x_{t-l}, l)dl \approx \sum_{l=l_0}^{L} f.w(x_{t-l}, l) \tag{1.4}$$

Ainsi, pour une histoire d'exposition donnée au temps t pour $l = l_0$, ..., L :

$$q_{x_t} = [x_{t-l_0}, ..., x_{t-l}, ., x_{t-L}]^T$$

et, en supposant une relation exposition-réponse linéaire, on a :

$$S(q_{x_t}; \eta) = q_{x_t}^T C\eta = W_{x_t}^T \eta$$

où C est une matrice obtenue après application d'une transformation de base à la fonction $w(l)$, η est le vecteur de coefficients associés et $W_{x_t} = q_{x_t}^T C$.

Sous l'hypothèse d'une relation exposition-réponse non linéaire, la matrice R_{x_t} est

obtenue après application d'une seconde transformation de base à la fonction $f(x)$. Puis, un produit tensoriel

$$A_{x_t} = (1^T_{v_l} \bigotimes R_{x_t}) \bigodot (C \bigotimes 1^T_{v_x}) \tag{1.5}$$

est défini et permet de former la base croisée :

$$S(q_{x_t}; \eta) = (1^T_{(L-l_0+1)} A_{x_t})\eta = W^T_{x_t}\eta \tag{1.6}$$

avec

— W_{x_t} : matrice des éléments de la base croisée,

— v_l et v_x : nombre de colonnes des matrices C et R_{x_t} respectivement,

— $\bigotimes$ et $\bigodot$: produits matriciels respectifs de Kronecker et Hadamard.

Le problème d'estimation de la fonction $f.w(x, l)$ se réduit donc au choix des bases de fonctions (définies pour les vecteurs $q_{x_t} = [x_{t-l_0}, ..., x_{t-l}, ., x_{t-L}]^T$ et $l = [l_0, .., l, ..., L]^T$) pour estimer les formes des relations exposition-réponse et temps-réponse respectivement. Les étapes de cette modélisation sont implémentées dans le package *dlnm* du logiciel R (Gasparrini, 2011).

Avant de pouvoir choisir le modèle adéquat pour caractériser une exposition médicamenteuse, il est nécessaire de capturer ses différentes variations jusqu'à la survenue éventuelle d'évènements chez les sujets. Pour le faire, les enquêtes de cohorte (qui permettent de suivre un groupe d'individus qui sont exempts de maladie au début du suivi pendant une période de temps spécifique jusqu'à l'apparition d'une certaine maladie) sont généralement utilisées. Ceci permet d'obtenir des délais de survenue d'évènement et donc des données de survie.

1.2 Analyse des données de survie

L'analyse des données de survie s'intéresse au délai écoulé depuis une origine fixée (naissance, début d'un traitement, inclusion dans une cohorte, moment d'un diagnostic,...) jusqu'à la survenue d'un évènement d'intérêt (décès, maladie, guérison d'une maladie, évolution d'une maladie, ...).

Dans cette partie, nous introduirons les notations qui seront utilisées et présenterons les concepts de base, les fonctions statistiques d'intérêt ainsi que les différentes méthodes dites classiques (les méthodes non paramétriques de Kaplan-Meier (Kaplan & Meier, 1958) et de Nelson-Aalen (Aalen, 1978 ; Nelson, 1972), les méthodes paramétriques et le modèle semi-paramétrique de Cox (Cox, 1972) d'analyse de survie.

1.2.1 Concepts de base

L'analyse de survie nécessite une définition précise de l'évènement d'intérêt, mais aussi du début et de la fin de la période de suivi et un choix pertinent de l'échelle de temps.

1) Date d'origine et date d'entrée

La date d'origine est le point de départ du suivi qui doit être défini sans ambiguïté, comparable entre les individus, et idéalement cliniquement pertinent. Des exemples typiques sont la date de naissance, le début d'un traitement ou le moment d'un diagnostic. La date de début du suivi ne correspond pas toujours à une date "cliniquement significative" mais à une date administrative "d'entrée dans l'étude". Les individus ne sont pas toujours suivis à partir de la date origine, mais peuvent parfois être observés à partir d'un moment ultérieur, on parle d'*entrée retardée* dans l'étude.

2) Date de point

C'est la date au-delà de laquelle l'étude est arrêtée et on ne tient plus compte des informations sur les sujets.

3) Échelle de temps

L'échelle de temps la plus souvent utilisée est le temps de suivi, c'est-à-dire le délai depuis la date d'entrée dans l'étude, à l'image des essais cliniques dans lesquels la date d'entrée correspond à la date de randomisation. Cependant, l'âge est une autre échelle

de temps couramment utilisée en épidémiologie pour étudier l'incidence d'un évènement de santé ou la mortalité. Certains auteurs le recommandent par rapport au temps de suivi en particulier lorsque le risque de survenue de l'évènement d'intérêt peut être fortement influencé par l'âge des individus pour mieux contrôler l'effet de l'âge (Korn et al., 1997 ; Thiébaut & Bénichou, 2004).

4) Censure et troncature

Une spécificité des données de survie qui doit être prise en compte dans leur analyse est la présence des données incomplètes. Les phénomènes de censure et de troncature sont à l'origine de ces données incomplètes et doivent être pris en compte dans l'écriture de la vraisemblance dans les méthodes d'analyse. Il existe plusieurs formes de censure et de troncature mais nous ne présenterons que les censures à droite et par intervalle ainsi que la troncature à gauche qui sont les formes de censure et de troncature (en particulier lorsque l'âge est l'échelle de temps) les plus courantes en analyse de survie.

- La censure à droite correspond à la situation où le suivi d'un sujet se termine avant que l'évènement d'intérêt ne se soit produit. Ainsi, un individu peut être censuré à droite pour plusieurs raisons telles que : l'étude se termine à un moment pré-spécifié qui précède le moment où l'évènement se produit chez le sujet (censure administrative à droite), le sujet interrompt le suivi avant que l'évènement ne se produise pour une raison quelconque (perdus de vue) ou le sujet subit un autre évènement qui empêche l'observation de l'évènement d'intérêt (risque concurrent).

La plupart des méthodes d'analyse des données de survie supposent que le temps de censure est indépendant du temps de survie : c'est l'hypothèse d'une censure non informative. Cette hypothèse est généralement raisonnable dans le cas de la censure administrative à droite, puisque la fin du suivi est pré-spécifiée. Cependant, dans le cas d'une censure aléatoire à droite ou en présence de risque concurrent, elle est moins clairement vérifiée.

- La censure par intervalle correspond à la situation où l'on ne sait pas à quel moment exact l'évènement d'intérêt est apparu chez un sujet mais seulement que cela s'est produit dans un intervalle de temps spécifique. La censure par intervalle est généralement observée lorsque le suivi est réalisé avec des rendez-vous de suivi réguliers. Ainsi, la seule information disponible sur le temps de survie est donnée par les dates des rendez-vous entre lesquelles l'évènement d'intérêt s'est produit.

- La troncature correspond à la situation où certains sujets ne sont pas observables et on n'étudie qu'un sous-échantillon. En effet, le temps de survie est dit tronqué à

gauche si son observation est conditionnelle à un autre évènement c'est-à-dire s'il n'est observable qu'après un temps précis. Par exemple, lorsqu'on étudie le délai depuis la naissance jusqu'à l'apparition d'un évènement de santé et que les sujets ne sont pas suivis depuis leur date de naissance, le temps de survie est tronqué à gauche car seuls les sujets vivants et indemnes de l'évènement de santé d'intérêt à la date d'inclusion sont observables (COMMENGES et al., 1998).

1.2.2 Notations

Pour un individu i d'un échantillon de n sujets $i = 1, ..., n$, considérons :
- le temps de survie ou délai avant survenue de l'évènement d'intérêt T_i
- le temps de censure C_i
- le temps d'observation $X_i = min(T_i, C_i)$
- la variable indicatrice δ_i indiquant si, au temps X_i, l'évènement d'intérêt s'est produit ou non :

$$\delta_i = I(T_i \leq C_i) = \begin{cases} 1 & \text{si l'évènement est observé pour le sujet } i \\ 0 & \text{sinon} \end{cases} \tag{1.7}$$

- le temps V_i $(V_i < X_i)$ où l'individu est entré dans l'étude. Si le sujet i est inclus dans l'étude à la date d'origine choisie, alors $V_i = 0$
- un vecteur de valeurs de p covariables $\mathbf{Z}_i(t) = (Z_{i1}(t), ..., Z_{ip}(t))$; $t \in [V_i, X_i]$, dont certaines peuvent être dépendantes du temps et d'autres fixes dans le temps.

Il existe deux approches pour la notation des données de survie. Une approche traditionnelle pour décrire les données de survie est donnée par le vecteur $(X_i, V_i, \delta_i, \mathbf{Z}_i(t))$ et elle fonctionne bien pour le délai jusqu'à l'apparition d'un seul évènement (par exemple le décès toutes causes confondues, l'exemple classique le plus simple des données de survie). On peut également s'intéresser aux variations de l'état (sain, malade, mort,...) d'un sujet au cours du temps, ce qui conduit à la notation empruntée à la théorie des processus de comptage qui s'adapte à des situations plus complexes telles que les processus multi-états (représentant l'état occupé par un individu au cours du temps). L'exemple précédent est un processus multi-états simple avec deux états (vivant et mort) et une transition entre ces deux états.

Soient les processus suivants (Kalbfleisch & Prentice, 2002) :

$$Y_i(t) = I(V_i < t \leq X_i) = \begin{cases} 1 & \text{si le sujet } i \text{ est à risque à l'instant } t \\ 0 & \text{sinon} \end{cases} \tag{1.8}$$

et

$$N_i(t) = I(V_i < X_i \leq t, \delta_i = 1) \tag{1.9}$$

$$= \begin{cases} 1 & \text{si le sujet } i \text{ a expérimenté l'évènement au temps } t \\ 0 & \text{sinon} \end{cases} \tag{1.10}$$

Le nombre d'individus encore à risque de subir l'évènement à l'instant t est donné par :

$$Y(t) = \sum_{i=1}^{n} Y_i(t) \tag{1.11}$$

et le nombre d'évènements observés au temps t, est donné par :

$$N(t) = \sum_{i=1}^{n} N_i(t) \tag{1.12}$$

En définissant le changement d'état $dN_i(t) = N_i(t) - N_i(t - dt)$ du sujet i entre les instants $t - dt$ et t, l'observation de l'évènement au temps t peut alors être représentée par $dN_i(t) = 1$.

Dans la suite, nous utiliserons la notation traditionnelle pour décrire les données de survie car nous nous limiterons dans cette thèse au risque de survenue d'un seul évènement indésirable associé à une exposition médicamenteuse.

1.2.3 Fonctions statistiques d'intérêt pour l'analyse de survie

Le temps de survie ou délai avant survenue de l'évènement T est la variable centrale de l'analyse des données de survie. T est une variable aléatoire continue positive dont la distribution peut être caractérisée par plusieurs fonctions définies pour $t \geq 0$ (t correspondant aux réalisations de T).

1) La densité de probabilité $f(t)$ représente la probabilité dite instantanée que l'évènement survienne dans un intervalle de temps infinitésimal après t. Elle est définie par :

$$f(t) = \lim_{dt \to 0} \frac{\mathbb{P}(t \leq T < t + dt)}{dt} \tag{1.13}$$

2) La fonction de répartition $F(t)$ est la probabilité que l'évènement apparaisse avant le temps t. Elle correspond à :

$$F(t) = \mathbb{P}(T \leq t) = \int_0^t f(u)du. \tag{1.14}$$

Elle est croissante. En effet, la probabilité d'avoir eu l'évènement avant l'origine ($t = 0$) est égale à 0 ($F(0) = 0$). Au fur et à mesure que le temps de suivi s'allonge, le nombre d'individus ayant connu l'évènement augmente et donc sa valeur augmente.

3) La fonction de survie $S(t)$ est la probabilité de survivre au-delà d'un temps t. Elle représente ainsi la proportion d'individus encore susceptibles de connaître l'évènement d'intérêt à un moment donné. La fonction $S(t)$ est toujours décroissante. En effet, à la date d'origine ($t = 0$), aucun individu n'a connu l'évènement d'intérêt et sa valeur est égale à 1 ($S(0) = 1$). Au fur et à mesure que le temps s'écoule et que le nombre d'individus ayant connu l'évènement augmente, sa valeur diminue. Elle est définie par :

$$S(t) = \mathbb{P}(T > t) = 1 - F(t) \tag{1.15}$$

4) La fonction de risque instantané $h(t)$ représente la probabilité que l'évènement d'intérêt survienne dans un intervalle de temps infinitésimal après t, sachant qu'il n'a jamais été observé avant t. Elle est au centre de l'analyse des données de survie car elle permet d'étudier si et quand un individu a présenté l'évènement depuis le début de son suivi. Elle est définie par :

$$h(t) = \lim_{dt \to 0} \frac{\mathbb{P}(t \leq T < t + dt | T \geq t)}{dt} = \frac{f(t)}{S(t)} \tag{1.16}$$

5) La fonction de risque cumulé correspond à :

$$H(t) = \int_0^t h(u)du = -\log(S(t)). \tag{1.17}$$

Il ressort des définitions et des relations ci-dessus que la connaissance de l'une des cinq fonctions permet de retrouver les autres.

1.2.4 Vraisemblance

Soit θ le vecteur des paramètres décrivant le risque de base d'un modèle de survie. Les estimateurs de ces paramètres sont obtenus en maximisant la vraisemblance du modèle choisi. Cette vraisemblance (L) représente la probabilité d'observer l'échantillon d'après le modèle et est le produit des n contributions individuelles ($L_i, i = 1, ..., n$) à cette vraisemblance :

$$L = \prod_{i=1}^{n} L_i \tag{1.18}$$

Soit x_i le temps d'observation du sujet i. En présence de données censurées, la contribution du sujet i à la vraisemblance dépend de l'état de son observation :

- Si le temps d'observation est un temps d'évènement ($\delta_i = 1$), c'est-à-dire le sujet i a connu l'événement d'intérêt à l'instant x_i, alors sa contribution à la vraisemblance sera $f(x_i; \theta)$ où f représente la densité de probabilité de la variable aléatoire T ;

- Si le temps d'observation est un temps de censure à droite ($\delta_i = 0$), alors sa contribution à la vraisemblance sera $S(x_i; \theta)$ où S représente la fonction de survie associée à T.

Ainsi, en supposant des temps d'évènement et de censure indépendants, la vraisemblance de l'ensemble des n observations s'écrit :

$$L = \prod_{i=1}^{n} f(x_i; \theta)^{\delta_i} \ S(x_i; \theta)^{1-\delta_i} = \prod_{i=1}^{n} h(x_i; \theta)^{\delta_i} \ S(x_i; \theta) \tag{1.19}$$

En présence de données tronquées à gauche, la contribution individuelle d'un sujet i dont l'observation est tronquée à gauche en V_i est :

$$L_i = \begin{cases} \dfrac{f(x_i; \theta)}{S(V_i; \theta)} & \text{si } \delta_i = 1 \\ \\ \dfrac{S(x_i; \theta)}{S(V_i; \theta)} & \text{si } \delta_i = 0 \end{cases} \tag{1.20}$$

En présence de données censurées par intervalle, la contribution individuelle d'un

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

$$L_i = \begin{cases} S(r_i;\theta) - S(l_i;\theta) & \text{sans troncature à gauche} \\ \dfrac{S(r_i;\theta) - S(l_i;\theta)}{S(V_i;\theta)} & \text{avec troncature à gauche} \end{cases} \tag{1.21}$$

1.2.5 Méthodes d'estimation non paramétriques

Elles permettent d'estimer les fonctions caractéristiques de la distribution des temps de survie sans faire d'hypothèse a priori sur celle-ci. Les méthodes non paramétriques prennent en compte les données censurées à droite et sont souvent utilisées comme première étape dans une analyse de survie pour générer des statistiques descriptives non biaisées.

1.2.5.1 Estimateur de Kaplan-Meier de la fonction de survie

Soient un échantillon $t_j, j = 1, ..., J$ de temps d'évènement observés, r_j le nombre d'individus encore à risque de subir l'évènement à l'instant t_j et d_j le nombre d'individus ayant connu l'évènement d'intérêt au temps t_j ($d_j = 1$ en cas d'évènement unique ou $d_j > 1$ en cas d'ex-æquo c'est-à-dire s'il y'a plusieurs évènements au même temps t_j).

L'estimateur de Kaplan-Meier, parfois appelé estimateur *produit-limite* (Kaplan & Meier, 1958) au temps t, est défini par :

$$\hat{S}(t) = \prod_{j:t_j \leq t} \left(1 - \frac{d_j}{r_j}\right) \tag{1.22}$$

Il représente ainsi une fonction constante par morceaux (fonction en escalier) qui présente des sauts aux seuls temps t_j où l'on observe un évènement.

Un estimateur de la variance de l'estimateur de Kaplan-Meier est donné par la formule de Greenwood (Greenwood, 1926) :

$$\widehat{Var}(\hat{S}(t)) = \hat{S}^2(t) \sum_{j:t_j \leq t} \frac{d_j}{(r_j - d_j) r_j} \tag{1.23}$$

1.2.5.2 Estimateur de Nelson-Aalen de la fonction de risque cumulé

L'estimation de la fonction de risque cumulé peut être faite à partir de l'estimateur de Kaplan-Meier en utilisant la relation (2.11) qui lie la fonction de survie et la fonction

de risque cumulé :

$$H(t) = -\log(S(t)) \Longrightarrow \hat{H}_{Br}(t) = -\log \hat{S}(t) \tag{1.24}$$

$$= -\sum_{j:t_j \leq t} \log\left(1 - \frac{d_j}{r_j}\right) \tag{1.25}$$

L'estimateur obtenu est appelé estimateur de Breslow (Breslow, 1972, 1974). Cependant, l'estimateur préférentiel pour la fonction de risque cumulé est celui de Nelson-Aalen (Aalen, 1978 ; Nelson, 1972) :

$$\hat{H}_{NA}(t) = \sum_{j:t_j \leq t} \frac{d_j}{r_j} \tag{1.26}$$

qui permet d'obtenir un autre estimateur de la fonction de survie (appelé estimateur de Fleming-Harrington) $\hat{S}_{FH}(t) = \exp(-\hat{H}_{NA}(t))$. Les estimateurs de Breslow et de Nelson-Aalen sont asymptotiquement équivalents.

1.2.6 Méthodes d'estimation paramétriques

Elles permettent d'estimer les fonctions caractéristiques de la distribution des temps de survie en supposant que cette distribution appartient à une famille donnée de lois paramétriques : exponentielle, de Weibull, gamma, de Gompertz, log-normale, log-logistique et gaussienne inverse (Bagdonavicius & Nikulin, 2001). Nous présenterons les lois exponentielle, de Weibull (les plus utilisées) et de Gompertz qui seront utilisées dans une autre partie de ce manuscrit.

1.2.6.1 Loi exponentielle

La distribution exponentielle, qui ne dépend que d'un paramètre $\lambda > 0$, est la seule distribution continue qui suppose un risque instantané constant. Pour $t \geq 0$ et $\lambda > 0$, elle est caractérisée par :

$$h(t|\lambda) = \lambda,$$

$$S(t|\lambda) = \exp(-\lambda t),$$

$$f(t|\lambda) = \lambda \exp(-\lambda t),$$

$$F(t|\lambda) = 1 - \exp(-\lambda t),$$

$$H(t|\lambda) = \lambda t.$$

1.2.6.2 Loi de Weibull

La distribution de Weibull dépend de deux paramètres : un paramètre d'échelle α et un paramètre de forme γ. Pour $t \geq 0$ et avec $\alpha > 0$ et $\gamma > 0$, les fonctions caractéristiques d'une loi de Weibull sont définies comme suit :

$$\begin{aligned}
h(t|\alpha,\gamma) &= \alpha\gamma t^{\gamma-1}, \\
S(t|\alpha,\gamma) &= \exp(-\alpha t^{\gamma}), \\
f(t|\alpha,\gamma) &= \alpha\gamma t^{\gamma-1}\exp(-\alpha t^{\gamma}), \\
F(t|\alpha,\gamma) &= 1-\exp(-\alpha t^{\gamma}), \\
H(t|\alpha,\gamma) &= \alpha t^{\gamma}.
\end{aligned}$$

Ainsi, la loi de Weibull est une généralisation de la loi exponentielle de paramètre α que l'on retrouve en prenant $\gamma = 1$. De plus, le risque instantané $h(t)$ est une fonction strictement croissante si $\gamma > 1$ et strictement décroissante si $\gamma < 1$.

1.2.6.3 Loi de Gompertz

Elle est très utilisée pour la distribution du taux de mortalité. Pour $t \geq 0$ et les paramètres d'échelle $\alpha > 0$ et de forme $\gamma \in (-\infty, \infty)$, elle est caractérisée par :

$$\begin{aligned}
h(t|\alpha,\gamma) &= \alpha\exp(\gamma t), \\
S(t|\alpha,\gamma) &= \exp\left[\frac{\alpha}{\gamma}(1-\exp(\gamma t))\right], \\
f(t|\alpha,\gamma) &= \alpha\exp(\gamma t)\exp\left[\frac{\alpha}{\gamma}(1-\exp(\gamma t))\right], \\
F(t|\alpha,\gamma) &= 1-\exp\left[\frac{\alpha}{\gamma}(1-\exp(\gamma t))\right], \\
H(t|\alpha,\gamma) &= \frac{\alpha}{\gamma}(\exp(\gamma t)-1).
\end{aligned}$$

Les modèles paramétriques peuvent être utilisés dans les analyses de régression des données de survie lorsque les effets des variables sur le temps de survie doivent être étudiés. L'estimation des paramètres des modèles paramétriques peut être effectuée en utilisant la méthode du maximum de vraisemblance.

1.2.7 Modèle semi-paramétrique de Cox

Le modèle de Cox à risques proportionnels est l'approche multivariable la plus couramment utilisée pour analyser les données de survie en recherche médicale (Ryan & Woodall, 2005). C'est un modèle de régression qui permet d'évaluer l'effet des covariables sur la distribution du temps de survie. En effet, ce modèle se concentre sur la fonction de risque, définissant le risque au temps t en fonction de covariables connues en utilisant un modèle multiplicatif de la forme (Cox, 1972) :

$$h(t|\mathbf{Z}_i) = h_0(t)\exp(\beta^T\mathbf{Z}_i) \tag{1.27}$$

où i est l'indice du sujet,
$h_0(t)$ est la fonction de risque instantané de base, c'est-à-dire le risque instantané des sujets lorsque toutes les covariables sont nulles ou à leur niveau de référence,
$\mathbf{Z}_i$ est le vecteur de covariables supposées fixes dans le temps,
β est le vecteur des coefficients de régression.

C'est un modèle semi-paramétrique puisqu'il est constitué d'une première partie définie par h_0 sur laquelle aucune hypothèse paramétrique n'est faite et d'une seconde partie définie avec une fonction explicite des covariables faisant intervenir des coefficients de régression ou paramètres inconnus.

Le risque qu'un individu i ayant des caractéristiques $\mathbf{Z}_i$ connaisse l'évènement d'intérêt à l'instant t par rapport au même risque pour un individu $i*$ ayant des caractéristiques $\mathbf{Z}_{i*}$ est défini par le ratio :

$$\frac{h(t|\mathbf{Z}_i)}{h(t|\mathbf{Z}_{i*})} = \frac{h_0(t)\exp(\beta^T\mathbf{Z}_i)}{h_0(t)\exp(\beta^T\mathbf{Z}_{i*})} = \exp(\beta^T(\mathbf{Z}_i - \mathbf{Z}_{i*})) \tag{1.28}$$

Ce rapport des risques instantanés est constant au cours du temps : c'est l'hypothèse des risques proportionnels qui caractérise le modèle de Cox.

1.2.7.1 Vraisemblance partielle de Cox

L'intérêt de ce modèle de Cox semi-paramétrique réside dans la séparation de la fonction de risque instantané de base $h_0(t)$ et des coefficients de régression β. Ainsi, le modèle de Cox évite les hypothèses paramétriques sur la forme de $h_0(t)$ pour estimer les paramètres de régression du modèle en maximisant la vraisemblance dite partielle.

La vraisemblance partielle peut être construite en considérant la probabilité conditionnelle qu'un sujet i avec un ensemble de covariables $\mathbf{Z}_i$ subisse l'évènement d'intérêt à un temps particulier t_i, sachant qu'un évènement a été observé à ce temps t_i. Cela suppose

implicitement qu'un seul évènement se produit à chaque temps d'évènement, c'est-à-dire qu'il n'y a pas de temps d'évènements ex-æquos. Cette probabilité ou contribution à la vraisemblance partielle s'écrit :

$$L_i(\beta) = \frac{h_0(t_i)\exp(\beta^T \mathbf{Z}_i)}{\sum_{k \in R(t_i)} h_0(t_i)\exp(\beta^T \mathbf{Z}_k)} = \frac{\exp(\beta^T \mathbf{Z}_i)}{\sum_{k \in R(t_i)} \exp(\beta^T \mathbf{Z}_k)} \tag{1.29}$$

La fonction de risque instantané de base qui est totalement inconnue, est ainsi éliminée et la vraisemblance partielle complète comprenant tous les J temps d'évènement est définie par (Cox, 1972, 1975) :

$$L(\beta) = \prod_{i=1}^{J} \frac{\exp(\beta^T \mathbf{Z}_i)}{\sum_{k \in R(t_i)} \exp(\beta^T \mathbf{Z}_k)} \tag{1.30}$$

qui, à chaque temps d'évènement observé, compare les valeurs des covariables de l'individu ayant connu l'évènement à celles de tous les individus encore à risque à ce temps d'évènement (Klein et al., 2014) dont l'ensemble est représenté au temps t_i par $R(t_i)$. Cet ensemble contient également l'individu ayant connu l'évènement.

La log-vraisemblance partielle est donnée par :

$$\log(L(\beta)) = \sum_{i=1}^{J} \left\{ \beta^T \mathbf{Z}_i - \log\left(\sum_{k \in R(t_i)} \exp(\beta^T \mathbf{Z}_k) \right) \right\} \tag{1.31}$$

Pour estimer les coefficients, nous maximisons la vraisemblance partielle (ou la log-vraisemblance partielle) en utilisant l'équation de score :

$$U(\beta) = \frac{\partial \log(L(\beta))}{\partial \beta} = \sum_{i=1}^{J} \left\{ \mathbf{Z}_i - \sum_{k \in R(t_i)} \frac{\mathbf{Z}_k \exp(\beta^T \mathbf{Z}_k)}{\sum_{k \in R(t_i)} \exp(\beta^T \mathbf{Z}_k)} \right\} = 0 \tag{1.32}$$

Nous pouvons donc déterminer l'estimateur du maximum de vraisemblance partielle $\hat{\beta}$ tel que : $U(\hat{\beta}) = 0$.

1.2.7.2 Extensions du modèle de Cox à risques proportionnels

Dans de nombreuses études, certaines covariables peuvent d'une part ne pas vérifier l'hypothèse des risques proportionnels du modèle de Cox et d'autre part changer de valeur au cours du temps. Des extensions du modèle de Cox existent pour prendre en compte ces spécificités.

1) Extensions du modèle de Cox pour les risques non proportionnels

Le modèle de Cox dans le cas des risques non proportionnels (modèle de Cox avec des effets dépendants du temps) permet au coefficient de régression de varier avec le temps et est défini par

$$h(t|Z) = h_0(t)\exp(\beta(t)Z) \tag{1.33}$$

pour une covariable Z, $\beta(t)$ étant une fonction du temps. Nous présentons deux façons de modifier ce modèle pour assouplir l'hypothèse de proportionnalité des risques : l'introduction dans le modèle d'une interaction de la covariable d'intérêt avec le temps et la stratification du modèle.

a) Interaction avec le temps

Lorsqu'une covariable Z ne satisfait pas l'hypothèse des risques proportionnels du modèle de Cox $h(t|Z) = h_0(t)\exp(\beta Z)$, on peut ajouter au modèle une autre covariable qui est définie comme une interaction entre cette covariable Z et une fonction du temps pré-spécifiée $f(t)$ (Cox, 1972). Le nouveau modèle est de la forme :

$$h(t|Z) = h_0(t)\exp(\beta Z + \gamma f(t)Z). \tag{1.34}$$

Cette méthode permet à la fois d'identifier dans le modèle les covariables qui ne vérifient pas l'hypothèse des risques proportionnels et d'apporter une solution au problème. En effet, si la covariable ajoutée au modèle s'avère être statistiquement significative, cela indique que l'hypothèse des risques proportionnels n'est pas satisfaite pour une covariable donnée (c'est-à-dire que son effet varie avec le temps). L'inclusion de l'interaction dans le modèle permet d'interpréter les paramètres en tenant compte du fait que l'influence de la covariable sur le risque n'est pas constante. En ce qui concerne la forme de l'interaction (représentée par la fonction f), différentes fonctions peuvent être utilisées : linéaire ($f(t) = t$), logarithmique ($f(t) = \log(t)$), quadratique ($f(t) = t^2$)), inverse ($f(t) = 1/t$), ...

b) Modèle de Cox stratifié

Pour affranchir certaines covariables de l'hypothèse des risques proportionnels du modèle de Cox, l'une des solutions est de stratifier le risque de base par rapport à ces covariables. Soit une variable catégorielle qui ne vérifie pas l'hypothèse des risques

proportionnels. Les modalités de cette variable constituent les strates dans lesquelles seront classés les sujets étudiés (chaque sujet appartenant à une seule strate).

Dans le modèle stratifié, le risque de base dépend de la strate considérée et l'effet des autres covariables est identique dans chaque strate. On a donc :

$$h_s(t|\mathbf{Z}) = h_{0s}(t)\exp(\beta^T\mathbf{Z}). \tag{1.35}$$

avec $h_s(t|\mathbf{Z})$ et $h_{0s}(t)$ représentant respectivement la fonction de risque conditionnelle à l'ensemble des covariables et la fonction de risque de base dans la strate s.

La vraisemblance partielle du modèle stratifié est donnée par le produit des vraisemblances partielles de chaque strate et on a :

$$L(\beta) = \prod_{s=1}^{S}\prod_{i=1}^{J_s}\frac{\exp(\beta^T\mathbf{Z}_i)}{\sum_{k\in R_s(t_i)}\exp(\beta^T\mathbf{Z}_k)} \tag{1.36}$$

où S est le nombre de strates, J_s est le nombre d'évènements dans la strate s et $R_s(t_i)$ est l'ensemble des sujets à risque à l'instant t_i appartenant à la strate s.

Après la stratification d'une covariable, il n'est plus possible d'estimer son effet dans ce modèle stratifié mais l'estimation de son interaction avec une autre covariable reste possible. Cependant, la précision des estimations des coefficients diminue avec le nombre de strates. Ainsi, la stratification du risque de base par rapport à une covariable n'est pas adaptée si la covariable présentant un écart à l'hypothèse des risques proportionnels est une variable explicative d'intérêt.

2) Extension du modèle de Cox avec des covariables dépendantes du temps

Une covariable qui peut changer de valeur au fil du temps est dite dépendante du temps. Par exemple, l'exposition à un traitement médicamenteux peut varier dans le temps en terme de doses et de durée (on peut observer des périodes de diminution ou d'augmentation des doses, d'interruption ou de reprise du traitement).

Un modèle de Cox à risques proportionnels avec des covariables dépendantes du temps peut être défini par :

$$h(t|\mathbf{Z}(t)) \quad = \quad h_0(t)exp(\beta^T\mathbf{Z}(t)) \tag{1.37}$$

où $h_0(t)$ est la fonction de risque de base à l'instant t quand toutes les covariables sont nulles ou à leur niveau de référence, $\mathbf{Z}(t)$ est le vecteur de covariables à l'instant t et β est le vecteur des coefficients de régression associés à $\mathbf{Z}(t)$.

Dans ce cas, la vraisemblance partielle est donnée par :

$$L(\beta) = \prod_{i=1}^{J} \frac{\exp(\beta^T \mathbf{Z}_i(t))}{\sum_{k \in R(t_i)} \exp(\beta^T \mathbf{Z}_k(t))} \tag{1.38}$$

et l'estimateur du maximum de vraisemblance partielle de Cox peut être écrit comme la solution de l'équation :

$$U(\beta) = \sum_{i=1}^{J} \left\{ \mathbf{Z}_i(t) - \sum_{k \in R(t_i)} \frac{\mathbf{Z}_k(t) \exp(\beta^T \mathbf{Z}_k(t))}{\sum_{k \in R(t_i)} \exp(\beta^T \mathbf{Z}_k(t))} \right\} = 0 \tag{1.39}$$

Ces deux équations ont la même structure que celles présentées précédemment (équations 1.30 et 1.32) en remplaçant $\mathbf{Z}$ par $\mathbf{Z}(t)$.

L'hypothèse des risques proportionnels est conservée car l'effet de la covariable ne varie pas au cours du temps.

Par ailleurs, un modèle de Cox avec des effets dépendants du temps peut également être transformé en un modèle à risques proportionnels avec des covariables dépendantes du temps, en utilisant par exemple des régressions splines pour réexprimer ces effets (Abrahamowicz et al., 1996 ; Guess, 2006 ; Hauptmann et al., 2000 ; Platt, 2004).

1.2.7.3 Gestion des temps de survie ex-æquos

La vraisemblance partielle telle que définie précédemment pour estimer les paramètres du modèle de Cox repose sur l'hypothèse de temps distincts, c'est-à-dire que le temps de survie est continu. Cependant, en pratique, il est souvent difficile que l'enregistrement des temps d'évènements soit précis, ce qui conduit à observer, pour des individus différents, des temps d'évènements identiques ou ex-æquos. De ce fait, la présence d'évènements concomitants (évènements enregistrés à des dates identiques) nécessite de modifier l'expression de la vraisemblance partielle pour les prendre en compte et des méthodes existent pour le faire. Nous présentons les trois principales.

1) Méthode exacte

La méthode exacte suppose que les temps de survie ex-æquos résultent d'une mesure imprécise du temps et qu'il existe un ordre chronologique réel de survenue des évènements. Ainsi, en présence de temps de survie ex-æquos, l'expression exacte de la vraisemblance partielle considère toutes les permutations possibles des évènements

observés simultanément au temps t_j (PETO, 1972).

Comme exemple simple, considérons deux sujets 1 et 2 qui présentent simultanément l'évènement au temps t_j. Leur contribution à la vraisemblance partielle de Cox au temps t_j s'écrit :

$$\frac{\exp(\beta^T \mathbf{Z}_1)}{\sum\limits_{i \in R(t_j)} \exp(\beta^T \mathbf{Z}_i)} \times \frac{\exp(\beta^T \mathbf{Z}_2)}{\sum\limits_{\substack{i \in R(t_j) \\ i \neq 1}} \exp(\beta^T \mathbf{Z}_i)} + \frac{\exp(\beta^T \mathbf{Z}_2)}{\sum\limits_{i \in R(t_j)} \exp(\beta^T \mathbf{Z}_i)} \times \frac{\exp(\beta^T \mathbf{Z}_1)}{\sum\limits_{\substack{i \in R(t_j) \\ i \neq 2}} \exp(\beta^T \mathbf{Z}_i)}$$

La vraisemblance partielle exacte est définie par (KALBFLEISCH & PRENTICE, 2002) :

$$L_E(\beta) = \prod_{j=1}^{J} \frac{\exp(\beta^T \mathcal{Z}_j)}{\sum\limits_{p \in P_j} \prod_{q=1}^{d_j} \left[\sum\limits_{i \in R(t_j, p, q)} \exp(\beta^T \mathbf{Z}_i) \right]} \tag{1.40}$$

où J désigne le nombre total d'évènements survenus dans la cohorte, d_j est le nombre d'évènements survenus au temps t_j, P_j est l'ensemble des $d_j!$ permutations, p est un élément de P_j et est défini comme $p = (e_1, ..., e_{d_j})$, $R(t_j, p, q) = R(t_j) - \{e_1, ..., e_{q-1}\}$ et $\mathcal{Z}_j = \sum_{i \in D_j} \mathbf{Z}_i$ avec D_j l'ensemble des sujets présentant l'évènement au temps t_j.

A mesure que le nombre de temps de survie ex-æquos augmente, la méthode exacte devient très compliquée à mettre en œuvre et nécessite un temps de calcul de plus en plus long (ALLISON, 2010). En effet, si seulement 8 individus ont le même temps d'évènement, alors leur contribution à la vraisemblance partielle à ce temps correspondra à 40 320 (8!) termes (40 320 ordres possibles à évaluer). Pour remédier à ce problème, plusieurs approximations ont été proposées parmi lesquelles celles de Breslow et d'Efron qui sont très utilisées.

2) Approximation de Breslow

Une approximation de la vraisemblance partielle en présence des ex-æquos qui simplifie considérablement les calculs qui en résultent a été proposée par Breslow (BRESLOW, 1974). Cette approximation de la méthode exacte permet de ne pas avoir à considérer chaque permutation et suppose que l'ensemble à risque est identique pour tous les évènements survenus simultanément à un temps t_j de sorte que l'expression de la vraisemblance partielle soit :

$$L_B(\beta) = \prod_{j=1}^{J} \frac{\exp(\beta^T \mathcal{Z}_j)}{\left[\sum_{i \in R(t_j)} \exp(\beta^T \mathbf{Z}_i)\right]^{d_j}} \tag{1.41}$$

Toutefois, si le nombre d'évènements ex-æquos à une date donnée est relativement important, cette méthode peut ne pas donner une bonne approximation de la vraisemblance partielle exacte (Allison, 2010 ; Hertz-Picciotto & Rockhill, 1997). Cette approximation est utilisée par défaut par les logiciels SAS et Stata.

3) Approximation d'Efron

Une approximation alternative à celle de Breslow a été suggérée par Efron (Efron, 1977). Dans sa démarche, il considère qu'il y a une modification de l'ensemble à risque à chaque apparition d'un évènement. Pour prendre en compte cela, il choisit une permutation parmi les $d_j!$ possibles (chacune ayant la même probabilité d'être sélectionnée). Puis, dans le produit entre évènements (de la permutation), il affecte un poids aux observations associées à chaque évènement (de l'ensemble à risque) qui diminue à mesure que l'ordre d'apparition de l'évènement augmente. La vraisemblance partielle est ainsi approximée par :

$$L_{Ef}(\beta) = \prod_{j=1}^{J} \frac{\exp(\beta^T \mathcal{Z}_j)}{\prod_{q=1}^{d_j} \left[\sum_{i \in R(t_j)} \exp(\beta^T \mathbf{Z}_i) - \frac{q-1}{d_j} \sum_{i \in D_j} \exp(\beta^T \mathbf{Z}_i)\right]} \tag{1.42}$$

L'approximation d'Efron peut générer également un biais (moins important que celle de Breslow) dans l'estimation des paramètres, pour les ensembles de données comportant une fraction importante d'évènements ex-æquos à une date donnée dans l'ensemble à risque à cette date. Elle est utilisée par défaut avec le logiciel R.

En pratique, la méthode exacte, bien que chronophage (dans son étude publiée en 2014, Borucka a montré que le temps de calcul nécessaire pour obtenir les estimations des paramètres par le modèle de Cox à risques proportionnels était de 2,984s, 0,099s et 0,075s avec la méthode exacte et les approximations d'Efron et de Breslow respectivement), est dans la mesure du possible à privilégier. Sinon, plusieurs travaux ont montré qu'il fallait privilégier l'approximation d'Efron par rapport à celle de Breslow (Allison, 2010 ; Hertz-Picciotto & Rockhill, 1997) car elle génère un biais moins impor-

tant. En effet, pour un scénario d'une moyenne de 10 évènements ex-æquos dans un intervalle de temps donné, Hertz-Picciotto et Rockhill (1997) ont montré que le biais moyen des estimations obtenues à partir d'une régression de Cox était de -46,16 % avec l'approximation de Breslow tandis qu'il était de 0,42 % avec celle d'Efron.

Si le nombre d'évènements ex-æquos à une date donnée est très faible, les trois méthodes donnent des résultats très similaires. Pour les ensembles de données sans évènements ex-æquos, toutes les méthodes conduisent exactement aux mêmes résultats.

1.2.8 Problèmes liés aux données éparses et méthodes de résolution

Les estimations de l'effet des covariables par la méthode du maximum de vraisemblance (conditionnelle) telle que présentée ci-dessus (par les équations du score) peuvent être confrontées à plusieurs problèmes liés aux données éparses, notamment une estimation biaisée ou infinie du coefficient de régression et un échec fréquent de la convergence de la vraisemblance en raison de la séparation (Firth, 1993 ; Greenland et al., 2000 ; Heinze & Schemper, 2001).

Dans cette section nous présenterons les problèmes liés aux données éparses, les éléments contribuant à ces problèmes dans les analyses de régression avec maximum de vraisemblance ainsi que les méthodes pour les détecter et les résoudre.

1.2.8.1 Problèmes liés aux données éparses

Les effets des covariables sont généralement estimés par la méthode du maximum de vraisemblance dans les modèles de régression logistique (conditionnelle), de Cox ou de Poisson. Ces méthodes supposent que le nombre d'évènements observés est suffisant dans toutes les catégories de la variable d'exposition pour donner lieu à des estimations ajustées fiables (Greenland et al., 2016). Malheureusement, lorsque les évènements sont rares dans les données pour une certaine combinaison exposition-réponse, la méthode du maximum de vraisemblance peut produire des estimations qui s'éloignent des vraies valeurs des paramètres, même si aucun autre biais n'est présent. Le biais qui en résulte est appelé biais de petits échantillons ou biais de données éparses car il peut se produire dans des ensembles de données assez vastes avec un grand nombre de paramètres à estimer. À l'extrême, une valeur extrêmement grande ou infinie et une plage d'intervalle de confiance inhabituellement large peuvent être obtenues pour le paramètre estimé : c'est le problème de la séparation (quasi) complète dans un modèle de données éparses (Greenland et al., 2000 ; Heinze & Schemper, 2001 ; Rahman & Sultana, 2017). La séparation complète survient lorsque pour une catégorie ou strate d'une variable explicative, aucun individu ou tous les individus sont concernés par une modalité de la variable de réponse (voir A dans le tableau 1.1) tandis que la séparation

quasi complète survient lorsque pour une catégorie ou strate d'une variable explicative, très peu d'individus sont concernés par une modalité de la variable de réponse (voir B dans le tableau 1.1).

A

Statut d'exposition	Évènement		Total
	Oui	Non	
Exposé	7	2	9
Non exposé	0	21	21
Total	7	23	30

B

Statut d'exposition	Évènement		Total
	Oui	Non	
Exposé	7	2	9
Non exposé	1	20	21
Total	8	22	30

Tableau 1.1 – *Exemple hypothétique pour illustrer une séparation complète (A) ou quasi-complète (B)*

Une idée fausse courante est que ce biais est traité par l'appariement et la régression logistique conditionnelle, mais le biais des données éparses peut être grave avec ces méthodes même lorsqu'il y a un grand nombre d'ensembles appariés avec peu de covariables dans le modèle. En effet, les estimations s'éloignent de la valeur nulle en raison de petits effectifs d'observations au sein de la plupart des strates définies par des covariables catégorielles. Le biais qui en découle est ainsi dû à des ensembles appariés épars et peut facilement doubler l'ampleur des estimations de l'"odds ratio", même en l'absence de confusion, de biais de sélection ou d'erreur de mesure (Greenland et al., 2000). En raison de l'équivalence algébrique des vraisemblances partielles des modèles de régression logistique conditionnelle et de Cox, les mêmes biais peuvent affecter les analyses des modèles de Cox.

Les éléments suivants observés pendant l'examen ou l'analyse des données doivent attirer l'attention sur l'existence (ou la présence) éventuelle des données éparses et donc des biais sous-jacents dans les analyses de régression par maximum de vraisemblance (Greenland et al., 2016) :

(i) La présence d'un petit nombre d'évènements d'intérêt pour chaque variable étudiée, en particulier si tous (ou presque tous) les évènements tombent dans un groupe d'étude (ou catégorie d'exposition) et aucun (ou presque aucun) dans l'autre. Cet élément signale un problème de (quasi) séparation qui peut être facilement détecté en établissant un tableau croisé des modalités de chaque prédicteur avec celles de la variable de réponse, tel que présenté dans le tableau 1.1.

(ii) Des grandes valeurs de l'"odds ratio", souvent avec des intervalles de confiance larges, qui ne correspondent pas à des attentes raisonnables.

(iii) Une augmentation importante de l'ampleur de l'effet estimé après les ajustements statistiques qui tiennent compte des variables de confusion, car le contrôle

des variables de confusion rend généralement les estimations de l'ampleur de l'effet plus petites plutôt que plus grandes.

1.2.8.2 Méthodes de résolution

Il existe plusieurs méthodes pour corriger le biais des données éparses et le problème de séparation dans les analyses de régression avec maximum de vraisemblance, notamment les approches de pénalisation et les méthodes bayésiennes (Greenland et al., 2016 ; Greenland et al., 2000 ; Rahman & Sultana, 2017). Dans le cadre de cette thèse, nous ne présenterons que la méthode de pénalisation de Firth car elle est très utile pour résoudre les problèmes liés aux données éparses dans la régression logistique (conditionnelle) de Poisson et de Cox ; et est incorporée dans les principaux programmes de régression logistique sous les logiciels SAS (Heinze & Ploner, 2002, 2003 ; Heinze & Puhr, 2010), R (Heinze & Ladner, 2012 ; Heinze et al., 2022 ; Heinze et al., 2020 ; Kosmidis, 2013) et Stata (Coveney, 2008).

La pénalisation est une méthode très générale de stabilisation ou de régularisation des estimations, dans laquelle la fonction d'estimation est modifiée et donc partiellement contrainte par des fonctions de paramètres supplémentaires (Greenland & Mansournia, 2015).

L'approche de pénalisation la plus largement programmée semble être la méthode de Firth pour réduire le biais des petits échantillons. Firth a proposé une approche de vraisemblance pénalisée pour la réduction du biais dans la régression logistique non conditionnelle (Firth, 1993). Cette approche permet d'estimer l'effet des covariables β comme le maximum de la log-vraisemblance pénalisée (ou contrainte) par $r(\beta) = \log |I(\beta)|^{0,5}$, $I(\beta)$ désignant la matrice d'information de Fisher. En effet, pour réduire le biais d'une estimation obtenue par la méthode du maximum de vraisemblance, Firth a proposé de multiplier la vraisemblance par $|I(\beta)|^{0.5}$ (loi a priori de Jeffreys qui est une loi a priori non ou peu informative fondée sur l'information de Fisher). La vraisemblance pénalisée est alors de la forme :

$$L_p(\beta) = L(\beta)|I(\beta)|^{0,5} \tag{1.43}$$

et la log-vraisemblance pénalisée est :

$$\log(L_p(\beta)) = \log(L(\beta)|I(\beta)|^{0,5}) = \log(L(\beta)) + \log |I(\beta)|^{0,5} \tag{1.44}$$

où $L(\beta)$ est la vraisemblance non pénalisée. Cette approche permet non seulement de réduire le biais mais elle empêche également les estimations infinies provenant de pro-

blèmes de séparation (Firth, 1993 ; Heinze & Puhr, 2010 ; Heinze & Schemper, 2001). Néanmoins, elle ne minimise pas l'erreur quadratique moyenne (Firth, 1993). Dans la pratique, elle consiste à ajouter 0,5 dans chaque cellule d'un tableau croisé des modalités du prédicteur avec celles de la variable de réponse (Greenland et al., 2016).

Cette méthode a été étendue à la régression logistique conditionnelle (Heinze & Puhr, 2010 ; Sun et al., 2011).

1.3 Schémas des études observationnelles en pharmaco-épidémiologie

Plusieurs schémas d'études observationnelles sont utilisés en pharmaco-épidémiologie pour explorer les bénéfices (dont l'efficacité) ou les risques (effets délétères ou adverses) des produits de santé. Les études de cohorte et cas-témoins nichée sont couramment utilisées comme dans les autres domaines de l'épidémiologie. Cependant, pour pallier certaines de leurs limites, des schémas d'étude alternatifs reposant sur un échantillonnage des individus composant la cohorte sont également utilisés.

1.3.1 Schéma d'étude de cohorte

Les études de cohorte font partie des études observationnelles de type analytique. Elles consistent à suivre dans le temps un groupe de personnes bien définies, afin d'observer la survenue et/ou l'évolution d'un évènement de santé d'intérêt. Les délais avant la survenue de cet évènement de santé ainsi collectés sont étudiés à l'aide de l'analyse des données de survie et permettent de mettre en évidence des liens entre les facteurs de risque ou de protection (exposition médicamenteuse dans cette thèse) et l'évènement de santé étudié.

1) Principe du schéma d'étude

Une étude de cohorte peut être soit prospective, c'est-à-dire que des sujets a priori indemnes de(s) l'évènement(s) étudié(s) sont suivis et l'on recueille des données sur leurs expositions et leur état de santé, soit rétrospective, c'est-à-dire que le recueil des informations se fait a posteriori (à l'aide de données de registres par exemple). Le principe d'une étude de cohorte est de comparer l'incidence d'un évènement d'intérêt chez des sujets exposés à un facteur de risque à celle chez des sujets non exposés. L'incidence de l'évènement d'intérêt au sein des différents groupes est comparée afin de tester l'hypothèse d'association entre l'exposition au facteur étudié et la survenue de l'évènement.

2) Estimation de l'effet des covariables

En recherche médicale, l'analyse des données de la cohorte entière est généralement effectuée à l'aide du modèle de régression de Cox à risques proportionnels tel que présenté à la section (1.2.7) (Ryan & Woodall, 2005). D'autres modèles d'analyse sont parfois utilisés tels que la régression de Poisson (Breslow & Day, 1987 ; Breslow et al., 1983 ; Preston, 2000 ; Viel, 2004) ou la régression logistique groupée (D'Agostino et al., 1990) ; ils ne sont toutefois pas considérés dans le cadre de cette thèse.

3) Limites du schéma d'étude

Les études de cohorte proposent les meilleures conditions pour juger du rôle sur la santé des facteurs de risque ou de protection, en permettant de prendre en compte les évolutions temporelles et les interactions entre facteurs. Cependant, elles ont des limites et leur mise en œuvre n'est pas sans difficultés.

Les problèmes liés à la constitution de la cohorte et au suivi des sujets (par exemple l'attrition due aux sujets perdus de vue) peuvent entraîner un bais de sélection qui entame la validité des résultats d'analyse des données de la cohorte entière. Ce biais affectera également les résultats des études réalisées à partir d'un échantillon de la cohorte.

De plus, lorsque l'évènement de santé est rare ou à long délai d'apparition et la mesure de l'exposition difficile à obtenir, les études de cohorte sont moins efficientes car elles nécessitent la collecte des données sur un grand nombre d'individus pendant une longue période de temps (pour obtenir un nombre suffisant de cas), ce qui est difficile à mettre en œuvre et très coûteux.

Dans ces situations, d'autres approches, notamment les études basées sur de plus petits échantillons sont indispensables.

1.3.2 Schémas d'étude basés sur l'échantillonnage des témoins dans la cohorte

Ils consistent à recruter d'une part tous les cas (individus qui présentent l'évènement de santé d'intérêt), et d'autre part un échantillon aléatoire de témoins (individus qui ne présentent pas l'évènement de santé d'intérêt). Des données relatives à l'ensemble des sujets des deux groupes sont recueillies dans le but de mettre en évidence des facteurs de risque ou de protection vis-à-vis de l'évènement de santé étudié. Dans cette thèse nous présenterons deux d'entre eux : les schémas d'étude cas-témoins nichée et cas-cohorte.

1.3.2.1 Schéma cas-témoins niché

L'étude cas-témoins nichée proposée par Thomas en 1977 est une autre approche largement utilisée pour explorer l'association entre une exposition médicamenteuse et un évènement de santé (Billioti de Gage et al., 2012 ; Hernán et al., 2004 ; Le Vu et al., 2021). Elle découle d'une étude de cohorte dans laquelle les cas sont identifiés, puis comparés à un sous-échantillon de non-cas (témoins) sélectionnés dans la cohorte. Les témoins sont sélectionnés parmi les sujets de la cohorte à risque juste avant le moment où l'évènement de santé survient (appariement sur le temps).

1) Principe du schéma d'étude

L'étude cas-témoins nichée est menée comme une étude cas-témoins basée sur tous les cas et une sélection aléatoire de témoins appariés individuellement aux cas sur le temps d'apparition de l'évènement, en utilisant l'échantillonnage de l'ensemble à risque aussi appelé échantillonnage en densité d'incidence (Lubin & Gail, 1984 ; Prentice & Breslow, 1978). L'appariement des témoins aux cas peut en plus se faire éventuellement sur certaines covariables (Breslow & Day, 1980).

La sélection des témoins se fait indépendamment aux différents temps d'évènement. Ainsi, les témoins échantillonnés peuvent être sélectionnés comme témoins pour plus d'un cas et peuvent devenir ultérieurement des cas. La figure 1.1 ci-dessous illustre l'échantillonnage des témoins (Klein et al., 2014).

Il existe plusieurs façons de sélectionner les individus pour former l'ensemble des témoins à chaque temps d'évènement. Le schéma d'étude cas-témoins nichée standard tel que décrit précédemment, utilise une stratégie d'échantillonnage simple des témoins qui consiste à sélectionner aléatoirement et sans remise des témoins dans l'ensemble à risque à chaque temps d'évènement. Cette stratégie conduit à des estimations non biaisées du risque relatif. Une possibilité parfois envisagée, mais qui induit un biais, est de reporter le choix des témoins jusqu'à la fin de la période d'étude et de ne prendre que des témoins "purs", c'est-à-dire des individus qui ont été suivis pendant toute la période d'étude sans devenir des cas (Lubin & Gail, 1984). Cela exclut, en particulier, les individus sélectionnés comme témoins qui deviennent des cas dans le futur. D'autres possibilités connexes à éviter sont que les témoins qui connaissent ultérieurement un autre évènement pouvant être lié à l'exposition soient exclus ou que les personnes pouvant être sélectionnées comme témoins à un temps d'évènement doivent rester exemptes de l'évènement d'intérêt pendant une période spécifiée après leur sélection (Lubin & Gail, 1984).

L'appariement sur le temps d'évènement permet que l'exposition d'intérêt variant au

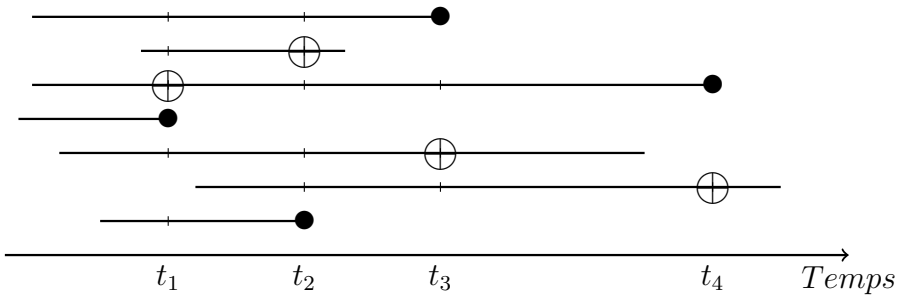


Figure 1.1 – *Échantillonnage des témoins dans une cohorte hypothétique de 7 individus pour une étude cas-témoins nichée avec un témoin par cas ; chaque individu est représenté par une ligne qui commence à un temps d'entrée et finit à un temps d'évènement ou de censure ;* • *temps d'évènement,* | *potentiels témoins appariés et* $\oplus$ *témoins sélectionnés*

cours du temps soit évaluée au moment de la survenue de l'évènement (tant pour le cas que pour ses témoins appariés), puis comparée entre le cas et ses témoins appariés. L'étude cas témoins nichée permet ainsi de prendre en compte une exposition d'intérêt variant au cours du temps et de mieux quantifier son effet, en évitant le recours à des techniques statistiques complexes (Etminan, 2004). L'appariement éventuel sur les covariables dans une étude cas-témoins nichée permet de contrôler les potentielles variables de confusion au stade de la sélection des témoins, un ajustement restant possible au moment de l'analyse.

2) Estimation de l'effet des covariables

L'analyse des données cas-témoins nichées (considérées comme un échantillon de la cohorte) peut être effectuée en utilisant le modèle de Cox à risques proportionnels (équation 1.37) pour exprimer l'association entre la distribution du temps de survie et les covariables fixes ou dépendantes du temps. L'approche classique pour estimer β (vecteur des coefficients de régression associés aux covariables $Z(t)$) avec des données cas-témoins niché consiste donc à maximiser une vraisemblance partielle (notée L_{CTN}) similaire à celle de l'équation (1.38) où les ensembles à risque complets $R(t_i)$ au temps t_i sont remplacés par les ensembles à risque échantillonnés $R'(t_i)$ au temps t_i (Liddell et al., 1977).

$$L_{CTN}(\beta) = \prod_{i=1}^{J} \frac{\exp(\beta^T \mathbf{Z}_i(t))}{\sum_{k \in R'(t_i)} \exp(\beta^T \mathbf{Z}_k(t))} \tag{1.45}$$

$R'(t_i)$ étant limité à un cas et à ses témoins appariés sur le temps t_i de survenue de l'évènement et éventuellement certaines covariables.

Cependant, en cas d'évènements enregistrés à des dates identiques, la vraisemblance partielle sera similaire à celles des équations (1.40), (1.41) ou (1.42) (en fonction de la méthode de gestion des temps de survie ex-æquos choisie) avec l'ensemble à risque au temps t_i limité à tous les cas ayant connu l'évènement au temps t_i ainsi que leurs témoins appariés sur ce temps t_i et éventuellement certaines covariables.

Ainsi, l'analyse des données cas-témoins nichées est effectuée en utilisant le modèle de Cox stratifié sur les ensembles à risque qui contiennent uniquement le(s) cas et ses (leurs) témoins appariés.

Il a été établi que la vraisemblance partielle de l'analyse cas-témoins nichée (vraisemblance partielle stratifiée) coïncide avec la vraisemblance conditionnelle pour l'analyse des données cas-témoins appariés sous un modèle de régression logistique (Prentice & Breslow, 1978).

Par conséquent, le modèle de régression logistique conditionnelle est équivalent au modèle de Cox stratifié et les données des études cas-témoins nichées peuvent être analysées à l'aide du modèle de régression logistique conditionnelle, en conditionnant sur l'ensemble à risque (ou les strates correspondantes). De plus, il a été démontré que des estimations du rapport de risque obtenues avec ce modèle pouvaient être non biaisées par rapport aux estimations obtenues avec l'analyse de la cohorte entière (Breslow et al., 1978 ; Goldstein & Langholz, 1992 ; Prentice & Breslow, 1978).

Cependant, comme l'analyse des données cas-témoins nichées est effectuée avec moins de données, une perte de puissance et de précision des estimations est attendue par rapport à celles obtenues avec l'analyse de la cohorte entière. En définissant l'efficacité relative comme le rapport entre la variance du paramètre estimé à partir de la cohorte entière et la variance du paramètre estimé à partir d'une étude cas-témoins nichée, sous l'hypothèse nulle d'absence de relation exposition-réponse et dans le cas d'une covariable unique, elle vaut $m/(m+1)$, m étant le nombre de témoins par cas dans un ensemble apparié (Goldstein & Langholz, 1992). Par exemple, si un seul témoin est sélectionné pour chaque cas, l'efficacité relative est de $1/2$, ce qui implique que la variance obtenue avec l'analyse des données cas-témoins nichées est deux fois plus grande que celle obtenue à partir de l'analyse de la cohorte entière.

Ainsi, en augmentant le nombre de témoins par cas, l'efficacité relative augmente et certains travaux ont montré que des estimations issues de l'analyse cas-témoins nichée s'améliorent dans ce cas pour se rapprocher de celles de l'analyse de la cohorte entière (Bertke et al., 2013 ; Breslow & Day, 1980 ; Essebag et al., 2005 ; Goldstein & Langholz, 1992 ; Pang, 1999). Habituellement, 1 à 4 ou 5 témoins par cas sont

utilisés car un nombre de témoins supérieur à 4 ou 5 n'améliore que légèrement l'efficacité relative (Gail et al., 1976 ; Taylor, 1986). Cependant, dans certaines situations telles qu'en présence d'un nombre réduit de cas dans la cohorte ou d'une forte relation exposition-réponse, des gains en efficacité relative et une réduction du biais peuvent être réalisés en échantillonnant plus de 4 ou 5 témoins par cas (Bertke et al., 2013 ; Pang, 1999).

3) Limites du schéma d'étude

Les formules mathématiques des vraisemblances étant similaires pour les études de cohorte et cas-témoins nichée, les estimations obtenues à partir des analyses cas-témoins nichées sont non biaisées et similaires à celles obtenues à partir de l'analyse de la cohorte entière (Prentice & Breslow, 1978) : c'est un atout majeur des études cas-témoins niché.

Cependant, les principaux désavantages des études cas-témoins nichées sont les précision et puissance réduites dues à l'échantillonnage des témoins et la possibilité de défauts dans la méthode d'échantillonnage ou de sa mise en oeuvre (Wacholder, 2009). Les études cas-témoins nichées ne permettent d'estimer généralement que des mesures relatives (rapport de risques) et non des mesures absolues (risque absolu) car on ne connaît pas le nombre exact de personnes-temps à risque (l'échantillonnage ayant modifié cette information en sélectionnant un nombre fixe de témoins dans chaque ensemble à risque). De plus, l'appariement sur le temps (de survenue de l'évènement) des cas et des témoins implique que les témoins ne peuvent pas être facilement réutilisés pour étudier un autre évènement d'intérêt (Wacholder, 1991).

Ces limites ont stimulé le développement d'autres schémas d'étude.

1.3.2.2 Schéma cas-cohorte

Comme dans les études cas-témoins nichées, les études cas-cohorte utilisent un échantillon de la cohorte pour évaluer l'association entre une exposition médicamenteuse et un évènement de santé. Cependant, alors que les études cas-témoins nichées sélectionnent les témoins au temps de survenue de chaque évènement, le schéma cas-cohorte sélectionne aléatoirement des sujets dans la cohorte entière, indépendamment du temps de survenue de l'évènement (Prentice, 1986). La sous-cohorte ainsi obtenue peut être utilisée pour étudier plusieurs évènements de santé (Wacholder, 1991).

1) Principe du schéma d'étude

La caractéristique principale d'une étude cas-cohorte est la sélection d'une sous-cohorte, qui est un échantillon aléatoire de la cohorte au début de l'étude, sélectionné en ignorant toute information obtenue pendant le suivi, et qui sert d'ensemble de témoins potentiels pour tous les cas.

L'étude comprend la sous-cohorte plus tous les cas supplémentaires, c'est-à-dire ceux qui ne font pas partie de la sous-cohorte. Une illustration de la procédure d'échantillonnage est donnée par la figure 1.2 (figure à rajouter). Bien que la sélection des sujets se fasse à la date d'entrée dans la cohorte, toute la période de suivi des sujets sélectionnés est incluse dans l'étude.

L'effectif de la sous-cohorte est basé sur une certaine proportion (appelée probabilité d'échantillonnage) spécifiée à l'avance.

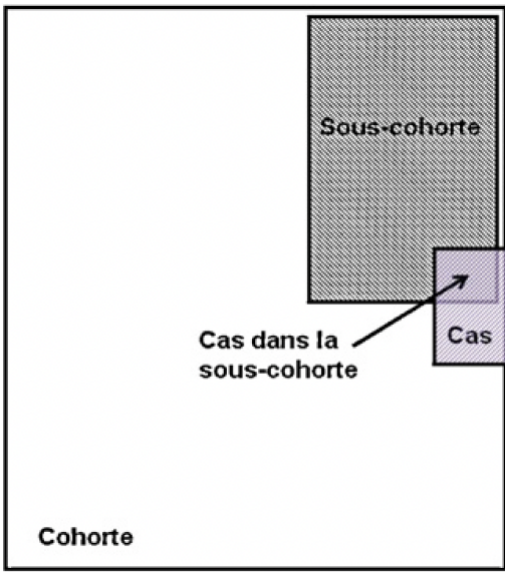


Figure 1.2 – *Échantillonnage pour le schéma cas-cohorte* (Marti & Chavance, 2013)

2) Estimation de l'effet des covariables

Le principe général de l'analyse des données cas-cohorte est la pondération des observations. En effet, les poids attribués aux sujets de l'échantillon cas–cohorte visent à reconstituer approximativement la cohorte entière à partir de ce seul échantillon.

Pour une analyse cas-cohorte standard, nous supposons le modèle à risques proportionnels de Cox pour le risque instantané de l'évènement d'intérêt défini à l'équation 2.27. L'analyse des données cas-cohorte consiste à comparer chaque cas à son temps de survenue de l'évènement avec les individus de la sous-cohorte qui sont toujours à risque à

ce temps-là en utilisant une pseudo-vraisemblance partielle (ou vraisemblance pondérée) définie par :

$$L_{CC}(\beta) = \prod_{i=1}^{J} \frac{\exp(\beta^T \mathbf{Z}_i(t)) w_i(t)}{\sum_{k \in R''(t_i)} \exp(\beta^T \mathbf{Z}_k(t)) w_k(t)} \tag{1.46}$$

avec $R''(t_i)$ représentant l'ensemble des sujets de l'échantillon cas-cohorte encore à risque au temps t_i et $w_i(t)$ le poids attribué au sujet i.

Classiquement, chaque cas a un poids $w = 1$ (car tous les cas de la cohorte sont inclus dans l'échantillon cas-cohorte) et chaque témoin a un poids $w = 1/p$ où p représente la probabilité d'échantillonnage (Lin & Ying, 1993).

L'analyse cas-cohorte permet ainsi d'estimer les rapports de risque mais aussi les risques absolus car les informations sur la population à risque sont maintenues via la probabilité d'échantillonnage de la sous-cohorte.

3) Limites du schéma d'étude

En présence d'une censure aléatoire à droite modérée ou lorsque les sujets peuvent entrer dans l'étude de cohorte à différents moments, l'échantillonnage cas-cohorte peut être beaucoup moins efficace que l'échantillonnage cas-témoins niché (Langholz & Thomas, 1990).

1.3.3 Schémas d'étude basés sur les cas seuls

Dans les études "cas seuls", contrairement aux études de cohorte, cas-témoins nichées ou cas-cohorte, seuls les cas sont nécessaires, ce qui élimine les biais éventuels de sélection des sujets témoins. De plus, dans ces études, chaque sujet cas est utilisé comme son propre témoin ; ceci permet d'annuler l'effet des facteurs de confusion invariables dans le temps. Plusieurs schémas d'étude "cas seuls" ont été proposés pour surmonter les problèmes des études de cohorte ou cas-témoins : série de cas, cas-croisé et "sequence symmetry analysis".

1.3.3.1 Série de cas

La méthode de la série de cas développée par Farrington permet d'étudier l'association entre une exposition intermittente et un évènement à survenue aiguë, rare ou potentiellement récurrent, en utilisant seulement les données des cas (Farrington, 1995). Elle est construite sur la méthodologie de la cohorte (exposition fixée, évènement aléatoire) et utilise le sujet cas comme son propre témoin (Farrington, 2004).

1) Principe et conditions d'application

Cette approche s'appuie sur une division de la période d'observation du sujet en périodes "à risque" et en périodes "témoins". Les périodes "à risque" sont des fenêtres de temps pendant ou après l'exposition, pendant lesquelles les individus sont considérés comme ayant un risque potentiellement élevé (ou réduit) de subir l'évènement d'intérêt en cas d'association. Toutes les autres fenêtres de temps de la période d'observation, c'est-à-dire avant, après ou entre les périodes "à risque", constituent les périodes "témoins" (voir l'illustration à la figure 1.3). Ces périodes sont souvent définies sur la base des connaissances a priori de l'effet de l'exposition étudiée sur la survenue de l'évènement de santé d'intérêt (Hocine & Chavance, 2010 ; Whitaker et al., 2009).

La méthode de la série de cas repose sur trois hypothèses principales qui doivent être respectées pour fournir des estimations valides et non biaisées
(Farrington, 1995 ; Whitaker et al., 2006 ; Whitaker et al., 2009).

- Les évènements éventuellement récurrents qui surviennent chez un même sujet au cours de la période d'étude sont indépendants. Cette hypothèse ne peut être vérifiée pour étudier par exemple la survenue d'épisodes successifs d'accident vasculaire cérébral (AVC) car la récurrence de l'AVC n'est pas indépendante de la première occurrence. Dans ce cas, seul le premier évènement est étudié (Petersen et al., 2016).

- L'occurrence d'un évènement ne doit pas modifier la probabilité des expositions ultérieures. Cependant, cela peut être le cas par exemple lorsqu'un évènement est une contre-indication de l'exposition ou lorsque la survenue d'un évènement peut affecter l'exposition. Une façon de corriger la violation de cette hypothèse est d'utiliser l'approche de pseudo-vraisemblance pour l'analyse des données (Farrington et al., 2008).

- L'occurrence d'un évènement ne doit pas censurer la période d'observation (les périodes d'observation doivent être indépendantes des temps d'évènement). Cette condition n'est pas vérifiée lorsque la survenue d'un évènement augmente le risque de mortalité à court terme (comme c'est le cas pour les infarctus du myocarde). Dans ce cas, on peut utiliser l'extension de la méthode de la série de cas qui supprime cette hypothèse d'indépendance en introduisant un terme supplémentaire dans la vraisemblance (Farrington et al., 2011) ou effectuer des analyses de sensibilité en excluant les cas qui sont décédés à la suite de l'évènement (Petersen et al., 2016).

2) Estimation de l'effet des covariables

On suppose que le nombre d'évènements pouvant survenir chez un sujet i pendant

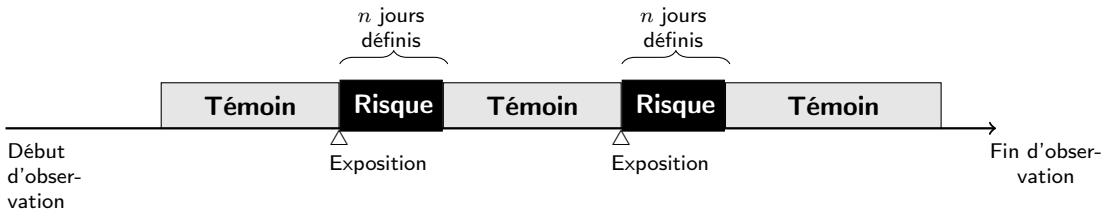


Figure 1.3 – *Périodes d'analyse pour le schéma série de cas*

sa période d'observation $(a_i, b_i]$, suit une loi de Poisson. Chaque période d'observation individuelle est divisée en groupe d'âge g. Pour une histoire d'exposition, les périodes à risque $k = 1, 2, ...$ sont définies et les périodes "témoins" sont indexées par $k = 0$.

Soient N_{igk} et e_{igk} respectivement le nombre d'évènements subis et le temps passé par l'individu i dans le groupe d'âge g et la période de risque k. Le taux d'incidence, désigné par λ_{igk}, est supposé être constant dans chaque intervalle et est donné par le modèle multiplicatif :

$$\lambda_{igk} = \exp(\phi_i + \alpha_g + \beta_k) \tag{1.47}$$

où ϕ_i est un effet pour l'individu i, α_g est un effet pour le groupe d'âge g et β_k est un effet pour la période "à risque" k. $\alpha_0 = 0$ et $\beta_0 = 0$, de sorte que l'incidence de base $\lambda_{i00} = \exp(\phi_i)$.

En conditionnant sur le nombre d'évènements $N_i = \sum_{g,k} N_{igk}$ survenus pour l'individu i pendant sa période d'observation, la vraisemblance conditionnelle résultante est multinomiale et est définie par (Whitaker et al., 2006 ; Whitaker et al., 2009) :

$$L_i(\alpha, \beta) = \prod_{g,k} \left\{ \frac{e_{igk} \exp(\alpha_g + \beta_k)}{\sum_{r,s} e_{irs} \exp(\alpha_r + \beta_s)} \right\}^{N_{igk}} \tag{1.48}$$

et la vraisemblance de la méthode de la série de cas pour J cas est :

$$L_{SC}(\alpha, \beta) = \prod_{i=1}^{J} \prod_{g,k} \left\{ \frac{e_{igk} \exp(\alpha_g + \beta_k)}{\sum_{r,s} e_{irs} \exp(\alpha_r + \beta_s)} \right\}^{N_{igk}} \tag{1.49}$$

Ainsi, les effets individuels ϕ_i s'annulent et c'est pour cela que la méthode est dite auto-contrôlée. Ces effets individuels comprennent tous les facteurs de confusion fixes dans le temps. Seul l'âge ou d'autres covariables dépendantes du temps doivent être inclus

dans le modèle, bien qu'il soit possible d'inclure les interactions entre l'exposition et les covariables fixes dans le temps.

3) Limites du schéma d'étude

Ce schéma d'étude n'est pas facilement adaptable à une proportion significative de traitements chroniques, tels que ceux qui nécessitent une administration à vie sans changement ni interruption, comme l'insuline sous-cutanée pour le diabète de type 1 (Grosso et al., 2011). De plus, il nécessite une variabilité dans le temps (ou en fonction de l'âge) de l'évènement (Whitaker et al., 2006). Enfin, le schéma d'étude de la série de cas ne peut fournir qu'une estimation du risque relatif, et non du risque absolu sans apports externes.

1.3.3.2 Schéma cas-croisé (*case-crossover*)

Basée sur les données des cas uniquement, la méthode cas-croisé a été développée par Maclure pour évaluer les effets transitoires d'une brève exposition sur le risque de survenue d'un évènement aigu (Maclure, 1991). Elle utilise chaque sujet comme son propre témoin dans une analyse similaire à celle d'une étude cas-témoins (Farrington, 2004).

1) Principe du schéma d'étude

Cette approche compare pour le même sujet cas, l'exposition étudiée dans une période précédant immédiatement la survenue de l'évènement (période "à risque" ou période "cas") avec celle d'une ou plusieurs périodes antérieures de même durée (période "témoin"). La période "à risque" et les périodes "témoins" peuvent éventuellement être séparées par une période de *wash-out*. Ces périodes d'analyse sont illustrées par la figure 1.4.

L'étude cas-croisé repose sur l'hypothèse selon laquelle, si un individu est exposé dans la période "à risque", alors il aura un risque augmenté (ou réduit) de survenue de l'évènement d'intérêt au moment observé s'il existe une association entre l'exposition et l'évènement. La définition de cette période nécessite une idée raisonnable du temps d'induction entre l'exposition étudiée et l'apparition de l'évènement d'intérêt, mais aussi du délai avant que le risque ne diminue à nouveau pour atteindre le risque de base sous-jacent (Maclure, 1991 ; Maclure & Mittleman, 2000). L'une des hypothèses principales de ce schéma en ce qui concerne particulièrement la pharmaco-épidémiologie est que la probabilité d'exposition à un traitement ou produit de santé doit être fixe au

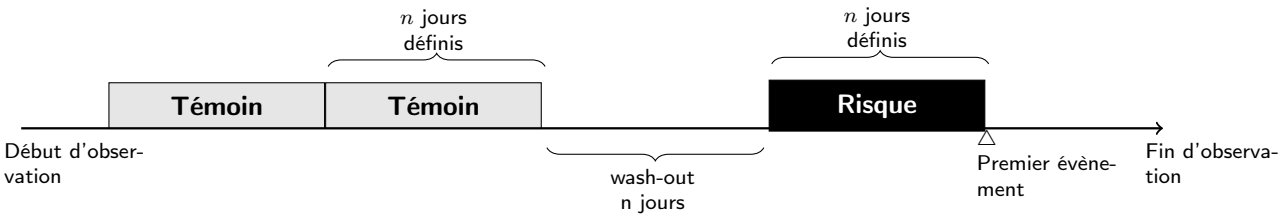


Figure 1.4 – *Périodes d'analyse pour le schéma cas-croisé*

cours du temps (Delaney & Suissa, 2009).

Ce schéma d'étude est particulièrement bien adapté pour les maladies aiguës à début bien défini, par exemple les accidents et les infarctus du myocarde. En revanche, les maladies à début insidieux comme le cancer se prêtent moins bien à une telle analyse.

2) Estimation de l'effet des covariables

Les données des études cas-croisé sont généralement analysées à l'aide d'un modèle de régression logistique conditionnelle ou d'un modèle de Cox stratifié équivalent, qui tient compte de la nature appariée des données pour l'estimation du risque (Mittleman et al., 1995).

Ainsi la vraisemblance logistique conditionnelle est donnée par :

$$L_{CCO}(\beta) = \prod_{i=1}^{J} \frac{\exp(\beta z_{i0})}{\sum_{m=0}^{M} \exp(\beta z_{im})} \tag{1.50}$$

où J est le nombre de sujets cas, M le nombre de périodes "témoins", z_{i0} l'exposition du cas i dans la période "à risque" et z_{im} l'exposition du cas i dans la période "à risque" pour $m = 0$ et dans la période "témoin" pour $m = 1, ..., M$.

Comme dans les études cas-témoins appariés où seules les paires discordantes contribuent à l'estimation de l'effet, dans les études cas-croisé, seuls les cas pour lesquels une variation de l'exposition entre la période "à risque" et la (les) périodes témoin(s) a été observée contribuent à l'estimation de l'effet (Suissa, 1995). Par conséquent, l'utilisation des études cas-croisé serait efficace pour évaluer l'effet d'une exposition qui varie dans le temps comme c'est souvent le cas pour une exposition médicamenteuse.

3) Limites du schéma d'étude

Bien qu'un biais éventuel de sélection des sujets témoins est impossible car chaque cas

est utilisé comme son propre témoin, il est toujours possible d'avoir un éventuel biais de sélection des périodes "témoins". Un biais de tendance temporelle peut se produire si les périodes "cas" et "témoins" sont très longues car il peut y avoir eu un changement dans les habitudes de prescription du médicament d'intérêt. Ceci entraînant une différence systématique de la prévalence de l'exposition entre la période "témoins" et la période "cas" (SCHNEEWEISS et al., 1997). Cette dernière limite a été levée par des extensions du schéma cas-croisé.

1.3.3.3 Extensions du schéma cas-croisé

1) Schéma *case-time-control*

Pour éliminer l'éventuel biais de tendance temporelle dans les estimations issues du schéma d'étude cas-croisé, Suissa (1995) a développé la méthode *case-time-control* qui prend en compte les variations de la prévalence de l'exposition au cours du temps (DONNAN & WANG, 2001 ; SCHNEEWEISS et al., 1997 ; SUISSA, 1995). Cette approche utilise un groupe de témoins et applique les analyses cas-croisé à la fois à ces témoins (période à risque définie au temps d'évènement du cas apparié) et aux cas. Le risque estimé chez les témoins est utilisé pour ajuster celui estimé chez les cas (le risque estimé chez les cas est divisé par celui estimé chez les témoins, pour éliminer l'effet de la tendance de l'exposition dont la probabilité varie au cours du temps). Cependant, s'il n'est pas bien apparié, le groupe de témoins peut réintroduire un biais de sélection.

2) Schéma *case-case-time-control*

Pour pallier cette limite du schéma d'étude *case-time-control*, le schéma *case-case-time-control* a été proposé par Wang (2011). Cette approche applique des analyses cas-croisé aux cas et à un groupe de témoins qui sont des futurs cas. Cette utilisation des futurs cas comme témoins pour les cas actuels permet de réduire, voire d'éliminer le biais provenant des tendances de l'exposition dont la prévalence varie au cours du temps (WANG et al., 2011). Le schéma d'étude *case-case-time-control* minimise en outre le risque d'introduire un biais de sélection en limitant les témoins aux futurs cas.

1.3.3.4 Analyse en symétrie de séquences

L'analyse en symétrie de séquences (SSA pour *Sequence symmetry analysis*) est un schéma d'étude basé sur les données des cas uniquement et reposant sur l'hypothèse selon laquelle, si un médicament est responsable d'un évènement indésirable, alors ce médicament sera prescrit plus souvent avant qu'après la survenue de cet évènement

(Hallas, 1996). Elle est utilisée à la fois comme méthode d'étude des effets indésirables spécifiques à l'utilisation des médicaments et comme outil d'exploration des bases de données de remboursement pour détecter des évènements indésirables inconnus et insoupçonnés liés aux médicaments (Tsiropoulos et al., 2009 ; Wahab et al., 2016).

1) Principe de l'étude

L'analyse de symétrie des séquences est basée sur l'examen des séquences d'évènements en relation avec l'initiation d'un traitement médicamenteux. En effet, en utilisant les données de prescription ou de remboursement disponibles sur une certaine période de temps, cette méthode vise à évaluer de potentielles associations entre un évènement indésirable (EI) d'intérêt et un traitement médicamenteux en dénombrant le nombre de séquences « initiation d'un traitement (E) puis EI d'intérêt (Y) : E $\rightarrow$ Y) » et « EI d'intérêt (Y) puis initiation d'un traitement (E) : Y $\rightarrow$ E ». L'EI pouvant être la prescription d'un autre traitement.

Sous l'hypothèse que le traitement médicamenteux ne déclenche pas l'évènement, la proportion d'observations pour lesquelles l'initiation d'un traitement précède l'EI d'intérêt devrait être similaire à la proportion d'observations pour lesquelles l'EI d'intérêt précède l'initiation d'un traitement. L'asymétrie de la distribution de cet EI d'intérêt avant et après l'initiation d'un traitement est donc utilisée pour évaluer l'association entre ce traitement et cet EI d'intérêt (Hallas, 1996).

2) Estimation de l'effet des covariables

La statistique d'intérêt pour la SSA est le rapport de séquences (*sequence ratio*, SR) qui est une mesure de l'asymétrie des séquences. Un rapport de séquences brut (*crude sequence ratio*, CSR) est d'abord calculé en divisant le nombre d'individus ayant la séquence « E $\rightarrow$ Y » par le nombre d'individus concernés par la séquence « Y $\rightarrow$ E » sur une fenêtre d'observation considérée. Ensuite, pour obtenir un rapport de séquences ajusté (*adjusted sequence ratio*, ASR, mesure d'intérêt pour la SSA), le rapport de séquences brut est divisé par un *null-effect sequence ratio* (NSR) afin de prendre en compte les possibles variations au cours du temps des prévalences à la fois de la survenue de l'évènement indésirable d'intérêt et de la prescription du traitement. Enfin, un IC à 95 % du rapport de séquences ajusté est déterminé (Hallas, 1996).

3) Limites du schéma d'étude

La SSA peut être affectée par des biais dus à des facteurs de confusion intra-personnels (qui varient au cours du temps) tels que les changements de comportement qui peuvent influencer l'ordre de prescription des traitements de référence et des traitements pour soigner l'évènement indésirable que l'on suspecte être induit par ces traitements de référence. De plus, il n'existe pas de méthode de calcul formelle pour ajuster les facteurs de confusion variant au cours du temps dans la SSA, contrairement aux analyses cas-croisé et de série de cas qui permettent d'ajuster les covariables dépendantes du temps, par exemple en les incorporant dans un modèle de régression logistique conditionnelle ou de poisson (Lai et al., 2017).

1.4 Algorithmes de génération de données de survie

Le modèle de Cox à risques proportionnels est couramment utilisé pour évaluer l'effet d'une covariable fixe ou dépendante du temps sur le délai de survenue d'un évènement en présence d'une censure à droite, qu'on retrouve très souvent dans les données de survie ou de cohorte en particulier. Des études de simulation sont de plus en plus souvent effectuées pour évaluer ses performances et ses propriétés statistiques dans diverses situations pré-définies.

Plusieurs approches et algorithmes permettant de générer des données de survie censurées à droite pour simuler le modèle de Cox à risques proportionnels avec des covariables fixes ou dépendantes du temps ont été proposés.

1.4.1 Approche de Bender et al.

Les travaux de Leemis (1987) ont montré que les temps de survie pouvaient être générés à partir de distributions connues en inversant la fonction de risque cumulé (Leemis, 1987). Bender et al. (2005) s'en sont inspirés pour présenter une formule générale décrivant la relation entre la fonction de risque et le temps de survie correspondant du modèle de Cox. Par leur approche, ils montrent comment avec des coefficients de régression connus et n'importe quelle fonction de risque de base non nulle, les temps de survie peuvent être générés pour simuler des modèles de Cox à risques proportionnels avec des covariables fixes dans le temps (Bender et al., 2005). En effet, en utilisant la formule suivante de la fonction de répartition du modèle de Cox (1.2.7)

$$F(t|\mathbf{Z}) = 1 - \exp[-H_0(t)\exp(\beta^T\mathbf{Z})], \tag{1.51}$$

où $H_0(t) = \int_0^t h_0(u)du$ est la fonction de risque cumulé de base, ils considèrent une variable aléatoire Y avec une fonction de répartition F. Alors, $U = F(Y) \sim \mathcal{U}[0;1]$ et $(1-U) \sim \mathcal{U}[0;1]$ également. Ainsi, avec T la variable aléatoire continue positive représentant le temps de survie et la formule 1.51, ils déduisent que :

$$U = \exp[-H_0(t)\exp(\beta^T \mathbf{Z})] \sim \mathcal{U}[0;1]. \tag{1.52}$$

Si $h_0(t) > 0$ pour tout t, alors H_0 peut être invsersée et le temps de survie T du modèle de Cox peut être exprimé par :

$$T = H_0^{-1}[-\log(U)\exp(-\beta^T \mathbf{Z})] \tag{1.53}$$

où H_0^{-1} est la fonction inverse de la fonction de risque cumulé de base H_0.

Ainsi, en appliquant la formule (1.53) après y avoir inséré l'inverse d'une fonction de risque cumulé de base appropriée, les nombres aléatoires uniformément distribués peuvent être transformés en temps de survie selon un modèle de Cox spécifique.

Parmi les distributions paramétriques connues, seules les distributions exponentielle, de Weibull et de Gompertz vérifient l'hypothèse des risques proportionnels avec le modèle de Cox.

En utilisant la formule (1.53), on peut ainsi générer des temps de survie pour les modèles de Cox avec des temps de survie distribués selon ces trois lois.

En effet, pour une distribution exponentielle de paramètre $\lambda > 0$, on a un modèle Cox-exponentiel défini par la fonction de risque

$$h(t|\mathbf{Z}) = \lambda \exp(\beta^T \mathbf{Z})$$

et les temps de survie sont obtenus par la formule

$$T = \frac{-\log(U)}{\lambda \exp(\beta^T \mathbf{Z})}.$$

Pour une distribution de Weibull de paramètres d'échelle $\lambda > 0$ et de forme $\gamma > 0$, on a un modèle Cox-Weibull défini par la fonction de risque

$$h(t|\mathbf{Z}) = \lambda\gamma t^{\gamma-1} \exp(\beta^T \mathbf{Z})$$

et les temps de survie sont obtenus par la formule

$$T = [\frac{-\log(U)}{\lambda \exp(\beta^T \mathbf{Z})}]^{1/\gamma}.$$

Et pour une distribution de Gompertz de paramètres d'échelle $\lambda > 0$ et de forme $\alpha \in (-\infty, \infty)$, on a un modèle Cox-Gompertz défini par la fonction de risque

$$h(t|\mathbf{Z}) = \lambda \exp(\alpha t) \exp(\beta^T \mathbf{Z})$$

et les temps de survie sont obtenus par la formule

$$T = \frac{1}{\alpha} \log[1 - \frac{\alpha \log(U)}{\lambda \exp(\beta^T \mathbf{Z})}].$$

Bender et al. (2005) ont ainsi proposé des solutions analytiques (*closed form*) pour générer des données de survie appropriées pour un modèle de Cox à risques proportionnels avec des covariables fixes dans le temps en utilisant les distributions exponentielle, de Weibull et de Gompertz. Cette méthode repose sur l'utilisation de la fonction inverse de la fonction de risque cumulé (Bender et al., 2005).

Sylvestre et Abrahamowicz (2008) ont pensé que l'inversion de la fonction de risque cumulé utilisée par Bender et al. (2005) s'avérerait difficile à étendre aux situations où les fonctions paramétriques ne peuvent pas être utilisées pour décrire le changement dans le temps ou lorsque les covariables ne sont pas définies sur toute la plage de temps.

Par conséquent, ils ont évalué l'extension de l'algorithme de permutation introduit par Abrahamowicz et al. (1996) et décrit en détail par MacKenzie et Abrahamowicz (2002) pour les covariables dépendantes du temps. Cet algorithme ne nécessite pas l'inversion de la fonction de risque cumulé et permet de prendre en compte un nombre quelconque de covariables fixes et variables dans le temps (Abrahamowicz et al., 1996 ; Mackenzie & Abrahamowicz, 2002 ; Sylvestre & Abrahamowicz, 2008).

1.4.2 Algorithme de permutation

L'extension de l'algorithme de permutation (Mackenzie & Abrahamowicz, 2002) pour générer des temps de survie conditionnés par des covariables dépendantes du temps nécessite les cinq étapes suivantes (Sylvestre & Abrahamowicz, 2008) :

(i) Génération des temps de survie $T_i, i = 1, .., n$ suivant une loi spécifiée au préalable, telle qu'exponentielle ou Weibull. La loi est supposée représenter la distribution des temps de survie dans la population entière de l'étude, indépendamment

des covariables.

(ii) Génération des temps de censure $C_i, i = 1, .., n$ suivant une loi spécifiée au préalable.

(iii) Classement des temps d'observation $X_i = min(T_i, C_i)$ par ordre croissant.

(iv) Génération des matrices individuelles des valeurs des covariables $\mathbf{Z}_s(t)$, $s = 1, .., n$. Chaque matrice comporte m lignes, chacune représentant un intervalle de temps de suivi et p colonnes correspondant à des covariables fixes et dépendantes du temps.

(v) Appariement un à un des temps d'observation (rangés par ordre croissant) avec les vecteurs de covariables individuels sélectionnés aléatoirement :

Si $X_i = T_i$ (temps de survie), un sujet s est aléatoirement choisi sur la base d'une probabilité dérivée de la vraisemblance partielle du modèle de Cox à risques proportionnels (1.37) :

$$p_{s,X_i} = \frac{\exp[\beta^T \mathbf{Z}_s(X_i)]}{\sum_{k \in R(X_i)} \exp[\beta^T \mathbf{Z}_k(X_i)]} \tag{1.54}$$

Si $X_i = C_i$ (temps de censure), un sujet censuré est sélectionné aléatoirement dans l'ensemble $R(X_i)$ (ensemble des individus qui n'ont pas encore été sélectionnés au temps x_i) avec une probabilité identique pour tous les sujets de $R(X_i)$: $P = 1/card(R(X_i))$.

Cette méthode permet de prendre en compte un nombre quelconque de covariables dépendantes du temps sans qu'il soit nécessaire de spécifier une forme fonctionnelle pour leur variation dans le temps ou d'inverser la fonction de risque cumulé. Cependant, l'approche proposée pour générer les temps de survie n'a pas de solution à "forme fermée".

Austin (2012) s'est inspiré de l'approche de Bender et al. (2005) pour montrer comment la technique d'inversion de la fonction de risque cumulé peut être étendue pour générer des temps de survie conditionnés par des covariables dépendantes du temps qui vérifient l'hypothèse des risques proportionnels du modèle de Cox (Austin, 2012).

1.4.3 Approche d'Austin

Pour étendre les résultats de Bender et al. (2005) pour la simulation des temps de survie (à partir des trois distributions paramétriques couramment utilisées) dans les situations où les covariables varient au cours du temps, Austin (2012) a considéré trois différents types de covariables dépendantes du temps : une covariable dichotomique

variant dans le temps représentée par le statut d'exposition qui peut changer au plus une fois de "non exposé" à "exposé" ; une covariable continue variant dans le temps telle que l'exposition cumulée à une dose fixe d'un médicament et une covariable dichotomique variant dans le temps telle que le statut d'exposition qui peut changer plusieurs fois de "non exposé" à "exposé" et de nouveau à "non exposé" (AUSTIN, 2012). Nous ne présenterons que les formules de génération des temps de survie pour une covariable dichotomique dépendante du temps à changement unique.

En considérant une seule covariable dépendante du temps $Z(t)$ et les autres covariables $\mathbf{X}$ fixes dans le temps, Austin (2012) utilise un modèle de Cox défini par :

$$h(t|(\mathbf{X}, Z(t))) = h_0(t)\exp(\beta_z Z(t) + \beta_x^T \mathbf{X})$$

où β_z et β_x^T représentent respectivement le coefficient de régression associé à $Z(t)$ et le vecteur des coefficients de régression associés au vecteur des covariables fixes $\mathbf{X}$. Il en déduit la fonction de risque cumulé :

$$H(t, \mathbf{X}, Z(t)) = \int_0^t h_0(u)\exp(\beta_z Z(u) + \beta_x^T \mathbf{X})du$$

Soit t_0 l'instant où le statut d'exposition représentant la covariable dépendante du temps passe de "non exposé" à "exposé". Pour $t < t_0$, $Z(t) = 0$ tandis que pour $t \geq t_0$, $Z(t) = 1$.

Après avoir déterminé les expressions de la fonction de risque cumulé et de son inverse selon les trois distributions paramétriques couramment utilisées sur chacun des intervalles $(0, t_0)$ et $[t_0, \infty)$ (AUSTIN, 2012) et pour $u \sim \mathcal{U}[0; 1]$, il peut appliquer la formule (1.53) pour générer des temps de survie conditionnés par une covariable dépendante du temps et des covariables fixes.

Ainsi, pour une distribution exponentielle de paramètre $\lambda > 0$,

$$T = \begin{cases} \frac{-\log(u)}{\lambda\exp(\beta_x^T\mathbf{X})} & \text{si } -\log(u) < \lambda\exp(\beta_x^T\mathbf{X})t_0 \\ \\ \frac{-\log(u)-\lambda\exp(\beta_x^T\mathbf{X})t_0+\lambda\exp(\beta_x^T\mathbf{X}+\beta_z)t_0}{\lambda\exp(\beta_x^T\mathbf{X}+\beta_z)} & \text{si } -\log(u) \geq \lambda\exp(\beta_x^T\mathbf{X})t_0 \end{cases}$$

Pour une distribution de Weibull de paramètres d'échelle $\lambda > 0$ et de forme $\gamma > 0$,

$$T = \begin{cases} [\frac{-\log(u)}{\lambda \exp(\beta_x^T \mathbf{X})}]^{1/\gamma} & \text{si } -\log(u) < \lambda \exp(\beta_x^T \mathbf{X}) t_0^\gamma \\ [\frac{-\log(u) - \lambda \exp(\beta_x^T \mathbf{X}) t_0^\gamma + \lambda \exp(\beta_x^T \mathbf{X} + \beta_z) t_0^\gamma}{\lambda \exp(\beta_x^T \mathbf{X} + \beta_z)}]^{1/\gamma} & \text{si } -\log(u) \geq \lambda \exp(\beta_x^T \mathbf{X}) t_0^\gamma \end{cases}$$

Et pour une distribution de Gompertz de paramètres d'échelle $\lambda > 0$ et de forme $\alpha \in (-\infty, \infty)$,

$$T = \begin{cases} \frac{1}{\alpha} \log[1 + \frac{\alpha(-\log(u))}{\lambda \exp(\beta_x^T \mathbf{X})}] & \text{si } -\log(u) < \frac{\lambda \exp(\beta_x^T \mathbf{X})}{\alpha} [\exp(\alpha t_0) - 1] \\ \frac{1}{\alpha} \log[\frac{\alpha(-\log(u))}{\lambda \exp(\beta_x^T \mathbf{X} + \beta_z)} - \frac{\exp(\alpha t_0) - 1 - \exp(\beta_z + \alpha t_0)}{\exp(\beta_z)}] & \text{si } -\log(u) \geq \frac{\lambda \exp(\beta_x^T \mathbf{X})}{\alpha} [\exp(\alpha t_0) - 1] \end{cases}$$

Austin (2012) a fourni ainsi des solutions analytiques (*closed form*) pour générer des temps de survie conditionnés par exactement une covariable dépendante du temps à partir de distributions exponentielle, de Weibull et de Gompertz. Toutefois, il n'existe pas de solution à "forme fermée" pour une distribution de Weibull avec une covariable continue variant dans le temps (Austin, 2013). Les approches développées par Austin (2012) sont limitées par la possibilité de ne prendre en compte qu'une seule covariable dépendante du temps.

Par ailleurs, Crowther et Lambert (2013) ont étendu la méthode de Bender et al. (2005) pour permettre des études de simulation plus complexes où soit l'intégrale pour obtenir la fonction de risque cumulé $H_0(t)$ est insoluble, soit $H_0(t)$ est non inversible.

1.4.4 Algorithme de Crowther et Lambert

Dans les approches de Bender et al. (2005) et Austin (2012), la génération des temps de survie à partir de la formule (1.53) repose sur la capacité d'inverser la fonction de risque cumulé, elle même devant être obtenue en intégrant la fonction de risque. Pour relever ces défis, Crowther et Lambert (2013) ont développé un algorithme combiné impliquant des techniques de recherche de racines et d'intégration numérique afin de fournir une méthode efficace de génération des temps de survie à partir d'une variété de distributions paramétriques complexes, incorporant toute combinaison d'effets dépendants du temps, de covariables variant dans le temps, d'entrée retardée, d'effets aléatoires et de covariables mesurées avec erreur (Crowther & Lambert, 2013).

En effet, si l'on souhaite définir une fonction de risque complexe qui ne peut pas être intégrée analytiquement pour obtenir la fonction de risque cumulé, alors on peut utiliser

des techniques d'intégration numérique telles que la quadrature de Gauss-Legendre pour avoir une approximation de l'intégrale. Ensuite, on a une fonction de risque cumulé qui ne peut pas être inversée directement pour obtenir les temps de survie, ce qui nécessite des techniques de recherche de racines. Il en résulte un algorithme général en deux étapes impliquant une intégration numérique imbriquée dans une procédure itérative de recherche de racines (Crowther & Lambert, 2013).

Bien qu'étendue, cette approche repose sur l'intégration numérique et peut être coûteuse en temps de calcul si le nombre de covariables est important (Montez-Rath et al., 2017).

Hendry (2014) a développé un algorithme général qui met en œuvre la méthode de Zhou (2001) qui fournit une solution analytique (*closed form*) pour générer des temps de survie censurés à droite pour un modèle de Cox avec un nombre quelconque de covariables à la fois fixes et variables dans le temps qui changent à des étapes à valeur entière de l'échelle de temps. (Hendry, 2014 ; Zhou, 2001).

1.4.5 Algorithme de Hendry

Zhou (2001) a développé une procédure pour générer des temps de survie dans le cadre d'un modèle de Cox avec des covariables dépendantes du temps qui utilise une transformation des variables aléatoires exponentielles par morceaux (Zhou, 2001). Hendry (2014) s'en est inspiré pour démontrer que les variables aléatoires exponentielles par morceaux avec un support $[a; b]$ telles que $0 < a < b$ (variables aléatoires exponentielles tronquées par morceaux), peuvent être générées par un algorithme d'acceptation-rejet où les réalisations en dehors du support sont écartées et celles qui sont à l'intérieur sont incluses. L'algorithme proposé par Hendry peut être résumé par les étapes suivantes (Hendry, 2014) :

(i) Définition d'une fonction g croissante monotone telle que $g(0) = 0$ et $g^{-1}(t)$ est différentiable. Elle doit être définie de telle sorte que $g^{-1}(t) = H_0(t)$, la fonction de risque de base cumulé.

(ii) Définition d'une valeur maximale pour l'échelle de temps $t \in \mathbb{N}$;

(iii) Définition d'une partition finie de l'échelle de temps $\{s_1, ..., s_J\} \subset \mathbb{N} : s_J \leq t$.

(iv) Définition des bornes de troncature $a, b \in \{s_1, ..., s_J\} : a < b < t$; où a la borne inférieure peut être par exemple la durée minimale à passer dans l'étude pour être éligible et b la borne supérieure correspond à la durée maximale admissible observée pour un individu.

(v) Génération des valeurs des covariables $\{\mathbf{Z}_{ij}\}_{j=1}^{t}$ pour chaque individu $i = 1, .., n$ et pour chaque morceau (intervalle) de l'échelle de temps partitionné.

(vi) Calcul des taux $\{\lambda_{ij}\}_{j=1}^{t} = \{\exp(\beta^T \mathbf{Z}_{ij})\}_{j=1}^{t}$.

(vii) Génération des variables aléatoires X_i exponentielles tronquées par morceaux de paramètres $\{\lambda_{ij}\}_{j=1}^{t}$ et calcul des temps de survie $T_i = g(X_i)$.

(viii) Définition de l'indicatrice de censure $\{\delta_i\}_{i=1}^{n}$ avec $\delta_i \in \{0, 1\}$.

Montez-Rath et al. (2017) fournissent des directives et des recommandations pour utiliser cet algorithme (Montez-Rath et al., 2017).

1.5 Revue des travaux comparant différents schémas d'étude

Plusieurs travaux ont comparé les différents schémas d'étude présentés précédemment, soit à partir des données réelles, soit en évaluant leurs performances et propriétés statistiques à l'aide de données simulées. Nous présentons quelques travaux (ayant comparé au moins deux de ces schémas d'étude) identifiés sur *Pubmed* et *Google Scholar* en utilisant des combinaisons des termes suivants : *cohort*, *nested case-control*, *case-crossover*, *self controlled case series*, *simulations*, *case-time-control*, *case-case-time-control*, *Sequence symmetry analysis*, *comparison*, *survival analysis*, *pharmacoepidemiology*, *time-dependent exposures*, *case-cohort*.

1.5.1 Travaux de simulation

Les travaux (tableau A.1 dans l'annexe A) sur les comparaisons des schémas d'étude à partir de données simulées s'intéressent aux expositions fixes et variables dans le temps, avec un seul ou plusieurs changements.

Dans ces travaux, nous avons constaté que des biais négatifs ou positifs ont été observés dans les estimations des schémas cas-témoins nichés (CTN) et cas-cohorte tandis qu'il n'y avait pas de biais dans celles de la cohorte entière (Austin et al., 2012 ; Bertke et al., 2013 ; Kim, 2015 ; Leffondré et al., 2003 ; Leffondré et al., 2010). Par exemple, pour une exposition dépendante du temps, le biais maximum était de 22 % pour l'analyse CTN tandis qu'il était de 3,3 % pour celle de la cohorte entière dans l'étude de Leffondré (2003). Cependant, ces biais pouvaient être réduits en augmentant :

- Le nombre de témoins par cas (Bertke et al., 2013 ; Kim, 2015) : le biais passait de 4 % pour un témoin par cas à environ 1 % pour 5 témoins par cas avec un $\log(HR) = 0,7$ et une exposition fixe (Kim, 2015) ;

- La taille de l'échantillon : pour un témoin par cas, le biais passait de 4 % à 2 % pour respectivement 500 et 1000 sujets dans la cohorte avec un $\log(HR) = 0{,}7$ et une exposition fixe (Kim, 2015) ;

- La proportion d'évènements : pour un témoin par cas, le biais passait de 26 % à 6 % pour respectivement 0,6 % et 2 % d'évènements, avec une exposition fixe (Bertke

et al., 2013). A contrario, dans l'une de ces études, il a été observé une augmentation du biais lorsque la proportion d'évènements augmentait, passant d'environ 9 % à 45 % pour 5 % et 25 % d'évènements respectivement, avec une exposition dépendante du temps et $HR = 1,25$ (Austin et al., 2012).

Les travaux ayant comparé les schémas de cohorte, cas-croisé, *case-time-control* et *case-case-time-control* ont montré que les estimations des schémas cas-croisé pouvaient être biaisées, en particulier pour de petites tailles d'échantillon et une faible proportion d'évènements (Brickner, 2015) et pour des expositions à long terme : biais absolu allant de 0,10 à 2,32 respectivement pour 10 % et 90 % de sujets restant exposés sur toute la période du suivi (Bykov et al., 2020). Ce biais dû aux expositions persistantes pouvait être réduit ou éliminé en utilisant le schéma *case-time-control* (Bykov et al., 2020). Cependant, l'utilisation d'un ensemble de témoins inapproprié peut laisser des biais dans les estimations des schémas *case-time-control* lorsque la prévalence de l'exposition varie au cours du temps. Dans ce cas, le schéma *case-case-time-control* permet d'ajuster ces tendances de l'exposition et d'obtenir des estimations non biaisées (Wang et al., 2011).

Un schéma d'étude autocontrôlé doit être utilisé de façon appropriée dans des conditions où ses hypothèses principales sont satisfaites afin d'éviter un éventuel biais dans les estimations résultantes. En effet, Takeuchi (2018) a comparé les performances statistiques des schémas d'étude cas-croisé, série de cas, et l'analyse en symétrie de séquences en considérant plusieurs hypothèses (la présence de tendances temporelles dans les expositions ou l'occurrence d'évènements, la censure des observations des patients basée sur la survenue de l'évènement, l'exclusion des patients qui ont connu le premier évènement avant leur première exposition et la spécification inadéquate de la durée de la période à risque) et a obtenu les résultats suivants (Takeuchi et al., 2018) :

- Aucun biais substantiel pour les estimations de tous les schémas considérés en présence des tendances temporelles de l'exposition ou de l'évènement ;

- Pas de biais substantiel, mais faible taux de recouvrement pour les estimations de la série de cas, après restriction de la période de suivi basée sur l'occurrence d'un évènement ;

- Biais substantiel et faible taux de recouvrement pour les estimations de la série de cas, après restriction des patients basée sur la survenue d'un évènement avant la première exposition ;

- Biais sévère pour les estimations du cas-croisé pour une période à risque erronément courte ;

- Biais sévère pour les estimations des 3 schémas d'étude pour une période à risque erronément longue.

1.5.2 Travaux à partir de données réelles

Ces travaux (tableau A.2 dans l'annexe A) sur les comparaisons des schémas d'étude à partir des données réelles s'intéressent aux expositions fixes et variables dans le temps, avec un seul ou plusieurs changements.

Contrairement aux résultats différents observés dans les études comparant les schémas de cohorte et cas-témoins niché à partir des données simulées, la majorité des études effectuées à partir de données réelles ont montré que les estimations obtenues par le schéma cas-témoins niché étaient similaires (mais moins précises) à celles obtenues par l'analyse de la cohorte entière, en particulier à partir de 4 témoins par cas (Billioti de Gage et al., 2012 ; Essebag et al., 2005 ; Hak et al., 2004). En effet, pour une exposition variable dans le temps, Essebag et al. (2005) ont montré qu'en utilisant n'importe quel nombre de 4 à 64 témoins par cas, la moyenne des HR (des estimations issues de 100 analyses CTN répétant l'échantillonnage aléatoire des témoins) était très similaire (HR moyen compris entre 2,02 et 2,19 pour les différents nombres de témoins par cas) à l'HR obtenu en utilisant la régression de Cox sur la cohorte entière (HR = 2,03). De plus, la précision des estimations des analyses CTN s'améliorait avec l'augmentation du nombre de témoins par cas (Essebag et al., 2005).

Des résultats presque semblables issus des schémas de cohorte, cas-témoins niché et série de cas ont été observés dans une étude analysant les effets indésirables des vaccins (Farrington et al., 1996) et une autre évaluant l'association entre les antidépresseurs et la fracture du col du fémur/hanche (de Groot et al., 2016). Cependant, les estimations de l'analyse du schéma cas-témoins niché étaient moins précises que celles de l'analyse du schéma de la série de cas (Farrington et al., 1996).

Les travaux ayant comparé les schémas cas-croisé et *case-time-control* ont montré que les résultats obtenus avec le schéma *case-time-control* étaient meilleurs que ceux obtenus avec le schéma cas-croisé. En effet, le schéma *case-time-control* peut mettre en exergue des associations entre une exposition médicamenteuse et un effet indésirable qui seraient passées inaperçues avec une analyse cas-croisé. C'est le cas par exemple dans une étude de Hernandez-Diaz et al. sur l'épidémiologie des maladies congénitales (Hernandez-Diaz, 2003) qui montre que l'analyse du schéma *case-time-control* faisait ressortir une association entre l'utilisation d'antagonistes de l'acide folique pendant les deuxième et troisième trimestres de la grossesse et le risque de malformations cardiovasculaires, tandis que l'analyse du schéma cas-croisé ne révélait aucune association. De plus, Nicholas et al. (2012) ont montré dans leur travail que le biais pouvant être observé dans les estimations issues de l'analyse du schéma cas-croisé peut être réduit en utilisant le schéma *case-time-control* (Nicholas et al., 2012). Une telle correction observée est conforme aux attentes de ce schéma (Suissa, 1995).

Malgré leurs avantages considérables, les schémas d'étude auto-contrôlés (cas-croisé, série de cas et *case-time-control*) peuvent fournir de moins bons résultats par rapport aux schémas classiques dans certaines situations. Par exemple, une étude de Nicholas et al. (2012) a montré que les résultats des estimations issues de ces schémas d'étude auto-contrôlés pouvaient être moins précis et davantage susceptibles d'être surévalués que ceux des schémas d'étude de cohorte et cas-témoins niché, lorsque les expositions à long terme étaient analysées (Nicholas et al., 2012).

1.6 Objectifs de la thèse

L'objectif général de cette thèse est d'évaluer les méthodes statistiques disponibles pour modéliser les expositions médicamenteuses et leurs effets variant dans le temps.

Il est généralement admis que les estimations obtenues à partir des analyses cas-témoins nichées sont similaires à celles obtenues à partir de l'analyse de la cohorte entière, avec seulement une perte modérée de précision (Breslow et al., 1983 ; Liddell et al., 1977). Cependant, tel qu'observé dans les travaux présentés précédemment, ces résultats ont été obtenus avec des comparaisons limitées à des covariables fixes dans le temps (pour des études de simulation) ou à un seul ensemble de données réelles. A notre connaissance, les seules études s'étant appuyées sur des simulations pour évaluer et comparer systématiquement les performances respectives des schémas cas-témoins nichés et cohorte entière pour estimer l'effet de l'exposition médicamenteuse qui varie au cours du temps ont révélé des biais potentiellement importants dans les analyses cas-témoins nichées (Austin et al., 2012 ; Leffondré et al., 2003 ; Leffondré et al., 2010). Cheung et al. (2019) ont montré que ces estimations biaisées obtenues dans deux de ces études de simulation du schéma cas-témoins niché étaient dues au biais des données éparses qui peut être contrôlé par la méthode de réduction de biais pour la régression logistique conditionnelle (Cheung et al., 2019).

Ayant ainsi constaté que très peu de travaux ont évalué et comparé systématiquement les performances respectives des schémas de cohorte et cas-témoins niché (pourtant, très utilisés pour l'analyse des données réelles) pour estimer l'effet d'une exposition médicamenteuse fixe ou variable au cours du temps, nous nous intéresserons particulièrement dans cette thèse à ces deux schémas d'étude classiques.

2 - Matériel et méthodes

Dans ce travail de thèse, notre démarche a été de partir de données de cohorte, réelles ou simulées, et de là de sélectionner l'ensemble des "cas" ayant présenté l'évènement au cours du suivi et d'échantillonner des "témoins" parmi les individus à risque. Nous avons aussi étudié des méthodes d'analyse basées sur les cas seuls mais n'en présentons pas les résultats dans ce mémoire. En effet, les résultats obtenus pour le schéma cas-croisé en particulier étaient décevants, voire contre-intuitifs : nous avons observé un biais qui augmentait avec le nombre de périodes témoins. Le temps a manqué pour explorer et comprendre d'où pouvaient provenir les biais observés. Nous avons donc préféré ne présenter ici que le travail abouti sur l'analyse cas-témoins nichée dans une cohorte.

2.1 Données réelles

Ce travail de thèse est à la fois motivé et illustré par l'analyse de la relation entre l'utilisation de traitements hormonaux de la ménopause (THM) et le risque de cancer du sein. En effet, l'utilisation de THM est aujourd'hui un facteur de risque établi de cancer du sein comme nous le précisons dans une première section. Cette exposition médicamenteuse illustre parfaitement la problématique qui nous intéresse de par sa dimension temporelle. Les données de la cohorte E3N (Etude Epidémiologique auprès de femmes de la Mutuelle Générale de l'Education Nationale) que nous utilisons dans ce travail sont particulièrement pertinentes pour analyser l'association entre utilisation de THM et risque de cancer du sein car elles concernent un grand échantillon de femmes, suivies depuis plus de 30 ans avec une mesure détaillée et actualisée de leur exposition.

2.1.1 Etat des connaissances sur l'association entre utilisation de THM et risque de cancer du sein

De nombreuses études s'accordent aujourd'hui à dire que l'utilisation des THM augmente le risque de cancer du sein. En 1997, le *Collaborative Group on Hormonal Factors in Breast Cancer* a publié une méta-analyse de 51 études épidémiologiques soulignant que les utilisatrices actuelles et récentes de THM présentaient un risque accru de cancer du sein (Collaborative Group on Hormonal Factors in Breast Cancer, 1997). En 2002, l'essai contrôlé randomisé aux Etats-Unis *Women's Health Initiative* a montré un risque accru de cancer du sein et d'évènements cardiovasculaires chez les femmes traitées par THM œstroprogestatif par rapport au groupe placebo (Rossouw

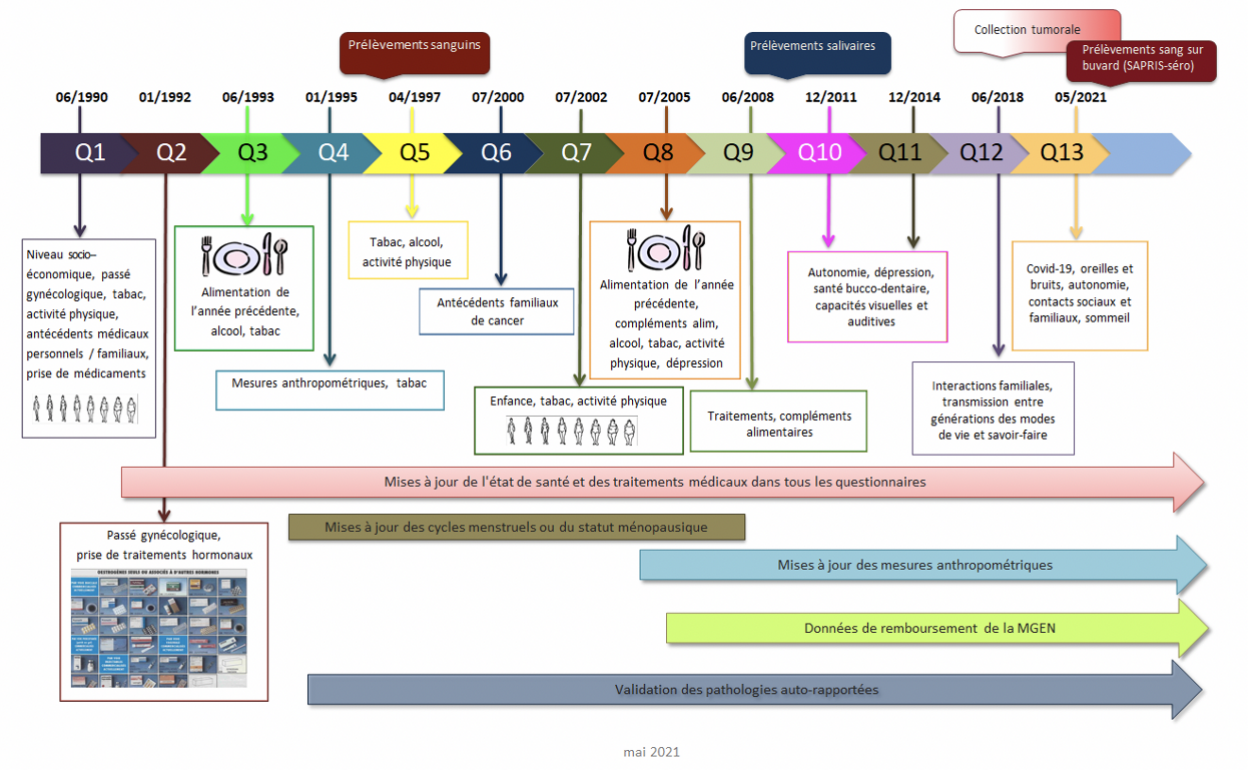
et al., 2002). Les risques globaux pour la santé excédant les bénéfices, l'essai a dû être arrêté prématurément, ce qui a suscité une inquiétude considérable chez les utilisatrices et les prescripteurs de THM du monde entier. Peu de temps après, une grande cohorte de femmes au Royaume-Uni, la *Million Women Study*, a également signalé un risque accru de cancer du sein associé à divers régimes de THM par rapport aux non-utilisatrices (Beral & Million Women Study Collaborators, 2003). Le risque excessif chez les utilisatrices a été confirmé dans la méta-analyse récemment mise à jour par le *Collaborative Group on Hormonal Factors in Breast Cancer*, qui a montré que ce risque augmente régulièrement avec la durée d'utilisation (Collaborative Group on Hormonal Factors in Breast Cancer, 2019). Bien que l'excès de risque soit largement réduit après l'arrêt de THM, il pourrait encore persister au-delà de 10 ans après cet arrêt chez les femmes traitées pendant une longue période (Collaborative Group on Hormonal Factors in Breast Cancer, 2019 ; Vinogradova et al., 2020). D'un point de vue épidémiologique, il s'agit d'une illustration typique d'une exposition dynamique dont les effets varient dans le temps.

2.1.2 Présentation générale de la cohorte E3N

L'étude E3N, Etude Epidémiologique auprès de femmes de la Mutuelle Générale de l'Education Nationale (MGEN) est une étude de cohorte prospective incluant 98 995 femmes nées entre 1925 et 1950 et suivies depuis 1989-1990 (Clavel-Chapelon & E3N Study Group, 2015). Elle s'intéresse principalement à l'impact du mode de vie, de l'alimentation, de l'environnement et des traitements (médicamenteux ou hormonaux) sur la santé des femmes, en particulier le risque de survenue de cancers. Les participantes sont suivies au moyen de questionnaires auto-administrés envoyés tous les deux à trois ans environ et certaines données relatives aux remboursements de soins (utilisation de médicaments en particulier) sont fournies par la MGEN depuis 2004 (voir figure 2.1). Le taux de réponse était d'au moins 75 % pour chaque questionnaire de suivi. La Commission Nationale de l'Informatique et des Libertés (CNIL) a approuvé le protocole de l'étude et les participantes ont signé un consentement écrit.

Nous précisons maintenant l'exposition et l'évènement qui nous intéressent pour l'illustration des travaux méthodologiques de cette thèse.

Figure 2.1 – *Dates d'envoi des auto-questionnaires et leur contenu* (source : www.e3n.fr)



2.1.3 Définition et mesure de l'exposition d'intérêt

Dans ce travail, nous nous intéressons à l'exposition aux THM comme facteur de risque de cancer du sein.

Les informations sur l'utilisation passée d'un quelconque THM ont été recueillies à partir du deuxième questionnaire Q2 envoyé en janvier 1992 et mises à jour dans chacun des questionnaires suivants. Un livret incluant les photos (voir figure 2.1) des spécialités commercialisées en France accompagnait le questionnaire de 1992 pour faciliter la mémorisation des participantes. Ces informations sur l'utilisation de THM comprenaient les noms des marques, la date de début et la durée d'utilisation (Fournier et al., 2014). Les THM incluaient toute utilisation non-vaginale d'estrogènes (sauf l'estriol) ou de tibolone.

Nous avons défini et utilisé pour nos analyses deux types d'exposition aux THM :

— fixe : le statut d'exposition des participantes au début du suivi est considéré comme fixe sur toute la période de l'étude ;

— variable dans le temps : les participantes "non exposées" au début du suivi peuvent changer de statut et devenir "exposées" au cours du suivi.

2.1.4 Définition et mesure de l'évènement d'intérêt

L'évènement d'intérêt pour notre étude est la survenue d'un cancer du sein invasif en premier cancer.

La survenue d'un cancer du sein a été principalement identifiée par auto-déclaration dans chaque questionnaire puis validée et codée après examen des rapports anatamo-pathologiques obtenus auprès des médecins. Quelques cas de cancers supplémentaires ont été identifiés à partir des déclarations des proches, des données de la MGEN ou après recherche des causes de décès auprès du CépiDC de l'Inserm (Clavel-Chapelon & E3N Study Group, 2015).

2.1.5 Sélection de la population d'étude

Pour notre analyse de l'association entre utilisation de THM et risque de cancer du sein, nous avons considéré le début du suivi à la date de remplissage du deuxième questionnaire Q2 envoyé en 1992 et la fin de l'étude à la date du diagnostic d'un cancer, à la date du dernier questionnaire rempli ou en juin 2008 (date de retour du dernier questionnaire considéré pour nos analyses). Parmi les 98 995 femmes de la cohorte E3N, nous avons exclu celles qui n'ont pas répondu au deuxième questionnaire de 1992, celles non ménopausées au début du suivi, celles qui n'ont plus été suivies après leur réponse au deuxième questionnaire et celles atteintes de cancer avant le début du suivi.

Notre population d'étude était donc composée de 38 091 femmes ménopausées qui étaient indemnes de tout cancer lorsqu'elles ont rempli un questionnaire détaillé sur leur utilisation passée d'un quelconque THM en 1992. Parmi elles, 17 194 (45,1 %) avaient déjà utilisé un THM au début de l'étude, tandis que 7 805 (20,5 %) ont commencé à utiliser un THM pendant le suivi. Au total, 2 261 (5,9 %) cancers du sein invasifs ont été diagnostiqués pendant les 532 925 personnes-années du suivi (incidence de 424 cas pour 100 000 personnes-années).

2.2 Étude de simulation

Nous avons utilisé des simulations pour étudier et comparer les propriétés statistiques des estimations obtenues au moyen des différentes méthodes d'analyse considérées. Pour chaque scénario considéré, nous avons généré 1000 échantillons aléatoires indépendants, chacun représentant une cohorte hypothétique de $n = 5\,000$ sujets suivis pendant $m = 2$ ans (730 jours).

Dans cette partie, nous décrivons le protocole de simulations qui comprend : la génération de covariables fixes dans le temps, la génération de l'exposition d'intérêt fixe

ou variable dans le temps, la génération des temps de censure, la génération des temps d'évènement en fonction de l'exposition et des covariables et les scénarios de simulation.

2.2.1 Génération des covariables fixes

Nous avons considéré deux covariables fixes dans le temps (mesurées à l'inclusion) :

— une variable continue, l'âge $(A_i \sim \mathcal{U}(40, 65))$ en années compris entre 40 et 65 ans (à l'image de la cohorte E3N) ;

— une variable binaire, le sexe $(S_i \sim Bernoulli(0, 4))$ avec 40 % d'hommes comme considéré dans l'étude de simulation de Sylvestre et Abrahamowicz (2008).

2.2.2 Génération de l'exposition

2.2.2.1 Génération du statut d'exposition

Le statut d'exposition Z_i d'un individu i, $i = 1, ..., n$ a été généré comme une variable binaire (qui prend la valeur 1 lorsqu'un individu est exposé et la valeur 0 s'il ne l'est pas) à partir d'une distribution de Bernoulli avec une probabilité p d'exposition à l'inclusion pré-spécifiée et maintenue inchangée pendant toute la période de suivi. Cette proportion p pouvait dépendre des covariables fixes. En fonction de l'âge et du sexe, elle était définie comme :

$$logit(p_i) = \alpha_0 + \alpha_A \times A_i + \alpha_S \times S_i$$

où A_i est l'âge en années du sujet i, S_i est le sexe du sujet i avec $S_i = 1$ pour les hommes et $S_i = 0$ pour les femmes. Les paramètres de la fonction logistique $\alpha_A = \log(2)/10$ et $\alpha_S = \log(2)$ reflètent la dépendance de l'exposition à l'âge et au sexe tandis que le paramètre $\alpha_0 = logit(p) - (\log(2)/10 \times \overline{A} + \log(2) \times \overline{S})$ peut être ajusté pour approcher une proportion voulue d'exposés dans la cohorte entière ($\overline{A}$ et $\overline{S}$ étant les valeurs moyennes des deux covariables fixes).

2.2.2.2 Exposition fixe dans le temps

Pour une exposition fixe dans le temps, les individus ayant un statut "exposé" ou "non exposé" sont considérés comme tels pendant toute la période d'observation.

2.2.2.3 Exposition variable dans le temps avec changement unique

Pour une exposition variable au cours du temps à changement unique, telle que l'initiation d'un traitement pendant le suivi, l'état d'exposition $Z_i(t)$ au temps t a été supposé dichotomique : le statut d'exposition des individus peut changer de "non exposé" à "exposé" au maximum une fois pendant l'intervalle de suivi. Une fois exposé, un

sujet reste exposé pendant toute la durée restante du suivi. Ainsi, pour chaque sujet exposé i identifié, nous avons généré son temps de début d'exposition $\tau(i)$ à partir d'une distribution uniforme sur l'intervalle $[0; m]$ où m = durée maximale du suivi pour la cohorte entière. Ensuite, nous avons fixé $Z_i(t) = 0$ pour $0 < t < \tau(i)$ et $Z_i(t) = 1$ pour $t \geq \tau(i)$.

2.2.2.4 Exposition variable dans le temps avec changements multiples

Pour ce type d'exposition, nous avons distingué deux cas : celui d'une exposition ponctuelle répétée (vaccin à plusieurs doses par exemple) et celui d'une exposition répétée de durée variable (traitement antibiotique par exemple).

Dans le cas d'une exposition ponctuelle répétée, le statut d'exposition $Z_i(t)$ prend successivement les valeurs 0 ("non exposé") et 1 ("exposé") avec un premier changement au temps $t_1(i)$, un deuxième de 1 à 0 au temps $t_2(i)$, un troisième au temps $t_3(i)$, etc. Nous avons généré les temps de changement de statut d'exposition successifs de façon aléatoire :

- $t_1(i) \sim \mathcal{U}(1, m)$: loi uniforme sur toute la période du suivi, $t_1(i)$ = début de la première période d'exposition, la durée d'exposition (d) étant fixée à 1 jour ;
- $t_2(i) \sim \mathcal{U}[t_1(i) + d, m]$: loi uniforme entre la fin de la première période d'exposition et le dernier jour du suivi, $t_2(i)$ = début de la deuxième période d'exposition ;
- etc.

De cette façon, le nombre de changements de statut d'exposition était lui-même aléatoire (variable d'un individu à l'autre), comme dans l'étude de Hernán et al. qui évaluait l'association entre le vaccin recombinant contre l'hépatite B et le risque de sclérose en plaques : le nombre de fois que les individus avaient été vaccinés variait de 0 à 3 ou plus pendant la période considérée (Hernán et al., 2004).

Dans le cas d'une exposition répétée de durée variable, la procédure de génération était similaire à celle du cas précédent mais avec une durée d'exposition aléatoire définie selon une distribution normale tronquée $(d \sim t\mathcal{N}(m = 14, sd = 2, l = 0, u = 28))$ de sorte que les valeurs soient comprises entre 0 et 28 jours et presque exclusivement dans l'intervalle $[0; 18]$.

2.2.3 Génération des temps de censure

Nous avons supposé une distribution uniforme pour générer les temps de censure supposés refléter les pertes de vue aléatoires au cours du suivi, avec une censure administrative supplémentaire de tous les sujets qui restaient à risque à l'issue des deux ans de suivi.

2.2.4 Génération des temps d'évènement

La simulation des temps d'évènement comme fonction de l'exposition fixe ou variable dans le temps a été effectuée à l'aide de l'algorithme de permutation spécialement développé et validé pour simuler les temps d'évènement conditionnellement aux effets et/ou covariables variant au cours du temps (Abrahamowicz et al., 1996 ; Mackenzie & Abrahamowicz, 2002 ; Sylvestre & Abrahamowicz, 2008). La fonction *permalgorithm* permettant d'appliquer cet algorithme est implémentée dans le package *PermAlgo* du logiciel R (Sylvestre et al., 2015) et nécessite le nombre d'individus n, la durée maximale de suivi pour ces individus m, une matrice contenant les valeurs possibles des covariables, le vecteur des coefficients de régression β correspondant à ces covariables, les temps d'évènement et de censure pour chaque individu. Pour nos simulations, nous avons supposé une loi exponentielle de paramètre λ constant (choisi selon la proportion d'exposés souhaitée) pour générer les temps d'évènement.

L'algorithme de permutation a été utilisé pour attribuer des temps d'évènement et de censure (classés par ordre croissant) aux vecteurs individuels des valeurs des covariables sur la base d'hypothèses préétablies concernant l'impact de ces covariables sur le risque de survenue d'évènement. Pour nos simulations, nous avons supposé le modèle de Cox défini par

$$h(t|Z, A, S) = h_0(t) \exp(\beta \times Z + \log(2)/10 \times A + \log(2) \times S)$$

ou

$$h(t|Z(t), A, S) = h_0(t) \exp(\beta \times Z(t) + \log(2)/10 \times A + \log(2) \times S)$$

pour une exposition fixe et variable dans le temps respectivement, où β désigne le paramètre associé à l'exposition d'intérêt.

Notre unité de temps pour l'analyse était d'un jour, de sorte que les temps d'évènement et de censure sont arrondis au jour le plus proche.

2.2.5 Scénarios de simulation

Nous nous sommes inspirés d'exemples publiés (Austin et al., 2012 ; Brickner, 2015) pour le choix des scénarios considérés pour nos simulations. Pour chacun des deux types d'exposition fixe et variable dans le temps, nous avons supposé deux valeurs différentes pour l'effet théorique de l'exposition : $\beta = \log(2) = 0{,}6931$ ou $\beta = \log(1{,}25) = 0{,}2231$, soit un *hazard ratio* (HR) de 2 ou 1,25 . Nous avons également fait varier les proportions de sujets exposés (0,10, 0,25 et 0,50) et de sujets ayant connu un évènement non censuré (0,05, 0,10 et 0,25) au cours des deux années de suivi. Cette dernière

proportion a été approximativement contrôlée en ajustant le paramètre d'incidence λ de la distribution exponentielle utilisée pour générer les temps d'évènement. Dans chacun des 18 scénarios ainsi considérés pour chaque type d'exposition fixe ou variable dans le temps, les 1000 ensembles de données de cohorte simulées ont été analysés.

Ce même protocole de simulation a été utilisé pour étudier les propriétés statistiques des méthodes d'analyse basées sur les schémas de cohorte, cas-témoins niché et cas-croisé. Les résultats obtenus étant surprenants, nous avons choisi, pour mieux comprendre la provenance des biais observés, de nous recentrer sur les schémas de cohorte et cas-témoins niché avec des modèles simples sans covariables, comprenant uniquement une exposition fixe ou variable dans le temps à changement unique, soit ($h(t|Z) = h_0(t)\exp(\beta \times Z)$ ou $h(t|Z(t)) = h_0(t)\exp(\beta \times Z(t))$) respectivement.

Dans le cadre de cette thèse, ce sont les résultats restreints à ces schémas et types d'exposition qui seront présentés.

2.3 Analyse statistique

Les données réelles et simulées ont été analysées à l'aide de deux approches : l'une correspondant à une étude de cohorte entière et l'autre à une étude cas-témoins nichée issue de la même cohorte.

2.3.1 Analyse des données de la cohorte entière

Les données réelles et simulées ont été analysées à l'aide d'un modèle à risques proportionnels de Cox non ajusté avec une exposition Z fixe ou $Z(t)$ variable dans le temps. Nous avons utilisé la durée de suivi comme échelle de temps pour l'analyse des données simulées et l'âge comme échelle de temps pour l'analyse des données réelles (Thiébaut & Bénichou, 2004). Les estimations de l'effet (log-HR) de l'exposition, ainsi que l'intervalle de confiance (IC) à 95 %, ont été obtenus à l'aide de la fonction *coxph* du package *survival* de R, avec l'approximation d'Efron (Efron, 1977) pour gérer les temps d'évènements ex-æquos. Nous nous sommes par ailleurs intéressés à l'approximation de Breslow (Breslow, 1974) en complément de celle d'Efron, car c'est l'option par défaut dans le logiciel SAS.

Nous avons quantifié formellement le nombre de temps d'évènements ex-æquos dans les jeux de données réelles et simulées à l'aide des formules suivantes. Soit M_j le nombre de temps d'évènements qui apparaissent j fois dans un ensemble de données. Le nombre

total d'évènements dans cet ensemble de données M est défini par :

$$M = \sum_{j \geq 1} j \times M_j$$

et le nombre K de temps d'évènements distincts par :

$$K = \sum_{j \geq 1} M_j.$$

Le nombre de temps d'évènements ex-æquos est obtenu par soustraction du nombre K de temps d'évènements distincts au nombre total M d'évènements de la cohorte simulée :

$$M - K = \sum_{j \geq 1} (j-1) \times M_j$$

2.3.2 Sélection et analyse des données cas-témoins nichées dans la cohorte

A partir des données de cohorte réelles ou simulées, des données pour une étude cas-témoins nichée ont été échantillonnées. Pour les données simulées, à chacun des 1000 jeux de données correspond une étude cas-témoins. Pour les données réelles issues de la cohorte E3N, nous avons fait le choix de répéter 100 fois l'échantillonnage cas-témoins dans la cohorte constituée (comme dans l'étude d'Essebag et al. (2005)) pour l'analyse de l'association entre utilisation de THM et risque de cancer du sein.

L'échantillonnage des données cas-témoins nichées a été effectué en identifiant d'abord tous les "cas" (sujets qui ont eu l'évènement pendant le suivi) et leur temps d'évènement correspondant dans la cohorte, puis en sélectionnant aléatoirement par densité d'incidence (dans l'ensemble à risque de chaque cas considéré) 1 ou 5 témoins à apparier à chaque cas. Le choix de ces deux scénarios du nombre de témoins par cas s'est fait en s'inspirant de celui proposé par Austin (2012). Pour l'appariement 5 :1, les témoins ont été sélectionnés de manière imbriquée, de sorte que le témoin sélectionné pour l'appariement 1 :1 faisait partie des cinq témoins sélectionnés. Cette procédure d'échantillonnage a été effectuée en utilisant la fonction *ccwc* du package *Epi* de R.

Les échantillons cas-témoins nichés avec 1 ou 5 témoins par cas ont été analysés à l'aide du modèle de régression logistique conditionnelle pour estimer l'association entre l'exposition et le risque de survenue de l'évènement d'intérêt, en conditionnant sur les ensembles appariés. Les estimations de l'effet (log-*Odds Ratio*) de l'exposition pour la régression logistique conditionnelle, équivalentes au log-HR, ont été obtenues à l'aide de la fonction *clogit* du package *survival* de R, avec l'approximation d'Efron pour gérer les

ex-æquos (Efron, 1977) au sein des strates.

2.3.3 Exploration des biais dans l'analyse des données cas-témoins nichées

2.3.3.1 Biais dus aux données éparses

Pour explorer la possibilité de biais dus aux données éparses, nous avons ré-analysé les échantillons cas-témoins nichés en utilisant la régression logistique conditionnelle avec la vraisemblance pénalisée de Firth (Firth, 1993). Nous avons utilisé le code R d'abord développé par Sun et al. pour étendre l'approche de réduction de biais de Firth à la régression logistique conditionnelle (Sun et al., 2011), puis étendu par Cheung et al. pour un nombre illimité de variables d'exposition (Cheung et al., 2019).

2.3.3.2 Biais dus aux temps d'évènements ex-æquos

Nous avons ensuite exploré la possibilité de biais dus à la gestion des temps d'évènements ex-aequos selon deux stratégies :

— **Modification de la méthode de gestion des ex-æquos**
Pour explorer l'impact du choix d'une méthode de gestion des temps d'évènements ex-æquos sur les estimations obtenues par les analyses de régression logistique conditionnelle des études cas-témoins nichées, nous avons refait les analyses en considérant deux méthodes alternatives à l'approximation d'Efron : la méthode exacte (Peto, 1972) et l'approximation de Breslow (1974).

— **Modification de la sélection des cas et témoins appariés**
La fonction *ccwc* que nous avons utilisée pour la sélection des témoins appariés à chaque cas regroupe tous les cas qui ont subi l'évènement d'intérêt au même moment ainsi que leurs témoins respectifs dans la même strate, ce qui peut entraîner la présence d'évènements ex-æquos dans les strates lorsque deux cas ou plus ont subi l'évènement au même moment.
En nous inspirant de la procédure d'échantillonnage des données cas-témoins nichées dans une cohorte de la macro SAS %nCCsampling (Desai et al., 2016), nous avons trouvé un moyen de nous affranchir des problèmes liés à la présence d'ex-æquos dans les strates. En effet, nous avons modifié le code de la fonction *ccwc* en utilisant un identifiant unique pour chaque cas de façon à ce que les cas qui ont vécu l'évènement au même moment soient alloués à des strates distinctes. Ainsi nous avons veillé à ce que chaque strate ne comprenne qu'un seul cas, ce qui permet d'éviter tout ex-æquos. Le code modifié de la fonction *ccwc* est présenté en matériel supplémentaire de l'article soumis (annexe B du mémoire). Les échantillons cas-témoins nichés sélectionnés à l'aide de cette fonction R modi-

fiée ont été analysés par régression logistique conditionnelle avec l'approximation d'Efron, les trois méthodes de gestion des ex-æquos (l'approximation de Breslow et la méthode exacte) aboutissant au même résultat étant donné l'absence d'ex-æquos dans les strates. Cette approche a été appelée "approche *ccwc* modifiée".

Pour augmenter le nombre d'ex-æquos dans les données réelles, nous avons réalisé des analyses secondaires en considérant le mois plutôt que le jour comme unité de temps.

2.3.4 Critères de comparaison pour les données simulées

Pour chaque scénario considéré, nous avons comparé les performances des méthodes statistiques utilisées pour analyser les données cas-témoins nichées et celles de la cohorte entière en utilisant les critères suivants :

— le biais relatif par rapport à la valeur théorique de β utilisée dans la procédure de génération des données

$$b = \frac{\bar{\hat{\beta}} - \beta}{\beta}$$

avec

$$\bar{\hat{\beta}} = \frac{1}{1000} \sum_{k=1}^{1000} \hat{\beta}_k$$

la moyenne empirique des β_k estimés pour chaque simulation $k = 1, ..., 1000$;

— la moyenne $\bar{\hat{\sigma}}$ de tous les écarts-types estimés $\hat{\sigma}_k$ sur l'ensemble des simulations $k = 1, ..., 1000$ (SEE pour *average standard error estimator*)

$$\bar{\hat{\sigma}} = \frac{1}{1000} \sum_{k=1}^{1000} \hat{\sigma}_k;$$

— l'erreur quadratique moyenne (RMSE pour *Root Mean Squared Error*)

$$\sqrt{(\bar{\hat{\beta}} - \beta)^2 + var(\hat{\beta})}$$

où $var(\hat{\beta})$ est la variance empirique des effets estimés à partir des 1000 échantillons simulés ;

— la probabilité de couverture (ou taux de recouvrement) : proportion d'échantillons simulés pour lesquels l'intervalle de confiance à 95 %

$$\hat{\beta}_k \pm 1,96 \times \hat{\sigma}_k$$

de l'effet estimé contient la valeur théorique β. Pour 1000 échantillons simulés, on s'attend à un taux de recouvrement compris entre 0,9365 et 0,9635.

2.3.5 Critères de comparaison pour les données réelles

L'échantillonnage des témoins et les analyses CTN avec le modèle de régression logistique conditionnelle ont été répétés 100 fois pour 1 et 5 témoins par cas, en utilisant quatre approches différentes pour gérer les ex-æquos : la méthode exacte, les approximations de Breslow et d'Efron, ainsi que l'approche *ccwc* modifiée. Pour chaque méthode de gestion des ex-æquos, nous avons comparé les *hazard ratio* (HR) et leur intervalle de confiance à 95 % (IC 95 %) résultant des analyses CTN et de l'analyse de la cohorte entière selon les critères suivants :

— moyenne, minimum et maximum des HR obtenus sur 100 répétitions des analyses CTN ;
— IC 95 % empirique obtenu comme la moyenne des bornes des IC 95 % des HR ;
— différence relative du log-HR : différence entre la moyenne des log-HR issus des analyses CTN et le log-HR de l'analyse de la cohorte sur ce dernier ;
— largeur relative de l'IC 95 % : rapport de la largeur moyenne des IC 95 % des HR issus des analyses CTN sur la largeur de l'IC 95 % de l'HR obtenu par l'analyse de la cohorte entière.

3 - Résultats

Pour cette thèse, nous avons décidé de nous concentrer sur l'analyse cas-témoins nichée. En effet, si beaucoup de travaux existent sur la comparaison de ce schéma d'étude avec l'analyse de la cohorte entière, ceux-ci concernent pour une très grande majorité une exposition fixe dans le temps. Très peu d'articles s'étaient intéressés spécifiquement à une exposition variable dans le temps et un parmi ceux-ci montrait des résultats de simulation contre-intuitifs, à savoir que les estimations issues d'une analyse cas-témoins nichée pouvaient être biaisées, le biais augmentant avec la proportion d'évènements (Austin et al., 2012). Dans un premier temps, nous avons analysé les données simulées et réelles à l'aide des procédures disponibles dans le logiciel R (section 3.1). Ayant constaté des biais similaires à ceux rapportés précédemment, nous avons cherché à corriger ces biais à l'aide d'une méthode de réduction de biais (section 3.2). Nous avons ensuite recherché la cause de ces biais en lien avec la prise en compte des temps d'évènement ex-æquos (section 3.3).

3.1 Analyses de la cohorte entière et cas-témoins nichées (approximation d'Efron)

3.1.1 Données simulées

3.1.1.1 Exposition fixe dans le temps

La figure 3.1 A représente les biais relatifs des estimations du log-HR obtenues à partir des analyses de la cohorte entière et des données CTN pour une exposition fixe dans le temps et un HR théorique de 2. Le tableau 3.1 contenant l'ensemble des résultats de simulation complète ces résultats avec la moyenne des écarts-types estimés, l'erreur quadratique moyenne et la probabilité de couverture. Dans les analyses de la cohorte entière comme CTN, l'approximation d'Efron est l'approche utilisée par défaut pour la gestion des ex-æquos.

Dans tous les scénarios, les estimations du log-HR pour l'analyse de la cohorte entière étaient non biaisées (biais relatif $<1\,\%$ par rapport au log-HR théorique). Les estimations du log-HR pour les analyses des données CTN présentaient un biais négatif qui diminuait avec l'augmentation du nombre de témoins par cas (biais relatif maximal de $27\,\%$ pour un témoin par cas à $6\,\%$ pour cinq témoins par cas, figure 3.1 A et tableau 3.1). Ce biais avait tendance à augmenter avec la proportion d'évènements (pour $10\,\%$ de sujets

exposés et un témoin par cas, le biais relatif passait de 10 % à 27 % pour 5 % et 25 % d'évènements respectivement).

Les estimations issues de l'analyse de la cohorte entière présentaient une variance plus faible que celles des analyses CTN quel que soit le scénario (voir les colonnes SEE présentant les moyennes des écart-types estimés dans le tableau 3.1). La précision des estimations s'améliorait de façon générale pour les deux schémas d'étude à mesure que la proportion d'évènements, la prévalence de l'exposition et (pour les analyses CTN) le nombre de témoins par cas augmentaient.

Dans tous les scénarios, des probabilités de couverture proches de la valeur nominale (0,95) ont été observées pour le schéma de cohorte tandis que les probabilités de couverture pour le schéma d'étude CTN étaient inférieures à la valeur nominale, en particulier pour des proportions d'évènements plus élevées (scénarios pour lesquels des biais considérables ont été observés).

Des résultats similaires à ceux présentés ci-dessus avec un HR=2 ont été obtenus pour des simulations avec un HR=1,25 (figure 3.1 B, tableau 3.2). Dans tous les scénarios, les estimations du log-HR pour l'analyse de la cohorte entière étaient non biaisées sauf pour le scénario de faibles proportions de sujets exposés et d'évènements (respectivement 10 % et 5 %) pour lequel un léger biais négatif de 5 % a été observé (figure 3.1 B). Un biais négatif (mais d'ampleur réduite par rapport à celui observé pour un HR=2) augmentant avec la proportion d'évènements (pour 10 % de sujets exposés et un témoin par cas, le biais relatif passait de 7 % à 21 % pour 5 % et 25 % d'évènements respectivement) a également été observé dans les analyses CTN (figure 3.1 B, tableau 3.2). Cependant, le biais étant plus faible, les probabilités de couverture étaient plus proches de la valeur nominale avec un HR=1,25 qu'avec un HR=2.

Figure 3.1 – *Biais relatif du log-HR estimé dans les schémas de cohorte et CTN (pour 1 et 5 témoins par cas) en fonction de la proportion d'évènements et de la prévalence de l'exposition p, les temps ex-æquos étant gérés selon l'approximation d'Efron : exposition fixe dans le temps et hazard ratio théoriques de 2 (A) et 1,25 (B)*

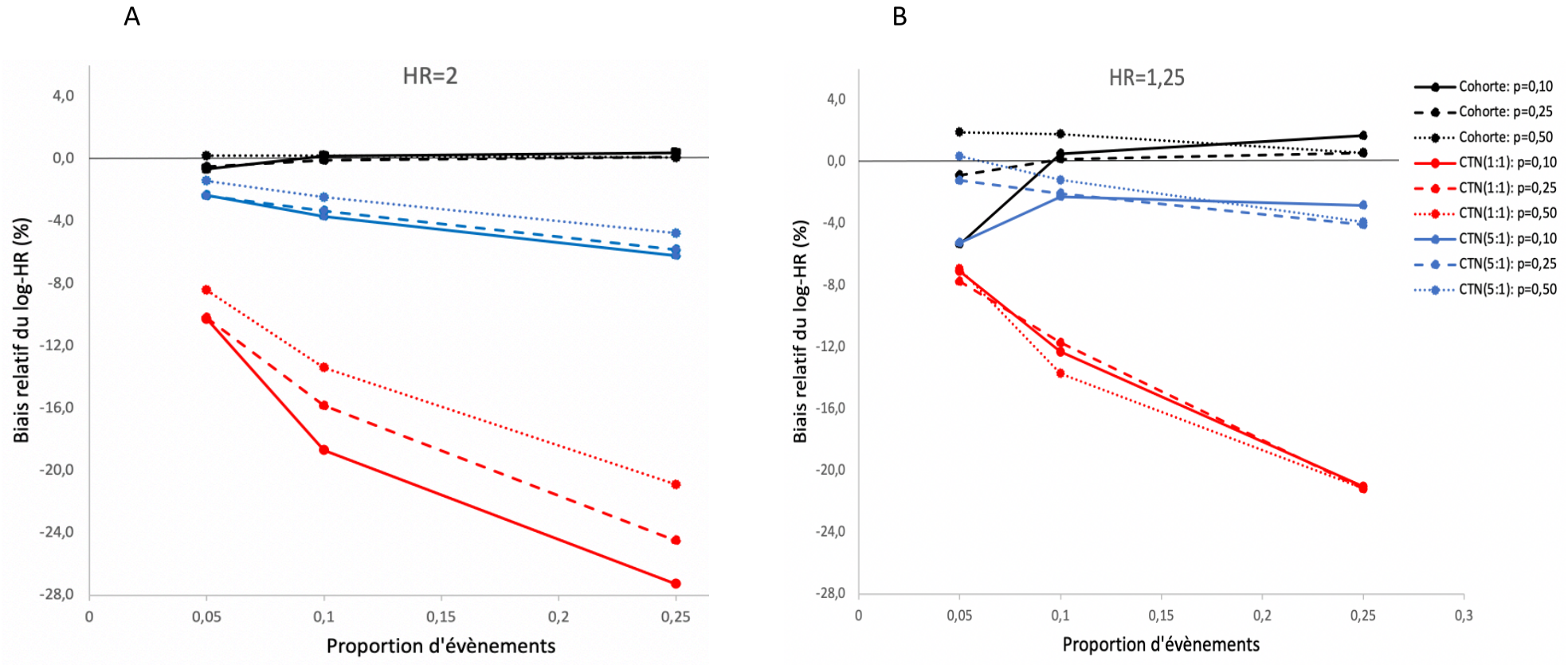


Tableau 3.1 – *Résultats des analyses de la cohorte entière et CTN (pour 1 et 5 témoins par cas) en fonction de la proportion d'évènements et de la prévalence de l'exposition p, les temps ex-æquos étant gérés selon l'approximation d'Efron : exposition fixe dans le temps et hazard ratio théorique de 2*

10% sujets exposés												
Schéma d'étude	5% d'évènements				10% d'évènements				25% d'évènements			
	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)
Cohorte	-0,636	0,167	0,166	95,3	0,200	0,116	0,113	95,5	0,406	0,076	0,076	95,4
CTN (1:1)	-10,256	0,244	0,251	93,8	-18,677	0,155	0,194	87,6	-27,263	0,091	0,205	45,7
CTN (5:1)	-2,311	0,188	0,184	95,0	-3,673	0,128	0,128	95,4	-6,190	0,081	0,092	91,6
25% sujets exposés												
Schéma d'étude	5% d'évènements				10% d'évènements				25% d'évènements			
	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)
Cohorte	-0,498	0,130	0,129	95,4	-0,076	0,090	0,089	95,8	0,137	0,058	0,057	96,0
CTN (1:1)	-10,193	0,182	0,187	94,0	-15,829	0,117	0,154	86,1	-24,480	0,068	0,181	28,4
CTN (5:1)	-2,361	0,143	0,145	95,4	-3,328	0,098	0,098	95,0	-5,812	0,061	0,073	91,0
50% sujets exposés												
Schéma d'étude	5% d'évènements				10% d'évènements				25% d'évènements			
	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)
Cohorte	0,207	0,135	0,137	94,8	0,235	0,093	0,091	94,8	0,089	0,059	0,059	95,5
CTN (1:1)	-8,408	0,177	0,187	93,9	-13,388	0,116	0,143	88,8	-20,890	0,068	0,158	42,1
CTN (5:1)	-1,397	0,144	0,147	94,5	-2,446	0,098	0,097	95,4	-4,752	0,061	0,070	91,5

SEE *average standard error estimator* ; RMSE *root mean squared error* ; PC probabilité de couverture ; CTN cas-témoins niché

Tableau 3.2 – *Résultats des analyses de la cohorte entière et CTN (pour 1 et 5 témoins par cas) en fonction de la proportion d'évènements et de la prévalence de l'exposition p, les temps ex-æquos étant gérés selon l'approximation d'Efron : exposition fixe dans le temps et hazard ratio théorique de 1,25*

10% sujets exposés												
Schéma d'étude	5% d'évènements				10% d'évènements				25% d'évènements			
	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)
Cohorte	-5,331	0,196	0,204	94,6	0,529	0,135	0,136	94,3	1,697	0,087	0,087	94,7
CTN (1:1)	-7,078	0,269	0,267	95,7	-12,308	0,173	0,169	95,8	-21,038	0,102	0,111	93,2
CTN (5:1)	-5,243	0,215	0,225	94,0	-2,251	0,145	0,144	95,6	-2,824	0,091	0,091	95,2
25% sujets exposés												
Schéma d'étude	5% d'évènements				10% d'évènements				25% d'évènements			
	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)
Cohorte	-0,865	0,140	0,139	94,8	0,166	0,097	0,099	94,3	0,562	0,062	0,062	96,4
CTN (1:1)	-7,751	0,189	0,188	94,9	-11,707	0,123	0,126	94,3	-21,183	0,072	0,085	91,1
CTN (5:1)	-1,206	0,152	0,147	96,1	-2,046	0,104	0,105	95,0	-4,071	0,065	0,065	95,4
50% sujets exposés												
Schéma d'étude	5% d'évènements				10% d'évènements				25% d'évènements			
	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)
Cohorte	1,921	0,128	0,128	95,4	1,795	0,088	0,087	95,0	0,579	0,057	0,057	95,0
CTN (1:1)	-6,928	0,169	0,165	94,7	-13,706	0,111	0,111	94,5	-21,156	0,065	0,080	88,3
CTN (5:1)	0,356	0,138	0,139	94,5	-1,175	0,094	0,092	95,2	-3,902	0,059	0,060	94,7

SEE *average standard error estimator* ; RMSE *root mean squared error* ; PC probabilité de couverture ; CTN cas-témoins niché

3.1.1.2 Exposition variable dans le temps

Des résultats similaires à ceux d'une exposition fixe dans le temps ont été obtenus pour une exposition variable dans le temps avec un HR=2 (tableau 3.3) ou un HR=1,25 (tableau 3.4).

Dans tous les scénarios, l'analyse de la cohorte entière a donné des estimations non biaisées (biais relatif <2,5 %). Cependant, les estimations du log-HR issues des analyses CTN présentaient toujours un biais négatif qui diminuait avec l'augmentation du nombre de témoins par cas (biais relatif maximal de 27 % pour un témoin par cas à 7 % pour cinq témoins par cas) mais augmentait avec la proportion élevée d'évènements (figure 3.2 A).

Dans tous les scénarios, les variances des estimations issues de l'analyse de la cohorte entière étaient plus faibles que celles des estimations des analyses CTN (SEE dans le tableau 3.3). Une amélioration de la précision des estimations a été observée pour les analyses de données des deux schémas d'étude à mesure que la proportion d'évènements, la prévalence de l'exposition et (pour les analyses CTN) le nombre de témoins par cas augmentaient.

Des probabilités de couverture proches de la valeur nominale (0,95) ont été observées pour le schéma de cohorte dans tous les scénarios alors que certaines probabilités de couverture du schéma d'étude CTN étaient inférieures à la valeur nominale, en particulier pour des proportions d'évènements plus élevées (10 % d'évènements pour un témoin par cas et 25 % d'évènements pour 1 et 5 témoins par cas, tableau 3.3).

Les résultats des simulations avec un HR=1,25 montrent que les estimations du log-HR pour l'analyse de la cohorte entière étaient non biaisées sauf pour les scénarios avec 5 % d'évènements. En effet, pour ces scénarios, on a observé un biais négatif compris entre 6 % et 8 % (figure 3.2 B, tableau 3.4). De même que pour un HR=2, nous avons observé, pour un HR=1,25, un biais négatif (mais d'ampleur réduite par rapport à celui observé pour un HR=2) qui augmentait avec la proportion d'évènements (pour 25 % de sujets exposés et un témoin par cas, le biais relatif passait de 13 % à 24 % pour 5 % et 25 % d'évènements respectivement) dans les analyses CTN (figure 3.2 B,, tableau 3.4).

Figure 3.2 – *Biais relatif du log-HR estimé dans les schémas de cohorte et CTN (pour 1 et 5 témoins par cas) en fonction de la proportion d'évènements et de la prévalence de l'exposition p, les temps ex-æquos étant gérés selon l'approximation d'Efron : exposition variable dans le temps et hazard ratio théoriques de 2 (A) et 1,25 (B)*

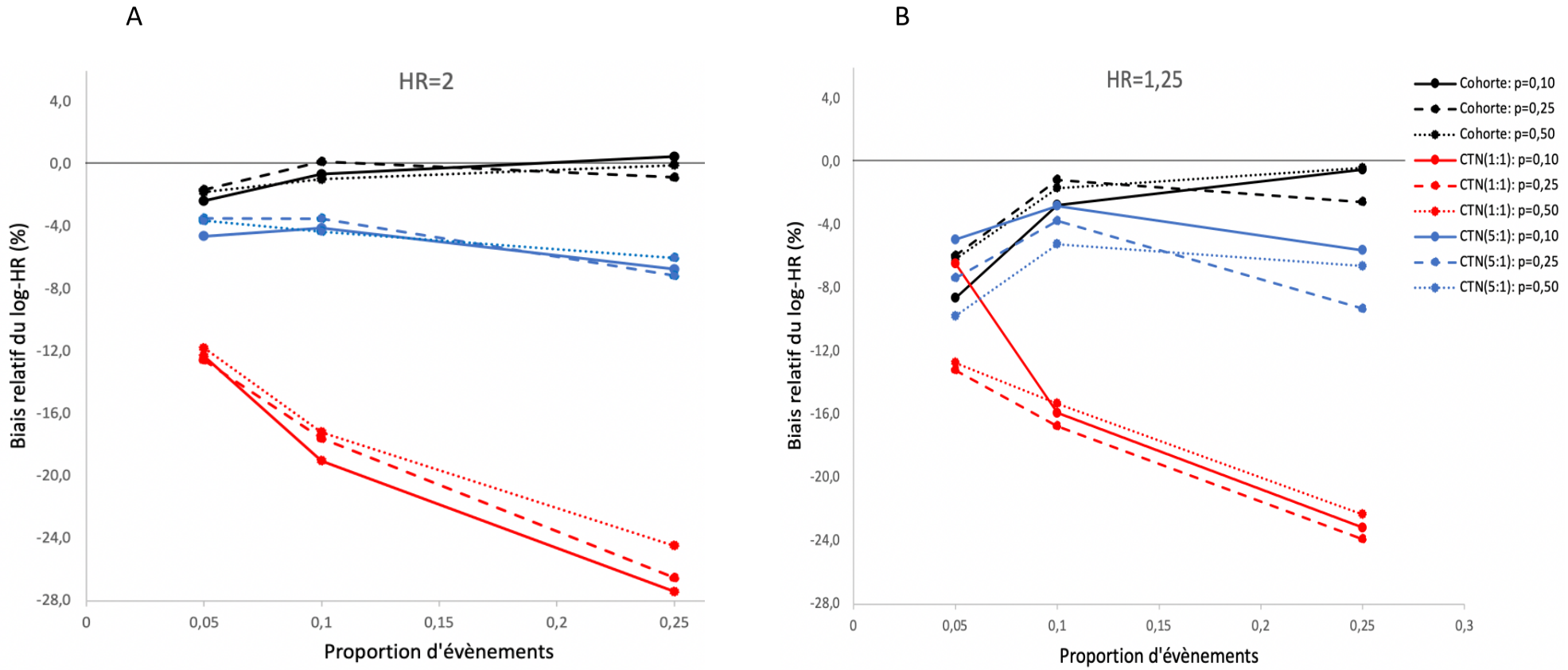


Tableau 3.3 – *Résultats des analyses de la cohorte entière et CTN (pour 1 et 5 témoins par cas) en fonction de la proportion d'évènements et de la prévalence de l'exposition p, les temps ex-æquos étant gérés selon l'approximation d'Efron : exposition variable dans le temps et hazard ratio théorique de 2*

10% sujets exposés												
Schéma d'étude	5% d'évènements				10% d'évènements				25% d'évènements			
	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)
Cohorte	-2,328	0,226	0,226	96,1	-0,611	0,157	0,146	96,9	0,506	0,104	0,100	96,7
CTN (1:1)	-12,283	0,335	0,328	95,2	-19,010	0,213	0,232	92,4	-27,395	0,125	0,218	70,6
CTN (5:1)	-4,591	0,256	0,264	94,1	-4,080	0,174	0,165	96,5	-6,716	0,111	0,116	95,3
25% sujets exposés												
Schéma d'étude	5% d'évènements				10% d'évènements				25% d'évènements			
	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)
Cohorte	-1,616	0,163	0,170	94,2	0,183	0,113	0,113	95,1	-0,820	0,074	0,074	94,5
CTN (1:1)	-12,551	0,234	0,251	93,0	-17,568	0,150	0,186	88,3	-26,514	0,088	0,200	45,9
CTN (5:1)	-3,458	0,182	0,191	94,3	-3,471	0,124	0,125	94,9	-7,119	0,079	0,091	90,8
50% sujets exposés												
Schéma d'étude	5% d'évènements				10% d'évènements				25% d'évènements			
	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)
Cohorte	-1,766	0,143	0,147	94,0	-0,938	0,099	0,099	94,3	-0,044	0,064	0,066	94,6
CTN (1:1)	-11,772	0,198	0,211	92,6	-17,178	0,128	0,171	85,7	-24,454	0,075	0,184	37,3
CTN (5:1)	-3,611	0,157	0,160	93,8	-4,305	0,107	0,111	94,1	-5,983	0,067	0,081	89,2

SEE *average standard error estimator* ; RMSE *root mean squared error* ; PC probabilité de couverture ; CTN cas-témoins niché

Tableau 3.4 – *Résultats des analyses de la cohorte entière et CTN (pour 1 et 5 témoins par cas) en fonction de la proportion d'évènements et de la prévalence de l'exposition p, les temps ex-æquos étant gérés selon l'approximation d'Efron : exposition variable dans le temps et hazard ratio théorique de 1,25*

10% sujets exposés												
Schéma d'étude	5% d'évènements				10% d'évènements				25% d'évènements			
	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)
Cohorte	-8,613	0,273	0,287	95,2	-2,719	0,188	0,177	97,1	-0,470	0,123	0,123	95,3
CTN (1:1)	-6,424	0,376	0,382	95,6	-15,887	0,242	0,234	96,4	-23,160	0,144	0,145	95,4
CTN (5:1)	-4,907	0,299	0,304	95,3	-2,774	0,203	0,195	96,5	-5,576	0,129	0,131	94,7
25% sujets exposés												
Schéma d'étude	5% d'évènements				10% d'évènements				25% d'évènements			
	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)
Cohorte	-5,921	0,185	0,184	95,7	-1,122	0,128	0,129	95,5	-2,531	0,084	0,085	94,3
CTN (1:1)	-13,185	0,252	0,250	95,1	-16,726	0,164	0,164	94,9	-23,891	0,098	0,110	92,0
CTN (5:1)	-7,342	0,202	0,206	95,4	-3,718	0,138	0,140	95,7	-9,278	0,088	0,091	94,6
50% sujets exposés												
Schéma d'étude	5% d'évènements				10% d'évènements				25% d'évènements			
	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)
Cohorte	-6,144	0,151	0,158	94,7	-1,637	0,104	0,105	94,8	-0,359	0,068	0,071	93,9
CTN (1:1)	-12,709	0,203	0,208	94,3	-15,293	0,132	0,140	93,7	-22,324	0,079	0,094	91,5
CTN (5:1)	-9,748	0,163	0,171	94,4	-5,197	0,112	0,112	94,5	-6,585	0,071	0,075	94,8

SEE *average standard error estimator* ; RMSE *root mean squared error* ; PC probabilité de couverture ; CTN cas-témoins niché

3.1.2 Données réelles

Le tableau 3.5 présente les résultats comparant les estimations obtenues pour 100 répétitions de l'analyse CTN à celles résultant de l'analyse de la cohorte entière. L'ensemble des analyses, que ce soit CTN ou sur la cohorte entière, ont montré que l'utilisation passée des THM, qu'elle soit mesurée à l'inclusion (exposition fixe dans le temps) ou actualisée tout au long du suivi (exposition variable dans le temps), était associée à une augmentation statistiquement significative, bien que modérée, du risque de cancer du sein (tableau 3.5). Les HR estimés pour la cohorte entière étaient de 1,23 (IC 95 %, 1,13 à 1,34) et de 1,49 (IC 95 %, 1,36 à 1,64) respectivement pour l'exposition fixe et variable dans le temps. Les estimations moyennes des analyses CTN avec cinq témoins par cas étaient plus proches de celles issues de l'analyse de la cohorte entière que les estimations moyennes des analyses CTN avec un témoin par cas (différence relative du log-HR d'environ 1,4 % contre 5 % et 0,4 % contre 4,4 % respectivement pour l'exposition fixe et variable dans le temps des THM, tableau 3.5). Pour les deux types d'exposition, la précision des estimations était meilleure pour les analyses de cohorte que pour les analyses CTN dont la précision s'améliorait avec l'augmentation du nombre de témoins par cas (voir colonne donnant la largeur relative des IC 95 % du tableau 3.5).

Tableau 3.5 – *Analyse de l'association entre utilisation de THM et risque de cancer du sein dans la cohorte E3N selon un schéma de cohorte ou CTN, le jour étant pris comme unité de temps et les temps ex-æquos étant gérés selon l'approximation d'Efron*

Schéma d'étude	Utilisation fixe dans le temps des THM				
	HR (IC 95%)	HR Minimum	HR Maximum	Différence relative des log-HR (%)	Largeur relative des IC 95%
Cohorte	1,229 (1,131 – 1,335)	-	-	Réf.	Réf.
CTN (1 :1)	1,217 (1,088 – 1,361)	1,113	1,344	-5,07	1,333
CTN (5 :1)	1,225 (1,120 – 1,341)	1,173	1,297	-1,39	1,081
Schéma d'étude	Utilisation variable dans le temps des THM				
	HR (IC 95%)	HR Minimum	HR Maximum	Différence relative des log-HR (%)	Largeur relative des IC 95%
Cohorte	1,494 (1,364 – 1,637)	-	-	Réf.	Réf.
CTN (1 :1)	1,470 (1,303 – 1,658)	1,282	1,636	-4,36	1,300
CTN (5 :1)	1,497 (1,356 – 1,651)	1,429	1,572	0,38	1,078

HR (IC 95 %) *hazard ratio* (avec un intervalle de confiance à 95 %) ; CTN cas-témoins niché ;
Résultats CTN pour 100 échantillonnages indépendants

Au vu des biais constatés pour les analyses CTN, nous avons recherché des explications. La première que nous avons explorée est celle d'un biais de données éparses pour lequel des méthodes de correction ont été proposées dans la littérature, comme nous l'avons décrit plus haut.

3.2 Correction des biais dus aux données éparses dans les analyses cas-témoins nichées par une méthode de réduction des biais

3.2.1 Données simulées

3.2.1.1 Exposition fixe dans le temps

L'ensemble des résultats de simulation sont présentés dans le tableau 3.6.

Dans tous les scénarios, nous avons constaté une diminution considérable des biais après recours à la méthode de Firth pour la réduction des biais. En effet, le biais relatif maximal était de seulement 5 % pour les analyses CTN après réduction des biais tandis qu'il était de 27 % pour les mêmes analyses utilisant l'approximation d'Efron (tableau 3.6).

La précision des estimations après réduction de biais était systématiquement inférieure à celle des estimations des mêmes analyses CTN avant réduction des biais. Ainsi, l'erreur quadratique moyenne (RMSE) n'était avantageuse pour la méthode de réduction de biais que pour les scénarios avec des proportions d'évènements plus élevées (25 %) pour lesquels des biais considérables avaient été observés dans un premier temps (voir les colonnes RMSE dans le tableau 3.6).

Les très faibles probabilités de couverture observées précédemment pour le schéma d'étude CTN avec l'approximation d'Efron se sont nettement améliorées après réduction des biais (tableau 3.6).

Des résultats similaires ont été observés pour un HR théorique égal à 1,25. Une diminution considérable des biais a été observée après réduction des biais selon la méthode de Firth (biais maximal de 4 % alors qu'il était de 21 % avec l'approximation d'Efron, tableau 3.7). Cependant, l'ampleur des biais issus des analyses CTN avec l'approximation d'Efron ayant diminué pour HR=1,25 par rapport à HR=2, l'erreur quadratique moyenne (RMSE) était plus favorable pour les estimations CTN avant qu'après réduction des biais pour tous les scénarios (tableau 3.7).

Tableau 3.6 – *Résultats des analyses CTN (pour 1 et 5 témoins par cas) avec réduction des biais selon la méthode de Firth : exposition fixe dans le temps et hazard ratio théorique de 2*

10% sujets exposés												
Schéma d'étude	5% d'évènements				10% d'évènements				25% d'évènements			
	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)
CTN (1:1) Ef	-10,256	0,244	0,251	93,8	-18,677	0,155	0,194	87,6	-27,263	0,091	0,205	45,7
CTN (1:1) RB	2,831	0,276	0,302	93,7	2,058	0,192	0,202	94,2	2,626	0,127	0,134	93,6
CTN (5:1) Ef	-2,311	0,188	0,184	95,0	-3,673	0,128	0,128	95,4	-6,190	0,081	0,092	91,6
CTN (5:1) RB	0,882	0,191	0,188	95,3	0,876	0,134	0,132	95,1	1,221	0,088	0,091	94,8
25% sujets exposés												
Schéma d'étude	5% d'évènements				10% d'évènements				25% d'évènements			
	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)
CTN (1:1) Ef	-10,193	0,182	0,187	94,0	-15,829	0,117	0,154	86,1	-24,480	0,068	0,181	28,4
CTN (1:1) RB	1,413	0,203	0,210	94,4	3,738	0,142	0,151	94,0	4,595	0,093	0,105	90,9
CTN (5:1) Ef	-2,361	0,143	0,145	95,4	-3,328	0,098	0,098	95,0	-5,812	0,061	0,073	91,0
CTN (5:1) RB	-0,152	0,146	0,148	94,8	0,451	0,101	0,100	95,7	0,872	0,066	0,067	94,0
50% sujets exposés												
Schéma d'étude	5% d'évènements				10% d'évènements				25% d'évènements			
	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)
CTN (1:1) Ef	-8,408	0,177	0,187	93,9	-13,388	0,116	0,143	88,8	-20,890	0,068	0,158	42,1
CTN (1:1) RB	0,849	0,191	0,205	94,8	2,784	0,133	0,141	93,9	4,432	0,086	0,099	91,2
CTN (5:1) Ef	-1,397	0,144	0,147	94,5	-2,446	0,098	0,097	95,4	-4,752	0,061	0,070	91,5
CTN (5:1) RB	-0,157	0,146	0,149	94,0	0,392	0,101	0,098	95,8	0,772	0,064	0,066	94,2

SEE *average standard error estimator* ; RMSE *root mean squared error* ; PC probabilité de couverture ; CTN cas-témoins niché ; Ef Efron ; RB réduction de biais ; les résultats des lignes grisées sont les mêmes que dans le tableau 3.1 et sont inclus ici pour faciliter la comparaison.

Tableau 3.7 – *Résultats des analyses CTN (pour 1 et 5 témoins par cas) avec réduction des biais selon la méthode de Firth : exposition fixe dans le temps et hazard ratio théorique de 1,25*

10% sujets exposés												
Schéma d'étude	5% d'évènements				10% d'évènements				25% d'évènements			
	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)
CTN (1:1) Ef	-7,078	0,269	0,267	95,7	-12,308	0,173	0,169	95,8	-21,038	0,102	0,111	93,2
CTN (1:1) RB	4,155	0,288	0,296	95,6	4,345	0,199	0,200	94,5	3,862	0,129	0,135	94,1
CTN (5:1) Ef	-5,243	0,215	0,225	94,0	-2,251	0,145	0,144	95,6	-2,824	0,091	0,091	95,2
CTN (5:1) RB	1,797	0,216	0,226	93,7	3,373	0,150	0,148	96,0	3,973	0,097	0,097	94,8
25% sujets exposés												
Schéma d'étude	5% d'évènements				10% d'évènements				25% d'évènements			
	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)
CTN (1:1) Ef	-7,751	0,189	0,188	94,9	-11,707	0,123	0,126	94,3	-21,183	0,072	0,085	91,1
CTN (1:1) RB	1,664	0,203	0,207	94,6	4,073	0,141	0,148	94,4	2,198	0,091	0,094	94,2
CTN (5:1) Ef	-1,206	0,152	0,147	96,1	-2,046	0,104	0,105	95,0	-4,071	0,065	0,065	95,4
CTN (5:1) RB	1,948	0,154	0,149	95,7	1,745	0,107	0,108	94,9	1,666	0,069	0,068	95,1
50% sujets exposés												
Schéma d'étude	5% d'évènements				10% d'évènements				25% d'évènements			
	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)
CTN (1:1) Ef	-6,928	0,169	0,165	94,7	-13,706	0,111	0,111	94,5	-21,156	0,065	0,080	88,3
CTN (1:1) RB	1,125	0,180	0,179	94,5	0,022	0,125	0,124	94,9	0,310	0,080	0,083	94,3
CTN (5:1) Ef	0,356	0,138	0,139	94,5	-1,175	0,094	0,092	95,2	-3,902	0,059	0,060	94,7
CTN (5:1) RB	1,598	0,140	0,141	94,7	1,621	0,096	0,094	95,2	1,145	0,062	0,063	95,3

SEE *average standard error estimator* ; RMSE *root mean squared error* ; PC probabilité de couverture ; CTN cas-témoins niché ; Ef Efron ; RB réduction de biais ; les résultats des lignes grisées sont les mêmes que dans le tableau 3.2 et sont inclus ici pour faciliter la comparaison.

3.2.1.2 Exposition variable dans le temps

Des résultats similaires à ceux d'une exposition fixe dans le temps ont été obtenus pour une exposition variable dans le temps avec HR=2 (tableau 3.8) ou HR=1,25 (tableau 3.9).

Dans tous les scénarios, les biais négatifs des estimations du log-HR issues des analyses CTN ont été considérablement réduits après l'utilisation de la méthode de réduction de biais, passant de 27 % à 3 % (tableau 3.8) pour un biais relatif maximal.

Dans tous les scénarios, la précision des estimations issues de l'analyse CTN avec la méthode de réduction de biais était à nouveau systématiquement inférieure à celle des estimations de l'analyse CTN avec l'approximation d'Efron. Les RMSE n'étaient ainsi avantageuses pour la méthode de réduction de biais que pour une plus grande proportion d'évènements (25 %, tableau 3.8).

Les probabilités de couverture du schéma d'étude CTN qui étaient précédemment inférieures à la valeur nominale avec l'approximation d'Efron ont considérablement augmenté vers la valeur nominale après réduction des biais (tableau 3.8).

Le tableau 3.9 présentant les résultats des simulations pour un HR=1,25 montre une diminution considérable des biais dans les estimations des analyses CTN avec la méthode de réduction des biais (biais maximal étant de 5 % alors qu'il était de 24 % avec l'approximation d'Efron, tableau 3.9). Cependant, la légère diminution de l'ampleur des biais issus des analyses CTN avec l'approximation d'Efron pour HR=1,25 (par rapport à HR=2) a rendu l'erreur quadratique moyenne (RMSE) plus favorable pour les estimations CTN avec cette méthode d'approximation qu'avec la méthode de réduction de biais pour tous les scénarios (tableau 3.9).

Tableau 3.8 – *Résultats des analyses CTN (pour 1 et 5 témoins par cas) avec réduction des biais selon la méthode de Firth : exposition variable dans le temps et hazard ratio théorique de 2*

10% sujets exposés												
Schéma d'étude	5% d'évènements				10% d'évènements				25% d'évènements			
	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)
CTN (1:1) Ef	-12,283	0,335	0,328	95,2	-19,010	0,213	0,232	92,4	-27,395	0,125	0,218	70,6
CTN (1:1) RB	-0,315	0,373	0,383	95,4	1,675	0,263	0,268	96,0	2,552	0,175	0,177	95,4
CTN (5:1) Ef	-4,591	0,256	0,264	94,1	-4,080	0,174	0,165	96,5	-6,716	0,111	0,116	95,3
CTN (5:1) RB	-0,429	0,259	0,267	93,7	0,915	0,182	0,171	95,6	0,820	0,121	0,117	96,3
25% sujets exposés												
Schéma d'étude	5% d'évènements				10% d'évènements				25% d'évènements			
	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)
CTN (1:1) Ef	-12,551	0,234	0,251	93,0	-17,568	0,150	0,186	88,3	-26,514	0,088	0,200	45,9
CTN (1:1) RB	-0,569	0,262	0,289	93,4	2,500	0,184	0,197	94,3	2,622	0,122	0,130	94,3
CTN (5:1) Ef	-3,458	0,182	0,191	94,3	-3,471	0,124	0,125	94,9	-7,119	0,079	0,091	90,8
CTN (5:1) RB	-0,668	0,185	0,194	94,2	0,730	0,129	0,130	94,3	-0,261	0,085	0,084	94,3
50% sujets exposés												
Schéma d'étude	5% d'évènements				10% d'évènements				25% d'évènements			
	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)
CTN (1:1) Ef	-11,772	0,198	0,211	92,6	-17,178	0,128	0,171	85,7	-24,454	0,075	0,184	37,3
CTN (1:1) RB	-1,052	0,218	0,230	94,2	1,186	0,153	0,164	93,9	3,279	0,100	0,114	91,8
CTN (5:1) Ef	-3,611	0,157	0,160	93,8	-4,305	0,107	0,111	94,1	-5,983	0,067	0,081	89,2
CTN (5:1) RB	-1,629	0,160	0,162	93,8	-0,765	0,111	0,113	95,0	0,395	0,072	0,076	94,3

SEE *average standard error estimator* ; RMSE *root mean squared error* ; PC probabilité de couverture ; CTN cas-témoins niché ; Ef Efron ; RB réduction de biais ; les résultats des lignes grisées sont les mêmes que dans le tableau 3.3 et sont inclus ici pour faciliter la comparaison.

Tableau 3.9 – *Résultats des analyses CTN (pour 1 et 5 témoins par cas) avec réduction des biais selon la méthode de Firth : exposition variable dans le temps et hazard ratio théorique de 1,25*

10% sujets exposés												
Schéma d'étude	5% d'évènements				10% d'évènements				25% d'évènements			
	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)
CTN (1:1) Ef	-6,424	0,376	0,382	95,6	-15,887	0,242	0,234	96,4	-23,160	0,144	0,145	95,4
CTN (1:1) RB	5,372	0,396	0,417	95,1	1,047	0,276	0,278	95,8	1,551	0,182	0,181	94,9
CTN (5:1) Ef	-4,907	0,299	0,304	95,3	-2,774	0,203	0,195	96,5	-5,576	0,129	0,131	94,7
CTN (5:1) RB	8,012	0,298	0,302	94,9	5,652	0,207	0,200	96,3	2,238	0,136	0,138	94,7
25% sujets exposés												
Schéma d'étude	5% d'évènements				10% d'évènements				25% d'évènements			
	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)
CTN (1:1) Ef	-13,185	0,252	0,250	95,1	-16,726	0,164	0,164	94,9	-23,891	0,098	0,110	92,0
CTN (1:1) RB	-3,690	0,269	0,273	95,2	-1,908	0,187	0,189	94,9	-1,126	0,123	0,128	93,8
CTN (5:1) Ef	-7,342	0,202	0,206	95,4	-3,718	0,138	0,140	95,7	-9,278	0,088	0,091	94,6
CTN (5:1) RB	-1,905	0,204	0,207	95,0	1,111	0,141	0,144	95,7	-3,356	0,093	0,095	95,1
50% sujets exposés												
Schéma d'étude	5% d'évènements				10% d'évènements				25% d'évènements			
	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)
CTN (1:1) Ef	-12,709	0,203	0,208	94,3	-15,293	0,132	0,140	93,7	-22,324	0,079	0,094	91,5
CTN (1:1) RB	-4,063	0,217	0,228	94,4	-1,096	0,151	0,159	93,2	0,077	0,098	0,104	94,2
CTN (5:1) Ef	-9,748	0,163	0,171	94,4	-5,197	0,112	0,112	94,5	-6,585	0,071	0,075	94,8
CTN (5:1) RB	-7,118	0,165	0,173	94,3	-1,753	0,115	0,115	94,8	-1,192	0,075	0,078	94,4

SEE *average standard error estimator* ; RMSE *root mean squared error* ; PC probabilité de couverture ; CTN cas-témoins niché ; Ef Efron ; RB réduction de biais ; les résultats des lignes grisées sont les mêmes que dans le tableau 3.4 et sont inclus ici pour faciliter la comparaison.

3.2.2 Données réelles

Le tableau 3.10 compare les estimations obtenues pour 100 répétitions de l'analyse CTN avec la méthode de Firth pour la réduction des biais à celles résultant de l'analyse de la cohorte entière. Après application de la méthode de réduction des biais pour les analyses CTN, l'utilisation des THM était toujours significativement associée à un risque modéré de cancer du sein (tableau 3.10). Cependant, pour un témoin par cas, les estimations moyennes des analyses CTN avec réduction des biais étaient plus proches de celles issues de l'analyse de la cohorte entière que les estimations moyennes des analyses CTN avec l'approximation d'Efron (différence relative du log-HR d'environ 2,6 % contre 5,1 % et 0,9 % contre 4,4 % respectivement pour l'utilisation fixe et variable dans le temps des THM, tableau 3.10). Pour cinq témoins par cas, cette diminution de la différence relative s'observait peu ou pas car l'augmentation du nombre de témoins par cas à elle seule avait déjà considérablement réduit la différence observée entre les estimations des analyses CTN (1 :1) avec l'approximation d'Efron et celles des analyses de la cohorte entière. La précision des estimations issues de l'analyse CTN avec réduction de biais était systématiquement inférieure à celle des estimations de l'analyse CTN avec l'approximation d'Efron. Néanmoins, elle s'améliorait avec l'augmentation du nombre de témoins par cas (voir colonne donnant la largeur relative des IC 95 % dans le tableau 3.10).

La nette amélioration des estimations obtenues après le recours à une méthode de réduction de biais semble suggérer que les biais observés pourraient résulter de données éparses. Cependant, cette explication ne nous a pas satisfaits car les biais persistaient quelle que soit la taille des cohortes simulées (résultats non présentés dans ce mémoire) et surtout quel que soit le nombre d'évènements. Il restait même contre-intuitif que les biais augmentent alors que la proportion d'évènements augmentait alors même que le risque de données éparses diminuait. Nous avons donc poursuivi nos investigations en nous intéressant à la gestion des temps d'évènement ex-æquos à la fois dans l'analyse des données de cohorte ou CTN et dans la sélection des témoins pour les analyses CTN.

Tableau 3.10 – *Analyse de l'association entre utilisation de THM et risque de cancer du sein dans la cohorte E3N selon les schémas de cohorte et CTN avec la méthode de Firth pour la réduction des biais, le jour étant pris comme unité de temps*

Schéma d'étude	Utilisation fixe dans le temps des THM				
	HR (IC 95%)	HR Minimum	HR Maximum	Différence relative des log-HR (%)	Largeur relative des IC 95%
Cohorte	1,229 (1,131 – 1,335)	-	-	Réf.	Réf.
CTN (1 :1) Ef	1,217 (1,088 – 1,361)	1,113	1,344	-5,07	1,333
CTN (1 :1) RB	1,236 (1,098 – 1,393)	1,125	1,383	2,61	1,444
CTN (5 :1) Ef	1,225 (1,120 – 1,341)	1,173	1,297	-1,39	1,081
CTN (5 :1) RB	1,228 (1,121 – 1,345)	1,175	1,302	-0,40	1,096
Schéma d'étude	Utilisation variable dans le temps des THM				
	HR (IC 95%)	HR Minimum	HR Maximum	Différence relative des log-HR (%)	Largeur relative des IC 95%
Cohorte	1,494 (1,364 – 1,637)	-	-	Réf.	Réf.
CTN (1 :1) Ef	1,470 (1,303 – 1,658)	1,282	1,636	-4,36	1,300
CTN (1 :1) RB	1,501 (1,322 – 1,704)	1,297	1,679	0,90	1,398
CTN (5 :1) Ef	1,497 (1,356 – 1,651)	1,429	1,572	0,38	1,078
CTN (5 :1) RB	1,501 (1,359 – 1,658)	1,432	1,580	1,10	1,092

HR (IC 95 %) *hazard ratio* (avec un intervalle de confiance à 95 %) ; CTN cas-témoins niché ; Ef Efron ; RB réduction de biais ; résultats CTN pour 100 échantillonnages indépendants ; les résultats des lignes grisées sont les mêmes que dans le tableau 3.5 et sont inclus ici pour faciliter la comparaison.

3.3 Prise en compte des ex-æquos dans les analyses de la cohorte entière et cas-témoins nichées

3.3.1 Données simulées

3.3.1.1 Exposition fixe dans le temps

Pour une exposition fixe dans le temps et un HR=2, la figure 3.3 représente le biais relatif des estimations du log-HR obtenues à partir des analyses CTN avec les approximations de Breslow et d'Efron et la méthode exacte pour la gestion des temps

d'évènement ex-æquos dans les strates des échantillons CTN. Le tableau 3.12 complète ces résultats avec la moyenne des écarts-types estimés, l'erreur quadratique moyenne et la probabilité de couverture.

Nous avons constaté que la fréquence des temps d'évènement ex-æquos était corrélée à la proportion d'évènements non censurés, allant d'une valeur médiane de 15 % à 53 % respectivement pour 5 % à 25 % de sujets qui ont observé un évènement, quelle que soit la prévalence de l'exposition (voir tableau 3.11).

Dans tous les scénarios, les estimations du log-HR étaient sans biais (biais relatif <1 %) pour l'analyse de la cohorte entière avec les approximations d'Efron ou de Breslow (tableau 3.12). En revanche, les estimations du log-HR pour les analyses CTN présentaient un biais négatif systématique, souvent important, avec l'approximation d'Efron et encore plus avec l'approximation de Breslow (figure 3.3 en A, B et C, D respectivement). Ce biais était beaucoup plus prononcé avec l'appariement 1 :1 qu'avec l'appariement 5 :1 (par exemple, le biais relatif maximal était de 41 % pour 1 témoin par cas contre 12 % pour 5 témoins par cas avec l'approximation de Breslow, figure 3.3 en A et B respectivement). Ce biais augmentait sensiblement avec la proportion d'évènements (par exemple, pour l'approximation de Breslow et 25 % de sujets exposés, le biais relatif était de 15 %, 25 % et 38 % respectivement pour 5 %, 10 % et 25 % de sujets ayant subi un évènement, figure 3.3 en A). De plus, le biais diminuait légèrement avec l'augmentation de la proportion de sujets exposés (par exemple, pour l'approximation d'Efron et 25 % de sujets ayant subi un évènement, le biais relatif était de 27 %, 24 % et 21 % pour respectivement 10 %, 25 % et 50 % de sujets exposés, figure 3.3 en C). Ces biais étaient presque entièrement éliminés lorsqu'on utilisait la méthode exacte pour la gestion des ex-æquos ou l'approche *ccwc* modifiée. En effet, le biais relatif maximal était seulement de 3 % pour les deux méthodes (figure 3.3 en E, tableau 3.12) contre 41 % et 27 % pour les approximations de Breslow et d'Efron, respectivement (figure 3.3 en A et C respectivement, tableau 3.12).

Dans tous les scénarios, les estimations basées sur l'analyse de la cohorte entière ont présenté une variance systématiquement plus faible que celles des analyses CTN (voir les colonnes SEE des écarts-types moyens estimés dans le tableau 3.12). Comme prévu, la précision des estimations s'est améliorée de façon générale pour les analyses des deux schémas d'étude à mesure que la proportion d'évènements, la prévalence de l'exposition et (pour les analyses CTN) le nombre de témoins par cas augmentaient. La précision des estimations issues de l'analyse CTN avec la méthode exacte ou l'approche *ccwc* modifiée était systématiquement inférieure à celle des estimations de l'analyse CTN avec l'approximation de Breslow ou d'Efron. Par conséquent, le compromis biais-variance n'était en faveur de la méthode exacte ou de l'approche *ccwc* modifiée que pour une plus grande proportion d'évènements, c'est-à-dire 25 % et peut-être 10 % (voir les colonnes

RMSE dans le tableau 3.12).

Dans tous les scénarios, nous avons observé des niveaux nominaux (0,95) pour la probabilité de couverture du schéma de cohorte tandis que le schéma d'étude CTN présentait des probabilités de couverture de niveaux inférieurs à la valeur nominale, en particulier pour les approximations de Breslow et d'Efron avec des proportions d'évènements plus élevées (scénarios pour lesquels des biais considérables ont été observés).

Les résultats de la simulation pour HR=1,25 étaient globalement similaires à ceux pour HR=2. Les estimations du log-HR pour l'analyse de la cohorte entière (avec les approximations de Breslow ou d'Efron) étaient non biaisées sauf pour le scénario avec 10 % de sujets exposés et 5 % d'évènements où on a observé un léger biais négatif de 5 % (tableau 3.13). Bien que leur ampleur soit quelque peu réduite, un biais négatif augmentant avec une proportion plus élevée d'évènements a de nouveau été observé dans les analyses CTN utilisant l'approximation d'Efron ou de Breslow. Cependant, comme leur biais était plus faible, le compromis biais-variance est devenu plus favorable et les probabilités de couverture étaient plus proches de la valeur nominale pour ces deux méthodes d'approximation avec un HR=1,25 (tableau 3.13).

Tableau 3.11 – *Distribution des temps d'évènements pour 1000 ensembles de données simulées en fonction des proportions d'évènements pour une exposition fixe dans le temps et un hazard ratio théorique de 2*

	5% d'évènements			10% d'évènements			25% d'évènements		
	Médiane	Min	Max	Médiane	Min	Max	Médiane	Min	Max
Nombre total d'évènements, M	249	202	313	521	456	593	1,265	1,173	1,363
Nombre de temps d'évènement distincts, K	211	174	259	373	331	413	600	567	636
Nombre de temps d'évènement uniques, M_1	177	149	213	255	211	296	223	178	263
Proportion des temps d'évènement ex-æquos, $(M\text{-}K)/M$	15%	8%	23%	28%	23%	33%	53%	50%	56%

Les résultats étaient identiques pour les trois proportions de sujets exposés considérées

Figure 3.3 – *Biais relatif du log-HR des analyses CTN (pour 1 et 5 témoins par cas) avec les approximations de Breslow (A et B respectivement) et d'Efron (C et D respectivement) et la méthode exacte (E et F respectivement) : exposition fixe dans le temps et hazard ratio théorique de 2*

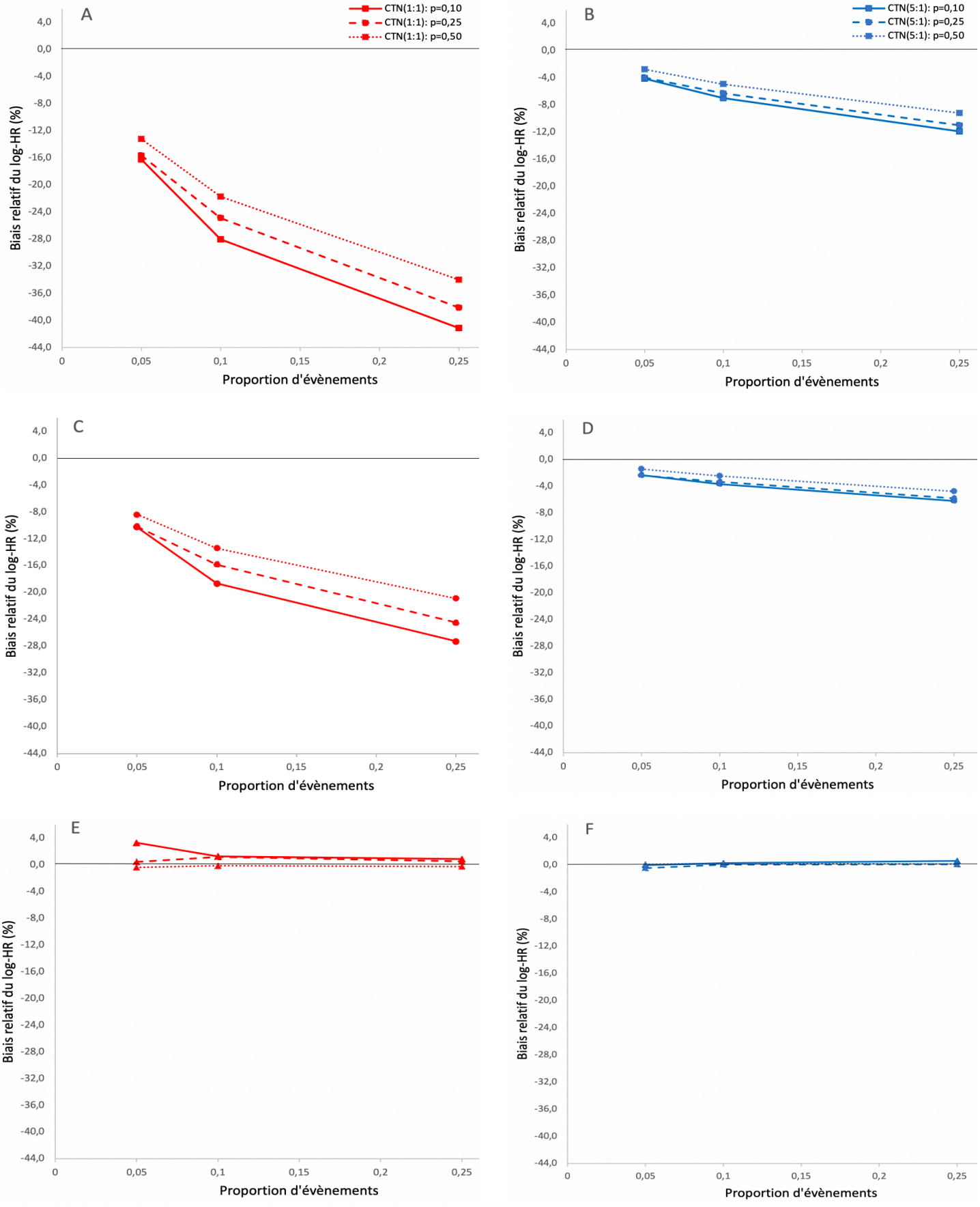


Tableau 3.12 – *Résultats des analyses de la cohorte entière (avec l'approximation d'Efron) et CTN (pour 1 et 5 témoins par cas) avec la méthode exacte, les approximations de Breslow et d'Efron et l'approche ccwc modifiée) pour une exposition fixe dans le temps et un hazard ratio théorique de 2*

10% sujets exposés												
Schéma d'étude	5% d'évènements				10% d'évènements				25% d'évènements			
	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)
Cohorte B	-0,641	0,167	0,166	95,3	0,190	0,116	0,113	95,5	0,379	0,076	0,075	95,5
Cohorte Ef	-0,636	0,167	0,166	95,3	0,200	0,116	0,113	95,5	0,406	0,076	0,076	95,4
CTN (1:1) B	-16,227	0,242	0,249	93,1	-28,039	0,153	0,231	80,5	-41,119	0,090	0,292	4,6
CTN (1:1) Ef	-10,256	0,244	0,251	93,8	-18,677	0,155	0,194	87,6	-27,263	0,091	0,205	45,7
CTN (1:1) Ex	3,326	0,278	0,293	94,7	1,309	0,191	0,194	95,2	0,878	0,125	0,124	94,6
CTN (1:1) *	3,059	0,279	0,284	95,3	1,105	0,192	0,188	96,2	1,135	0,127	0,129	94,0
CTN (5:1) B	-4,120	0,188	0,181	95,1	-6,987	0,127	0,129	95,5	-11,896	0,081	0,111	84,2
CTN (5:1) Ef	-2,311	0,188	0,184	95,0	-3,673	0,128	0,128	95,4	-6,190	0,081	0,092	91,6
CTN (5:1) Ex	-0,022	0,193	0,189	95,1	0,274	0,134	0,131	95,3	0,610	0,088	0,089	95,4
CTN (5:1) *	-0,591	0,193	0,188	95,9	0,614	0,134	0,126	96,0	1,062	0,088	0,087	95,2
25% sujets exposés												
Schéma d'étude	5% d'évènements				10% d'évènements				25% d'évènements			
	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)
Cohorte B	-0,502	0,130	0,129	95,4	-0,085	0,090	0,089	95,8	0,113	0,058	0,057	96,0
Cohorte Ef	-0,498	0,130	0,129	95,4	-0,076	0,090	0,089	95,8	0,137	0,058	0,057	96,0
CTN (1:1) B	-15,659	0,181	0,195	93,2	-24,881	0,116	0,197	72,8	-38,109	0,068	0,269	0,2
CTN (1:1) Ef	-10,193	0,182	0,187	94,0	-15,829	0,117	0,154	86,1	-24,480	0,068	0,181	28,4
CTN (1:1) Ex	0,481	0,202	0,201	95,3	1,209	0,139	0,136	95,9	0,550	0,090	0,088	94,8
CTN (1:1) *	1,640	0,203	0,200	96,1	0,731	0,140	0,139	95,1	0,887	0,091	0,091	95,7
CTN (5:1) B	-3,948	0,143	0,144	95,4	-6,259	0,098	0,102	94,0	-10,972	0,061	0,095	77,8
CTN (5:1) Ef	-2,361	0,143	0,145	95,4	-3,328	0,098	0,098	95,0	-5,812	0,061	0,073	91,0
CTN (5:1) Ex	-0,472	0,146	0,147	94,9	0,046	0,101	0,099	96,1	0,100	0,066	0,065	94,3
CTN (5:1) *	-0,007	0,146	0,146	95,5	0,319	0,101	0,101	95,2	0,188	0,066	0,065	96,2

Tableau 3.12, suite

50% sujets exposés												
Schéma d'étude	5% d'évènements				10% d'évènements				25% d'évènements			
	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)
Cohorte B	0,204	0,135	0,137	94,8	0,228	0,093	0,091	94,8	0,069	0,059	0,059	95,6
Cohorte Ef	0,207	0,135	0,137	94,8	0,235	0,093	0,091	94,8	0,089	0,059	0,059	95,5
CTN (1:1) B	-13,234	0,177	0,192	92,6	-21,738	0,115	0,180	76,0	-33,994	0,067	0,242	2,8
CTN (1:1) Ef	-8,408	0,177	0,187	93,9	-13,388	0,116	0,143	88,8	-20,890	0,068	0,158	42,1
CTN (1:1) Ex	-0,345	0,190	0,194	94,9	-0,110	0,130	0,126	95,9	-0,227	0,082	0,081	95,1
CTN (1:1) *	0,557	0,191	0,198	93,9	-0,523	0,131	0,130	95,6	0,277	0,084	0,083	95,8
CTN (5:1) B	-2,741	0,144	0,146	94,9	-4,922	0,098	0,099	94,5	-9,192	0,061	0,087	82,5
CTN (5:1) Ef	-1,397	0,144	0,147	94,5	-2,446	0,098	0,097	95,4	-4,752	0,061	0,070	91,5
CTN (5:1) Ex	0,093	0,146	0,149	94,0	0,292	0,101	0,098	96,0	0,119	0,064	0,065	94,9
CTN (5:1) *	0,342	0,146	0,147	94,9	0,221	0,101	0,100	95,3	-0,046	0,064	0,065	94,7

SEE *average standard error estimator* ; RMSE *root mean squared error* ; PC probabilité de couverture ; CTN cas-témoins niché ; B Breslow ; Ef Efron ; Ex Exact ; * approche *ccwc* modifiée ; les résultats des lignes grisées sont les mêmes que dans le tableau 3.1 et sont inclus ici pour faciliter la comparaison.

Tableau 3.13 – *Résultats des analyses de la cohorte entière (avec l'approximation d'Efron) et CTN (pour 1 et 5 témoins par cas) avec la méthode exacte, les approximations de Breslow et d'Efron et l'approche ccwc modifiée) pour une exposition fixe dans le temps et un hazard ratio théorique de 1,25*

10% sujets exposés												
Schéma d'étude	5% d'évènements				10% d'évènements				25% d'évènements			
	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)
Cohorte B	-5,335	0,196	0,204	94,6	0,520	0,135	0,136	94,3	1,674	0,087	0,087	94,7
Cohorte Ef	-5,331	0,196	0,204	94,6	0,529	0,135	0,136	94,3	1,697	0,087	0,087	94,7
CTN (1:1) B	-12,671	0,267	0,251	96,3	-22,057	0,172	0,155	96,9	-35,066	0,101	0,113	92,7
CTN (1:1) Ef	-7,078	0,269	0,267	95,7	-12,308	0,173	0,169	95,8	-21,038	0,102	0,111	93,2
CTN (1:1) Ex	5,851	0,293	0,300	95,7	5,051	0,201	0,201	94,4	4,004	0,129	0,135	94,3
CTN (1:1) *	4,955	0,293	0,300	95,2	2,498	0,201	0,208	94,5	2,194	0,129	0,129	96,4
CTN (5:1) B	-6,708	0,215	0,222	94,3	-5,176	0,145	0,140	95,9	-7,890	0,091	0,087	96,0
CTN (5:1) Ef	-5,243	0,215	0,225	94,0	-2,251	0,145	0,144	95,6	-2,824	0,091	0,091	95,2
CTN (5:1) Ex	-3,206	0,218	0,229	93,8	2,043	0,150	0,149	95,7	2,970	0,097	0,097	94,9
CTN (5:1) *	-4,253	0,218	0,225	94,5	2,392	0,150	0,150	95,7	1,606	0,097	0,096	95,4
25% sujets exposés												
Schéma d'étude	5% d'évènements				10% d'évènements				25% d'évènements			
	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)
Cohorte B	-0,868	0,140	0,139	94,8	0,158	0,097	0,099	94,3	0,541	0,062	0,062	96,4
Cohorte Ef	-0,865	0,140	0,139	94,8	0,166	0,097	0,099	94,3	0,562	0,062	0,062	96,4
CTN (1:1) B	-13,105	0,188	0,178	95,9	-20,948	0,122	0,120	95,8	-34,881	0,072	0,097	88,0
CTN (1:1) Ef	-7,751	0,189	0,188	94,9	-11,707	0,123	0,126	94,3	-21,183	0,072	0,085	91,1
CTN (1:1) Ex	2,036	0,204	0,207	94,6	3,894	0,141	0,147	94,7	1,832	0,090	0,093	94,3
CTN (1:1) *	1,440	0,204	0,200	95,6	2,763	0,141	0,143	94,6	3,986	0,091	0,091	95,1
CTN (5:1) B	-2,674	0,152	0,145	96,6	-4,796	0,103	0,102	95,3	-8,862	0,065	0,064	95,7
CTN (5:1) Ef	-1,206	0,152	0,147	96,1	-2,046	0,104	0,105	95,0	-4,071	0,065	0,065	95,4
CTN (5:1) Ex	0,549	0,155	0,150	95,8	1,025	0,107	0,108	95,0	1,345	0,069	0,068	95,1
CTN (5:1) *	-0,772	0,155	0,155	94,9	0,858	0,107	0,108	94,8	1,208	0,069	0,067	95,4

Tableau 3.13, suite

50% sujets exposés												
Schéma d'étude	5% d'évènements				10% d'évènements				25% d'évènements			
	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)
Cohorte B	1,917	0,128	0,128	95,4	1,787	0,088	0,087	95,0	0,559	0,057	0,057	95,1
Cohorte Ef	1,921	0,128	0,128	95,4	1,795	0,088	0,087	95,0	0,579	0,057	0,057	95,0
CTN (1:1) B	-12,072	0,169	0,157	96,0	-22,283	0,110	0,108	95,1	-34,468	0,065	0,094	81,8
CTN (1:1) Ef	-6,928	0,169	0,165	94,7	-13,706	0,111	0,111	94,5	-21,156	0,065	0,080	88,3
CTN (1:1) Ex	1,358	0,181	0,179	94,7	-0,199	0,125	0,123	94,9	-0,108	0,080	0,082	94,7
CTN (1:1) *	-0,782	0,182	0,179	94,9	1,389	0,125	0,122	95,4	1,597	0,080	0,081	94,4
CTN (5:1) B	-1,025	0,138	0,137	94,9	-3,720	0,094	0,090	95,7	-8,466	0,059	0,060	94,3
CTN (5:1) Ef	0,356	0,138	0,139	94,5	-1,175	0,094	0,092	95,2	-3,902	0,059	0,060	94,7
CTN (5:1) Ex	1,972	0,140	0,141	94,7	1,725	0,097	0,094	95,2	1,130	0,062	0,062	95,3
CTN (5:1) *	2,313	0,140	0,140	95,2	1,873	0,097	0,094	94,7	0,371	0,062	0,062	96,6

SEE *average standard error estimator* ; RMSE *root mean squared error* ; PC probabilité de couverture ; CTN cas-témoins niché ; B Breslow ;
Ef Efron ; Ex Exact ; * approche *ccwc* modifiée ; les résultats des lignes grisées sont les mêmes que dans le tableau 3.2 et sont inclus ici pour faciliter la comparaison.

3.3.1.2 Exposition variable dans le temps

Les résultats pour une exposition variable dans le temps étaient similaires à ceux obtenus pour une exposition fixe, que ce soit avec un HR=2 (tableau 3.15) ou un HR=1,25 (tableau 3.16). La figure 3.4 représente les biais relatifs des estimations du log-HR obtenues à partir des analyses CTN avec les approximations de Breslow et d'Efron et la méthode exacte.

La fréquence des temps d'évènement ex-æquos était comparable à celle observée avec une exposition fixe dans les neuf scénarios pour HR=2 (tableau 3.14) considérés.

L'analyse de la cohorte entière avec les approximations de Breslow et d'Efron a donné des estimations non biaisées (biais relatif <2,5 %, tableau 3.15), tandis que les estimations issues des analyses CTN avec les approximations de Breslow et d'Efron étaient à nouveau biaisées, parfois fortement (biais relatif maximal de 41 % pour un témoin par cas et 12 % pour 5 témoins par cas pour l'approximation de Breslow, figure 3.4 en A et B respectivement). La même tendance d'un biais croissant à mesure que la proportion d'évènements augmente était apparente (par exemple, pour l'approximation d'Efron et pour 10 % de sujets exposés, le biais est passé de 12 % à 19 % puis à 27 % pour respectivement 5 %, 10 % et 25 % de sujets ayant subi un évènement, figure 3.4 en C) . Il n'y avait cependant pas de tendance claire avec la proportion de sujets exposés. L'approximation de Breslow pour la gestion des temps d'évènement ex-æquos a entraîné un biais plus important dans les estimations du schéma CTN que l'approximation d'Efron dans tous les scénarios (figure 3.4 en A, B et C, D respectivement, tableau 3.15). Ces biais ont été considérablement réduits avec l'utilisation de la méthode exacte ou de l'approche *ccwc* modifiée avec un seul cas dans chaque strate (biais relatif maximal de 41 % et 27 % pour les approximations de Breslow et d'Efron respectivement à 2 % pour la méthode exacte et l'approche *ccwc* modifiée, figure 3.4 en A, C et E respectivement, tableau 3.15).

Le tableau 3.15 montre des écarts-types moyens estimés (SEE) et des erreurs quadratiques moyennes (RMSE) plus faibles pour les estimations issues de l'analyse de la cohorte entière que pour celles de l'analyse CTN dans tous les scénarios. La précision des estimations s'est améliorée pour les deux schémas d'étude à mesure que la proportion d'évènements, la prévalence de l'exposition et, pour les analyses CTN, le nombre de témoins par cas augmentaient. Comme avec l'exposition fixe dans le temps, pour l'analyse CTN avec les approximations de Breslow et d'Efron où les estimations ont produit un biais plus important, les RMSE étaient plus grandes pour une proportion plus élevée d'évènements et un témoin par cas et les SEE étaient systématiquement plus petites que celles de l'analyse CTN avec la méthode exacte ou l'approche *ccwc* modifiée (tableau 3.15).

De même que ce que nous avons observé pour une exposition fixe, la probabilité de couverture des estimations de la cohorte était proche de 95 % pour tous les scénarios, tandis que celle des estimations CTN était en dessous de la valeur nominale, en particulier pour les approximations de Breslow et d'Efron avec des proportions plus élevées d'évènements et de sujets exposés (scénarios pour lesquels les biais les plus importants ont été observés).

Les résultats des simulations avec un HR=1,25 révèlent que les estimations du log-HR pour l'analyse de la cohorte entière étaient non biaisées. Cependant, pour les scénarios avec 5 % d'évènements, on a observé un biais négatif compris entre 5 % et 9 % (tableau 3.16). Nous avons également observé un biais négatif (légèrement plus faible que celui obtenu pour un HR=2) qui augmentait avec la proportion d'évènements (pour l'approximation de Breslow, 25 % de sujets exposés et un témoin par cas, le biais relatif passait de 18 % à 37 % pour 5 % et 25 % d'évènements respectivement) dans les résultats des analyses CTN (tableau 3.16).

Tableau 3.14 – *Distribution des temps d'évènements pour 1000 ensembles de données simulées en fonction des proportions d'évènements et de sujets exposés pour une exposition variable dans le temps et un hazard ratio théorique de 2*

10% sujets exposés									
	5% d'évènements			10% d'évènements			25% d'évènements		
	Médiane	Min	Max	Médiane	Min	Max	Médiane	Min	Max
Nombre total d'évènements, M	249	202	314	521	462	592	1,265	1,172	1,354
Nombre de temps d'évènement distincts, K	211	174	259	373	334	420	600	569	630
Nombre de temps d'évènement uniques, M_1	178	147	212	255	218	298	223	182	259
Proportion des temps d'évènement ex-æquos, $(M-K)/M$	15%	8%	22%	28%	23%	33%	53%	49%	56%
25% sujets exposés									
	5% d'évènements			10% d'évènements			25% d'évènements		
	Médiane	Min	Max	Médiane	Min	Max	Médiane	Min	Max
Nombre total d'évènements, M	249	202	302	522	453	585	1,265	1,179	1,366
Nombre de temps d'évènement distincts, K	211	175	251	372	330	416	600	571	631
Nombre de temps d'évènement uniques, M_1	177	147	216	255	212	300	223	184	263
Proportion des temps d'évènement ex-æquos, $(M-K)/M$	15%	8%	23%	28%	22%	35%	53%	49%	56%
50% sujets exposés									
	5% d'évènements			10% d'évènements			25% d'évènements		
	Médiane	Min	Max	Médiane	Min	Max	Médiane	Min	Max
Nombre total d'évènements, M	249	204	300	522	451	582	1,266	1,171	1,367
Nombre de temps d'évènement distincts, K	211	174	248	372	332	407	600	566	628
Nombre de temps d'évènement uniques, M_1	177	137	213	255	215	293	223	184	264
Proportion des temps d'évènement ex-æquos, $(M-K)/M$	15%	8%	23%	29%	23%	35%	53%	50%	56%

Figure 3.4 – *Biais relatif du log-HR des analyses CTN (pour 1 et 5 témoins par cas) avec les approximations de Breslow (A et B respectivement) et d'Efron (C et D respectivement) et la méthode exacte (E et F respectivement) : exposition variable dans le temps et hazard ratio théorique de 2*

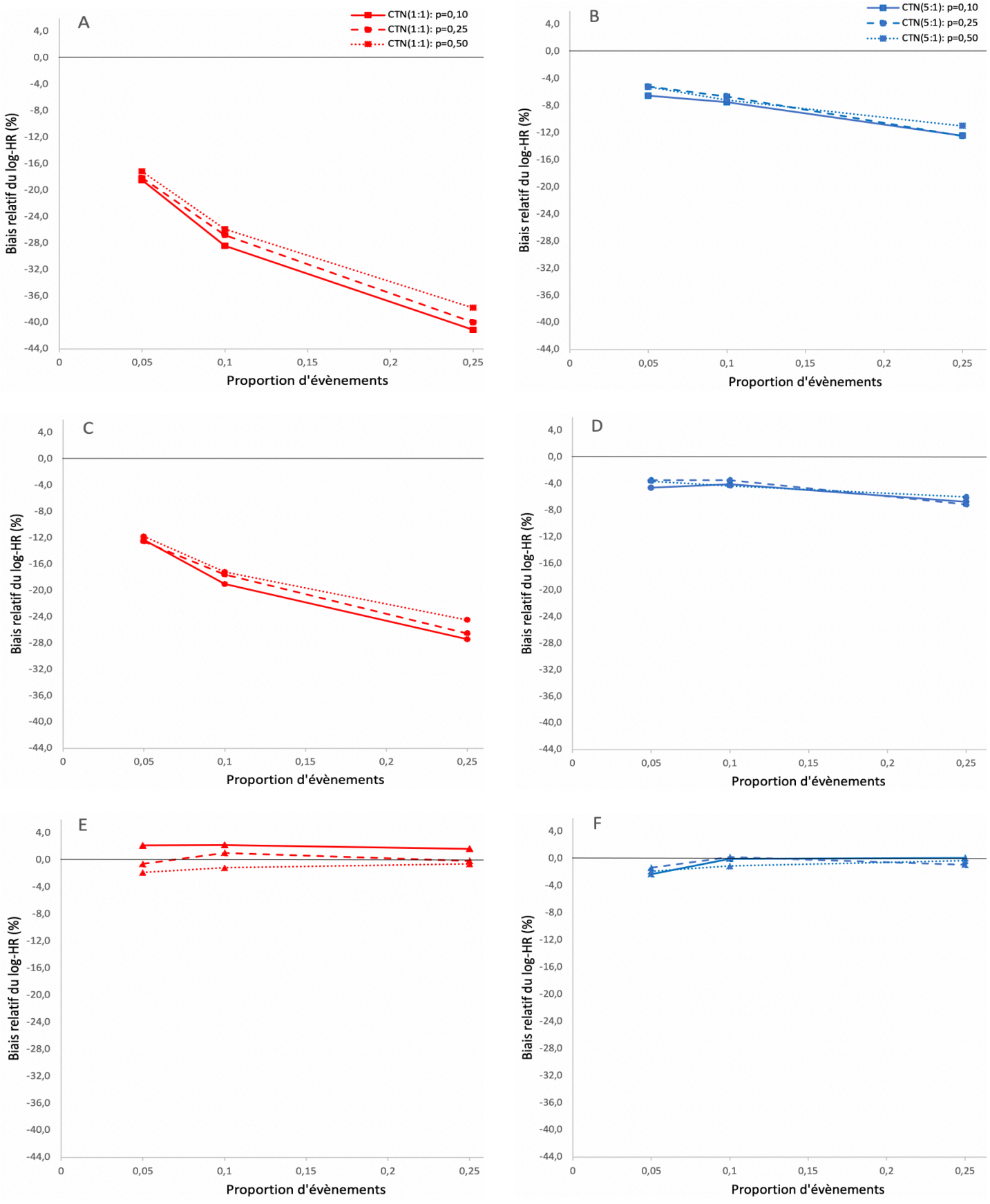


Tableau 3.15 – *Résultats des analyses de la cohorte entière (avec l'approximation d'Efron) et CTN (pour 1 et 5 témoins par cas) avec la méthode exacte, les approximations de Breslow et d'Efron et l'approche ccwc modifiée) pour une exposition variable dans le temps et un hazard ratio théorique de 2*

10% sujets exposés												
Schéma d'étude	5% d'évènements				10% d'évènements				25% d'évènements			
	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)
Cohorte B	-2,333	0,226	0,226	96,1	-0,621	0,157	0,146	96,9	0,478	0,104	0,100	96,8
Cohorte Ef	-2,328	0,226	0,226	96,1	-0,611	0,157	0,146	96,9	0,506	0,104	0,100	96,7
CTN (1:1) B	-18,475	0,331	0,318	95,9	-28,404	0,210	0,258	89,5	-41,108	0,123	0,297	29,2
CTN (1:1) Ef	-12,283	0,335	0,328	95,2	-19,010	0,213	0,232	92,4	-27,395	0,125	0,218	70,6
CTN (1:1) Ex	2,200	0,385	0,390	96,0	2,265	0,265	0,264	96,0	1,689	0,174	0,168	96,0
CTN (1:1) *	2,188	0,385	0,383	96,1	2,061	0,266	0,265	95,5	1,602	0,175	0,182	94,7
CTN (5:1) B	-6,503	0,256	0,259	94,7	-7,464	0,174	0,164	96,7	-12,415	0,111	0,130	92,0
CTN (5:1) Ef	-4,591	0,256	0,264	94,1	-4,080	0,174	0,165	96,5	-6,716	0,111	0,116	95,3
CTN (5:1) Ex	-2,263	0,263	0,271	93,7	-0,003	0,183	0,172	96,1	0,127	0,121	0,116	96,5
CTN (5:1) *	-1,924	0,263	0,259	95,8	-0,526	0,183	0,173	96,8	0,481	0,121	0,123	95,7
25% sujets exposés												
Schéma d'étude	5% d'évènements				10% d'évènements				25% d'évènements			
	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)
Cohorte B	-1,620	0,163	0,170	94,2	0,173	0,113	0,113	95,1	-0,845	0,074	0,074	94,7
Cohorte Ef	-1,616	0,163	0,170	94,2	0,183	0,113	0,113	95,1	-0,820	0,074	0,074	94,5
CTN (1:1) B	-18,136	0,232	0,252	92,4	-26,770	0,148	0,223	79,6	-39,979	0,087	0,284	4,0
CTN (1:1) Ef	-12,551	0,234	0,251	93,0	-17,568	0,150	0,186	88,3	-26,514	0,088	0,200	45,9
CTN (1:1) Ex	-0,552	0,263	0,279	93,8	1,085	0,182	0,185	94,9	-0,093	0,119	0,116	95,6
CTN (1:1) *	0,992	0,265	0,273	94,2	2,183	0,183	0,187	95,4	-0,482	0,120	0,119	94,7
CTN (5:1) B	-5,158	0,182	0,189	94,3	-6,632	0,123	0,127	94,0	-12,451	0,078	0,112	84,5
CTN (5:1) Ef	-3,458	0,182	0,191	94,3	-3,471	0,124	0,125	94,9	-7,119	0,079	0,091	90,8
CTN (5:1) Ex	-1,344	0,186	0,195	94,3	0,226	0,129	0,129	94,5	-0,887	0,085	0,083	94,4
CTN (5:1) *	-1,313	0,186	0,192	94,3	0,074	0,129	0,132	94,6	-1,081	0,085	0,087	94,1

Tableau 3.15, suite

50% sujets exposés												
Schéma d'étude	5% d'évènements				10% d'évènements				25% d'évènements			
	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)
Cohorte B	-1,770	0,143	0,147	94,0	-0,947	0,099	0,099	94,3	-0,067	0,064	0,066	94,6
Cohorte Ef	-1,766	0,143	0,147	94,0	-0,938	0,099	0,099	94,3	-0,044	0,064	0,066	94,6
CTN (1:1) B	-17,164	0,196	0,217	91,7	-25,875	0,127	0,210	72,6	-37,784	0,074	0,268	2,3
CTN (1:1) Ef	-11,772	0,198	0,211	92,6	-17,178	0,128	0,171	85,7	-24,454	0,075	0,184	37,3
CTN (1:1) Ex	-1,807	0,218	0,221	94,7	-1,127	0,150	0,151	94,2	-0,543	0,097	0,099	94,4
CTN (1:1) *	-0,228	0,220	0,220	95,7	-0,444	0,152	0,151	94,8	-0,233	0,098	0,097	95,7
CTN (5:1) B	-5,204	0,157	0,159	93,9	-7,162	0,107	0,115	92,9	-10,977	0,067	0,100	79,4
CTN (5:1) Ef	-3,611	0,157	0,160	93,8	-4,305	0,107	0,111	94,1	-5,983	0,067	0,081	89,2
CTN (5:1) Ex	-1,843	0,160	0,162	93,7	-1,079	0,111	0,112	95,0	-0,287	0,072	0,074	94,4
CTN (5:1) *	-2,668	0,160	0,165	94,2	-1,209	0,111	0,108	95,8	-0,235	0,072	0,074	94,6

SEE *average standard error estimator* ; RMSE *root mean squared error* ; PC probabilité de couverture ; CTN cas-témoins niché ; B Breslow ;
Ef Efron ; Ex Exact ; * approche *ccwc* modifiée ; les résultats des lignes grisées sont les mêmes que dans le tableau 3.3 et sont inclus ici pour faciliter la comparaison.

Tableau 3.16 – *Résultats des analyses de la cohorte entière (avec l'approximation d'Efron) et CTN (pour 1 et 5 témoins par cas) avec la méthode exacte, les approximations de Breslow et d'Efron et l'approche ccwc modifiée) pour une exposition variable dans le temps et un hazard ratio théorique de 1,25*

10% sujets exposés												
Schéma d'étude	5% d'évènements				10% d'évènements				25% d'évènements			
	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)
Cohorte B	-8,617	0,273	0,286	95,2	-2,728	0,188	0,177	97,1	-0,492	0,123	0,123	95,3
Cohorte Ef	-8,613	0,273	0,287	95,2	-2,719	0,188	0,177	97,1	-0,470	0,123	0,123	95,3
CTN (1:1) B	-12,790	0,374	0,358	96,3	-25,603	0,240	0,213	97,6	-36,669	0,143	0,138	96,8
CTN (1:1) Ef	-6,424	0,376	0,382	95,6	-15,887	0,242	0,234	96,4	-23,160	0,144	0,145	95,4
CTN (1:1) Ex	9,574	0,413	0,436	94,9	2,756	0,281	0,282	95,8	2,068	0,183	0,182	95,1
CTN (1:1) *	8,916	0,415	0,447	94,9	11,889	0,283	0,278	96,4	1,857	0,183	0,188	95,2
CTN (5:1) B	-6,540	0,299	0,299	95,5	-5,828	0,203	0,189	97,1	-10,494	0,129	0,125	95,8
CTN (5:1) Ef	-4,907	0,299	0,304	95,3	-2,774	0,203	0,195	96,5	-5,576	0,129	0,131	94,7
CTN (5:1) Ex	-2,696	0,305	0,310	95,1	0,554	0,210	0,202	96,3	0,082	0,137	0,138	94,6
CTN (5:1) *	-5,320	0,304	0,316	96,1	-3,578	0,209	0,199	96,4	-1,479	0,137	0,138	95,2
25% sujets exposés												
Schéma d'étude	5% d'évènements				10% d'évènements				25% d'évènements			
	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)
Cohorte B	-5,924	0,185	0,184	95,7	-1,131	0,128	0,129	95,5	-2,552	0,084	0,085	94,4
Cohorte Ef	-5,921	0,185	0,184	95,7	-1,122	0,128	0,129	95,5	-2,531	0,084	0,085	94,3
CTN (1:1) B	-18,547	0,250	0,235	96,4	-25,853	0,163	0,152	96,3	-37,302	0,097	0,115	91,7
CTN (1:1) Ef	-13,185	0,252	0,250	95,1	-16,726	0,164	0,164	94,9	-23,891	0,098	0,110	92,0
CTN (1:1) Ex	-2,509	0,273	0,276	95,0	-1,371	0,189	0,189	95,1	-1,387	0,123	0,127	94,0
CTN (1:1) *	0,331	0,274	0,287	94,1	1,535	0,189	0,189	95,2	-0,792	0,123	0,125	94,9
CTN (5:1) B	-8,783	0,202	0,202	95,6	-6,492	0,138	0,136	96,1	-13,880	0,088	0,090	94,9
CTN (5:1) Ef	-7,342	0,202	0,206	95,4	-3,718	0,138	0,140	95,7	-9,278	0,088	0,091	94,6
CTN (5:1) Ex	-5,558	0,205	0,209	95,2	-0,586	0,142	0,144	95,6	-4,185	0,093	0,095	94,8
CTN (5:1) *	-5,777	0,205	0,201	95,3	-0,584	0,142	0,141	95,9	-3,252	0,093	0,094	94,2

Tableau 3.16, suite

50% sujets exposés												
Schéma d'étude	5% d'évènements				10% d'évènements				25% d'évènements			
	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)	Biais relatif (%)	SEE	RMSE	PC (%)
Cohorte B	-6,147	0,151	0,158	94,7	-1,645	0,104	0,105	94,8	-0,380	0,068	0,071	93,9
Cohorte Ef	-6,144	0,151	0,158	94,7	-1,637	0,104	0,105	94,8	-0,359	0,068	0,071	93,9
CTN (1:1) B	-17,945	0,202	0,198	95,3	-24,273	0,132	0,132	94,5	-35,610	0,078	0,103	87,1
CTN (1:1) Ef	-12,709	0,203	0,208	94,3	-15,293	0,132	0,140	93,7	-22,324	0,079	0,094	91,5
CTN (1:1) Ex	-3,622	0,219	0,228	94,3	-0,983	0,151	0,159	93,4	-0,405	0,098	0,103	94,4
CTN (1:1) *	-5,338	0,219	0,222	95,5	-1,120	0,151	0,155	94,9	0,493	0,098	0,100	94,7
CTN (5:1) B	-11,170	0,163	0,169	94,5	-7,818	0,112	0,110	94,9	-11,170	0,071	0,074	94,0
CTN (5:1) Ef	-9,748	0,163	0,171	94,4	-5,197	0,112	0,112	94,5	-6,585	0,071	0,075	94,8
CTN (5:1) Ex	-8,245	0,166	0,174	94,2	-2,284	0,115	0,115	94,8	-1,519	0,075	0,078	94,7
CTN (5:1) *	-6,293	0,166	0,177	92,9	-1,318	0,115	0,115	95,7	-0,511	0,075	0,077	94,7

SEE *average standard error estimator* ; RMSE *root mean squared error* ; PC probabilité de couverture ; CTN cas-témoins niché ; B Breslow ;
Ef Efron ; Ex Exact ; * approche *ccwc* modifiée ; les résultats des lignes grisées sont les mêmes que dans le tableau 3.4 et sont inclus ici pour faciliter la comparaison.

3.3.2 Données réelles

3.3.2.1 Analyses avec une unité de temps égale au jour

En considérant l'âge des participantes à l'étude E3N en jours, il y a eu 1 984 âges distincts auxquels au moins un évènement (diagnostic du cancer du sein invasif) a été observé. Parmi ces âges, des évènements ex-æquos se sont produits à 253 âges distincts mais le nombre d'ex-æquos était relativement faible avec au maximum 4 cas de cancer du sein survenant au même âge (tableau 3.17).

Le tableau 3.18 compare les estimations obtenues après l'analyse de la cohorte entière à la moyenne des 100 estimations obtenues à l'aide de différentes stratégies d'analyse des données CTN. Toutes les analyses de la cohorte entière et des données CTN ont montré que les expositions aux THM, qu'elles soient fixes ou variables dans le temps, étaient associées à une augmentation statistiquement significative, bien que modérée, du risque de cancer du sein (tableau 3.18). Les estimations moyennes des analyses CTN pour un témoin par cas utilisant la méthode exacte ou l'approche *ccwc* modifiée étaient plus proches de celles obtenues à partir de l'analyse de la cohorte entière que les estimations moyennes des analyses CTN obtenues à l'aide des approximations de Breslow ou d'Efron (différence relative du log-HR d'environ 1 % contre 5 % à 10 %, et <1 % contre >4 % respectivement pour l'utilisation fixe et variable dans le temps des THM, tableau 3.18). Les analyses CTN utilisant les approximations de Breslow et d'Efron ont ainsi révélé des associations un peu plus faibles que celles des analyses de la cohorte entière. Ces différences étaient beaucoup plus faibles pour toutes les approches analytiques des données CTN avec 5 témoins par cas.

Tableau 3.17 – *Répartition des âges d'évènement (en jours) en fonction du nombre de fois qu'ils apparaissent (nombre d'ex-æquos)*

Nombre d'ex-æquos à chaque âge	Nombre d'âges distincts	Nombre total d'âges ex-æquos au moment du diagnostic du cancer du sein
1 (Pas d'ex-æquo)	1731	-
2	232	464
3	18	54
4	3	12
Total	1984	530

Tableau 3.18 – *Analyse de l'association entre utilisation de THM et risque de cancer du sein dans la cohorte E3N selon un schéma CTN, le jour étant pris comme unité de temps et les temps ex-œquos étant gérés selon l'approximation de Breslow, la méthode exacte et l'approche ccwc modifiée*

Schéma d'étude	Utilisation fixe dans le temps des THM					Utilisation variable dans le temps des THM				
	HR (IC 95%)	HR Minimum	HR Maximum	Différence relative des log-HR (%)	Largeur relative des IC 95%	HR (IC 95%)	HR Minimum	HR Maximum	Différence relative des log-HR (%)	Largeur relative des IC 95%
Cohorte	1,229 (1,131 – 1,335)	-	-	Réf.	Réf.	1,494 (1,364 – 1,637)	-	-	Réf.	Réf.
CTN (1 :1) B	1,204 (1,077 – 1,346)	1,109	1,327	-10,12	1,315	1,439 (1,276 – 1,623)	1,264	1,598	-9,61	1,270
CTN (1 :1) Ef	1,217 (1,088 – 1,361)	1,113	1,344	-5,07	1,333	1,470 (1,303 – 1,658)	1,282	1,636	-4,36	1,300
CTN (1 :1) Ex	1,232 (1,095 – 1,387)	1,124	1,375	1,06	1,428	1,498 (1,320 – 1,700)	1,296	1,686	0,45	1,391
CTN (1 :1) *	1,227 (1,091 – 1,381)	1,079	1,353	-0,96	1,418	1,499 (1,321 – 1,702)	1,340	1,669	0,64	1,396
CTN (5 :1) B	1,222 (1,117 – 1,337)	1,170	1,292	-2,83	1,078	1,488 (1,349 – 1,642)	1,421	1,562	-1,06	1,072
CTN (5 :1) Ef	1,225 (1,120 – 1,341)	1,173	1,297	-1,39	1,081	1,497 (1,356 – 1,651)	1,429	1,572	0,38	1,078
CTN (5 :1) Ex	1,228 (1,121 – 1,335)	1,175	1,300	-0,42	1,096	1,501 (1,359 – 1,658)	1,433	1,578	1,18	1,092
CTN (5 :1) *	1,231 (1,124 – 1,348)	1,171	1,304	0,87	1,098	1,503 (1,361 – 1,660)	1,435	1,575	1,54	1,094

HR (IC 95 %) *hazard ratio* (avec un intervalle de confiance à 95 %); CTN cas-témoins niché; B Breslow; Ef Efron; Ex Exact; * approche *ccwc* modifiée;
résultats CTN pour 100 échantillonnages indépendants; les résultats des lignes grisées sont les mêmes que dans le tableau 3.5 et sont inclus ici pour faciliter la comparaison.

3.3.2.2 Analyses avec une unité de temps égale au mois

Lorsque l'âge des participantes a été considéré en mois, le nombre d'ex-æquos a nettement augmenté. Il y a eu 363 âges distincts auxquels au moins un évènement (diagnostic du cancer du sein invasif) a été observé. Parmi ceux-ci, les évènements ex-æquos sont survenus à 309 âges et le nombre d'ex-æquos était élevé, avec jusqu'à 20 cas de cancer du sein survenant au même âge (tableau 3.19).

Les estimations issues des analyses CTN avec les approximations de Breslow et d'Efron différaient considérablement de celles issues de l'analyse de la cohorte entière (différence relative du log-HR jusqu'à 45 % et 23 % respectivement, pour un témoin par cas et une utilisation fixe dans le temps des THM, tableau 3.20). L'utilisation des approximations de Breslow et d'Efron pour gérer les ex-æquos dans les analyses CTN a entraîné des biais importants par rapport aux estimations produites par les analyses de la cohorte entière (tableau 3.20). L'utilisation de la méthode exacte ou de l'approche *ccwc* modifiée pour les analyses CTN a considérablement réduit cette différence (la différence relative du log-HR est passée à 2 % et 4 % pour un témoin par cas et l'utilisation fixe dans le temps des THM, tableau 3.20).

Tableau 3.19 – *Répartition des âges d'évènement (en mois) en fonction du nombre de fois qu'ils apparaissent (nombre d'ex-æquos)*

Nombre d'ex-æquos à chaque âge	Nombre d'âges distincts	Nombre total d'âges ex-æquos au moment du diagnostic du cancer du sein
1 (Pas d'ex-æquo)	54	-
2	43	86
3	33	99
4	20	80
5	24	120
6	33	198
7	20	140
8	26	208
9	25	225
10	22	220
11	22	242
12	6	72
13	13	169
14	7	98
15	5	75
16	4	64
17	2	34
18	1	18
19	1	19
20	2	40
Total	363	2207

Tableau 3.20 – *Analyse de l'association entre utilisation de THM et risque de cancer du sein dans la cohorte E3N selon un schéma de cohorte ou CTN, le mois étant pris comme unité de temps et les temps ex-œquos étant gérés selon les approximations de Breslow et d'Efron, la méthode exacte et l'approche ccwc modifiée*

Schéma d'étude	Utilisation fixe dans le temps des THM					Utilisation variable dans le temps des THM				
	HR (IC 95%)	HR Minimum	HR Maximum	Différence relative des log-HR (%)	Largeur relative des IC 95%	HR (IC 95%)	HR Minimum	HR Maximum	Différence relative des log-HR (%)	Largeur relative des IC 95%
Cohorte	1,227 (1,129 – 1,333)	-	-	Réf.	Réf.	1,491 (1,361 – 1,634)	-	-	Réf.	Réf.
CTN (1 :1) B	1,119 (1,027 – 1,221)	1,063	1,178	-44,93	0,951	1,260 (1,146 – 1,386)	1,178	1,338	-42,28	0,880
CTN (1 :1) Ef	1,164 (1,067 – 1,270)	1,088	1,243	-25,77	0,994	1,360 (1,236 – 1,496)	1,242	1,466	-23,20	0,955
CTN (1 :1) Ex	1,232 (1,095 – 1,386)	1,119	1,354	1,65	1,425	1,510 (1,332 – 1,711)	1,341	1,677	2,87	1,391
CTN (1 :1) *	1,217 (1,082 – 1,370)	1,063	1,331	-4,35	1,409	1,500 (1,321 – 1,703)	1,340	1,676	1,09	1,398
CTN (5 :1) B	1,190 (1,094 – 1,295)	1,126	1,241	-14,86	0,983	1,414 (1,289 – 1,551)	1,365	1,472	-13,45	0,960
CTN (5 :1) Ef	1,208 (1,111 – 1,315)	1,139	1,264	-7,55	0,999	1,453 (1,324 – 1,594)	1,399	1,518	-6,61	0,988
CTN (5 :1) Ex	1,226 (1,119 – 1,343)	1,149	1,287	-0,46	1,095	1,493 (1,352 – 1,648)	1,434	1,564	0,16	1,086
CTN (5 :1) *	1,231 (1,124 – 1,348)	1,170	1,284	1,40	1,100	1,495 (1,354 – 1,651)	1,419	1,557	0,52	1,089

HR (IC 95 %) *hazard ratio* (avec un intervalle de confiance à 95 %); CTN cas-témoins niché ; B Breslow ; Ef Efron ; Ex Exact ; * approche *ccwc* modifiée ;
Résultats CTN pour 100 échantillonnages indépendants

4 - Discussion et perspectives

4.1 Synthèse des résultats

L'objectif de cette thèse était d'évaluer les méthodes statistiques disponibles pour modéliser les expositions médicamenteuses variables dans le temps. Plus spécifiquement, il était d'examiner les performances des schémas d'étude de cohorte et cas-témoins niché pour estimer l'association entre une exposition médicamenteuse et le risque d'évènement indésirable, en particulier pour une exposition variable dans le temps.

Les résultats de l'étude de simulation que nous avons réalisée ont montré que les estimations obtenues de l'analyse de la cohorte entière étaient non biaisées dans tous les scénarios considérés. Cependant, les estimations issues des analyses CTN (avec l'approximation d'Efron) étaient substantiellement biaisées, en particulier lorsqu'un seul témoin était apparié à chaque cas. Ce biais des estimations CTN augmentait avec la proportion d'évènements.

Pour comprendre d'où venaient les biais dans les résultats des analyses CTN et y remédier éventuellement, nous avons dans un premier temps analysé les données CTN avec la méthode de Firth adaptée à la régression logistique conditionnelle pour la réduction des biais. Nous avons constaté une diminution considérable des biais dans les estimations suggérant ainsi que les biais observés pourraient provenir de données éparses.

Cependant, certains arguments nous ont fait comprendre que cette explication n'était pas satisfaisante. En effet, après avoir augmenté la taille des cohortes simulées (résultats non présentés dans ce mémoire) et surtout quel que soit le nombre d'évènements, les biais persistaient alors qu'en principe plus la proportion d'évènements augmente plus le risque de données éparses diminue. Nous avons donc, dans un second temps, pris en compte les temps d'évènement ex-æquos (pouvant apparaître dans les strates) en utilisant différentes méthodes : les approximations de Breslow et d'Efron (cette dernière étant celle utilisée par défaut dans les premières analyses CTN et de la cohorte entière), la méthode exacte et l'approche *ccwc* modifiée. Nous avons constaté que les estimations CTN avec l'approximation de Breslow étaient davantage biaisées qu'avec l'approximation d'Efron, tandis qu'avec la méthode exacte ou l'approche *ccwc* modifiée les estimations étaient presque exemptes de biais et assez proches de celles de l'analyse de la cohorte entière. Il est ainsi apparu que les biais initialement observés résultaient vraisemblablement de la présence de temps d'évènement ex-æquos dans les strates. Ces ex-æquos sont présents dans la cohorte entière mais relativement "dilués" de sorte que leur prise en compte impacte peu l'estimation de l'effet de l'exposition. En revanche, avec la méthode

de sélection des témoins utilisée pour l'extraction des données CTN, ces ex-æquos se retrouvaient dans des strates réduites aux cas et à leurs témoins appariés, de sorte que les approximations de Breslow et d'Efron se trouvaient très éloignées de la vraisemblance qu'elles étaient censées approcher.

Quel que soit le scénario considéré, les estimations basées sur l'analyse de la cohorte entière ont présenté une précision systématiquement plus élevée que celles des analyses CTN. Les estimations issues de l'analyse de la cohorte entière avaient un taux de recouvrement satisfaisant tandis que celles issues de l'analyse CTN présentaient de faibles taux de recouvrement, en particulier avec les approximations de Breslow et d'Efron pour des scénarios où des biais considérables avaient été observés. Tous ces résultats ont été observés autant pour une exposition variable que fixe dans le temps.

Les résultats des analyses de la cohorte E3N entière et CTN dans cette cohorte pour évaluer l'association entre l'exposition aux traitements hormonaux de la ménopause et le risque de cancer du sein ont révélé que les expositions aux THM, qu'elles soient fixes ou variables dans le temps, étaient associées à une augmentation statistiquement significative du risque de cancer du sein. L'utilisation de l'approximation d'Efron pour gérer les ex-æquos dans les analyses CTN a entraîné des différences importantes dans les estimations par rapport à celles issues des analyses de la cohorte entière, en particulier pour un témoin par cas. Cependant, l'augmentation du nombre de témoins par cas ou la gestion des ex-æquos avec la méthode exacte ou l'approche *ccwc* modifiée dans les analyses CTN a considérablement réduit cette différence.

4.2 Comparaison des résultats à ceux de la littérature existante

Les résultats très biaisés obtenus après les analyses CTN (1:1) (avec l'approximation d'Efron ou de Breslow) des données simulées par rapport à ceux des analyses de la cohorte entière contredisent la majorité des résultats de la littérature pour lesquels les analyses de données selon ces deux schémas d'étude produisent des estimations presque semblables (Billioti de Gage et al., 2012 ; Breslow et al., 1983 ; Essebag et al., 2005 ; Hak et al., 2004 ; Liddell et al., 1977). Cependant, de tels résultats ont été retrouvés dans les quelques travaux (à notre connaissance) qui comparaient ces deux schémas d'étude à l'aide des données simulées avec une exposition fixe ou variable dans le temps (Austin et al., 2012 ; Bertke et al., 2013 ; Leffondré et al., 2003). L'utilisation de la la méthode de Firth adaptée à la régression logistique conditionnelle pour la réduction des biais dans les analyses CTN a considérablement réduit ces biais et a ainsi montré qu'ils pouvaient être dus aux données éparses. Ces résultats sont similaires à ceux de l'étude de Cheung et al. (2019) qui ont montré que des estimations biaisées obtenues dans les

analyses CTN étaient dues au biais des données éparses qui pouvait être contrôlé par cette méthode de réduction de biais pour la régression logistique conditionnelle (Cheung et al., 2019).

La perte de précision des estimations issues des analyses CTN par rapport à celles des analyses de la cohorte entière était attendue et corrobore avec les résultats de travaux antérieurs (Breslow et al., 1983 ; Essebag et al., 2005 ; Hak et al., 2004 ; Liddell et al., 1977) car, dans le modèle d'analyse de cohorte, toutes les données sont utilisées alors que, dans celui de l'analyse CTN, seules les données sur les cas et les témoins sélectionnés sont utilisées.

Le biais encore apparent avec 5 témoins par cas était cependant moins attendu et contradictoire avec l'idée générale selon laquelle les analyses CTN avec un nombre suffisant de témoins par cas fournissent des estimations proches de celles des analyses de cohortes entières (Bertke et al., 2013 ; Essebag et al., 2005 ; Goldstein & Langholz, 1992 ; Pang, 1999). De plus, de manière contre-intuitive, le biais a eu tendance à augmenter avec une plus grande proportion d'évènements, un résultat similaire à celui rapporté par Austin et al. (2012) alors qu'on pourrait s'attendre à une estimation plus précise lorsque la richesse de l'information augmente comme observé dans les résultats de l'étude de Bertke et al. (2013). Dans notre cas cependant, les biais observés résultaient très probablement d'évènements ex-æquos et de la façon dont ils étaient traités. Ces biais étaient en effet nettement réduits (au prix d'une précision légèrement inférieure pour une proportion plus faible d'évènements) lorsque la méthode exacte (Peto, 1972) ou l'approche *ccwc* modifiée étaient utilisées au lieu de l'approximation de Breslow ou d'Efron.

La tendance observée d'une augmentation des biais avec une plus grande proportion d'évènements avec les approximations de Breslow et d'Efron est cohérente avec les résultats de travaux antérieurs montrant que ces approximations se détériorent lorsque la proportion d'évènements ex-æquos augmente (Farewell & Prentice, 1980). Dans notre cadre de simulation, la probabilité que deux évènements ou plus se produisent en même temps dans un délai limité de 730 jours augmente avec la proportion d'évènements, c'est-à-dire le nombre de cas dans une cohorte de taille fixe de 5 000 individus. L'augmentation de la proportion d'évènements ex-æquos a eu un fort impact négatif sur les estimations des analyses des données CTN, alors que les estimations de l'analyse de la cohorte entière ont été à peine affectées et sont restées pratiquement sans biais. En effet, dans les analyses de cohorte entière, la proportion d'ex-æquos dans les ensembles à risque à chaque temps d'évènement est considérablement diluée parmi tous les sujets appartenant à l'ensemble à risque à ce temps-là, alors que dans les analyses CTN, l'ensemble à risque est beaucoup plus petit car limité aux cas survenant à ce temps spécifique de

l'évènement et à leurs témoins appariés. Auparavant, Farewell et Prentice (1980) avaient montré par des simulations de données cas-témoins avec 5, 20 et 50 cas (c'est-à-dire des évènements ex-æquos) par strate que les estimations des analyses CTN avec l'approximation d'Efron, et dans une plus large mesure encore avec l'approximation de Breslow, sont affectées par un biais de sous-estimation systématique. Ils ont donc conseillé de ne pas utiliser ces deux méthodes d'approximation pour les études cas-témoins ou de cohorte dans lesquelles le nombre d'ex-æquos par temps d'évènement est important. D'autres auteurs ont préconisé l'utilisation de la méthode exacte (Peto, 1972) pour analyser les études cas-témoins lorsque la proportion de cas dans chaque strate cas-témoins n'est pas suffisamment faible (Langholz & Richardson, 2010). Ce problème peut également être évité dans les études cas-témoins en disposant les ensembles cas-témoins de manière à obtenir un seul cas par strate et, ainsi, à ne pas avoir d'évènements ex-æquos dans les strates. Cette solution exige toutefois que chacun des cas survenant au même moment de l'évènement puisse se voir attribuer un nombre suffisant de témoins appariés. Cela a motivé notre modification de la fonction *ccwc* dans R, conformément à la procédure d'échantillonnage réalisée dans la macro SAS %nCCsampling (Desai et al., 2016).

Nous craignons que les chercheurs ne soient pas conscients de la détérioration potentielle des estimations issues de l'analyse CTN en présence de nombreux évènements ex-æquos et que ce problème ne soit pas suffisamment pris en compte dans la pratique. En particulier, les chercheurs devraient être conscients des limites possibles des options par défaut utilisées dans les logiciels statistiques disponibles pour mettre en œuvre la régression logistique conditionnelle. Dans R, alors que la fonction *clogit* utilise par défaut la méthode exacte, elle appelle la fonction *coxph* qui utilise par défaut l'approximation d'Efron. Dans SAS, la régression logistique conditionnelle peut être effectuée à l'aide de la procédure *PHREG* avec une instruction STRATA, qui utilise l'approximation de Breslow par défaut, ou de la procédure *LOGISTIC* avec une instruction STRATA. Cette dernière procédure s'appuie sur la méthode exacte pour les données discrètes en supposant que les évènements se sont réellement produits exactement au même moment (Allison, 2010). Cette procédure est appelée, par exemple, par la macro %nCCsampling qui effectue un échantillonnage de densité d'incidence pour l'analyse CTN (Desai et al., 2016). Dans nos données simulées et réelles, en particulier lorsqu'un mois a été pris comme unité de temps, les ex-æquos résultaient d'une mesure imprécise du temps continu. Par conséquent, il convient d'utiliser la méthode exacte en supposant qu'il existe un ordre réel mais inconnu pour les temps d'évènements ex-æquos. Le calcul de la vraisemblance exacte, qui prend en compte tous les ordres possibles des évènements ex-æquos, peut être chronophage (Allison, 2010) pour les analyses de la cohorte entière mais, selon notre expérience, le temps de calcul était raisonnable pour les analyses CTN.

4.3 Comparaison entre l'étude de simulation et l'application aux données de la cohorte E3N

Certains scénarios de simulation permettaient de se rapprocher de certaines caractéristiques des données réelles. En effet, pour les données de la cohorte E3N, on avait 45 % de femmes exposées aux THM en début de suivi pour une exposition fixe et environ 6 % de cancers du sein invasifs et pour les données simulées, on avait un scénario de 50 % d'exposés et 5 % d'évènements.

Les deux types d'exposition (fixe et variable dans le temps) utilisés dans les simulations et l'application aux données de la cohorte E3N, bien que semblables, étaient quel que peu différents. Dans les données réelles, le premier type utilisait une exposition aux THM fixe basée sur le statut de base (statut d'exposition aux THM au début du suivi), ignorant ainsi les changements possibles de l'exposition au cours du suivi alors que l'exposition est restée fixe tout au long du suivi dans les données simulées. En ce qui concerne le deuxième type d'exposition, une proportion non négligeable de femmes étaient déjà exposées au début du suivi dans les données E3N alors que tous les sujets étaient considérés comme non exposés au début de leur suivi dans les données simulées. Cependant, bien que les caractéristiques de l'exposition aux THM n'aient pas été tout à fait identiques à celles de l'exposition définie dans les simulations, les résultats obtenus sur les données de la cohorte E3N étaient conformes à ceux des analyses des données simulées, quel que soit le type d'exposition. En effet, comme dans les résultats obtenus avec les données simulées, les analyses CTN des données E3N utilisant les approximations de Breslow et d'Efron ont révélé des associations un peu plus faibles que celles des analyses de la cohorte entière. De plus, ces résultats des estimations s'amélioraient pour toutes les approches analytiques des données CTN avec 5 témoins par cas.

4.4 Limites de l'étude

Dans nos simulations, l'exposition variable dans le temps change au plus une seule fois au cours du suivi et nous avons utilisé le modèle d'exposition actuelle pour la quantifier. Dans la vie réelle, l'exposition médicamenteuse est susceptible de varier plusieurs fois au cours du temps et plusieurs autres modèles d'exposition permettent de mieux prendre en compte cette variation et ainsi de mieux évaluer son association avec un évènement d'intérêt (Abrahamowicz et al., 2012 ; Gasparrini, 2014 ; Pazzagli et al., 2018 ; Sylvestre & Abrahamowicz, 2009). De même, pour la cohorte E3N, l'exposition variable dans le temps aux THM considérée change une seule fois de "non exposée" à "exposée" au cours du suivi, pourtant les informations mises à jour sur l'utilisation des

THM disponibles dans cette cohorte peuvent nous permettre de considérer une exposition aux THM qui varie plusieurs fois sur la période d'étude.

Seuls les résultats pour les schémas d'étude classiques de cohorte et CTN sont présentés dans ce mémoire alors qu'il y a plusieurs autres schémas d'étude utilisés en pharmaco-épidémiologie, en particulier les schémas cas seuls qui éliminent les biais éventuels de sélection des sujets témoins et l'effet des facteurs de confusion invariables dans le temps (Farrington, 1995 ; Hallas, 1996 ; Maclure, 1991) et dont il serait intéressant d'étudier les propriétés statistiques.

Les scénarios de simulation considérés nous ont permis d'évaluer uniquement un effet délétère de l'exposition médicamenteuse. Il serait intéressant d'évaluer également les propriétés statistiques de ces schémas d'étude pour l'estimation d'un effet bénéfique qui a son importance en pharmaco-épidémiologie.

Les performances des différentes approches n'ont pas été évaluées avec l'appariement ou l'ajustement sur un ensemble de divers facteurs de confusion fixes et/ou variables dans le temps. L'ajustement ou l'appariement sur plusieurs variables est généralement utilisée pour réduire les biais de confusion. Cependant, les biais ont été observés dans l'étude d'Austin (2012) malgré l'ajustement des modèles d'analyse CTN sur des covariables fixes dans le temps (Austin et al., 2012). Qu'en serait t-il avec l'appariement sur des covariables ? De même, l'analyse des données E3N ne prend pas en compte l'ajustement ou l'appariement sur des facteurs de confusion ; pourtant cette cohorte contient des informations sur des covariables fixes ou variables dans le temps qui peuvent être utilisées.

Enfin, nous avons ignoré les risques concurrents dus à d'autres évènements empêchant la survenue de l'évènement d'intérêt autant dans les analyses des données simulées que dans celles des données de la cohorte E3N (dans laquelle on a d'autres évènements tels que les décès et les autres types de cancers).

4.5 Perspectives

Bien que notre étude se soit limitée à une exposition variable dans le temps avec un seul changement de "non exposé" à "exposé" pendant le suivi, des travaux ultérieurs devraient étudier des expositions variables dans le temps avec des changements multiples (Desquilbet & Meyer, 2005), représentées par des modèles de mesure plus sophistiqués et plus flexibles développés pour mieux prendre en compte la complexité de la relation entre l'exposition médicamenteuse et les évènements de santé. En particulier, le modèle d'exposition cumulée pondérée (Sylvestre & Abrahamowicz, 2009) présenté dans des études comme étant plus adéquat pour représenter une exposition variable dans le temps

(Abrahamowicz et al., 2012 ; Pazzagli et al., 2018) peut être utilisé. Les *Distributed Lag Non-linear Models* (Gasparrini, 2014), qui généralisent le modèle d'exposition cumulée pondérée (WCE pour *weighted cumulative exposure*) en combinant deux fonctions pour modéliser de manière flexible la relation exposition-réponse pour l'intensité de l'exposition et la relation temps-réponse pour le temps écoulé depuis l'exposition, peuvent également être utilisés. Une étude évaluant les effets des expositions variables dans le temps avec des rapports de risque décroissants (Brickner, 2015) sur la survenue d'un évènement de santé serait intéressant à explorer.

L'adaptation de ces modèles d'exposition à l'analyse de l'association entre l'utilisation des THM et le risque de cancer du sein, tout en tirant parti du long suivi de la cohorte E3N, pourrait apporter un éclairage nouveau sur l'évaluation des effets indésirables des THM.

La prise en compte de l'ajustement sur des facteurs de confusion fixes ou variables dans le temps devrait être abordée autant dans les analyses des données simulées que réelles. Cependant, dans une étude avec des expositions qui varient au cours du temps, les modèles statistiques standard ajustés sur des facteurs de confusion variables dans le temps peuvent conduire à des estimations biaisées de l'effet lorsque ces facteurs de confusion sont affectés par les niveaux d'exposition antérieurs, c'est-à-dire qu'ils agissent comme des intermédiaires dans la relation entre l'exposition et l'évènement d'intérêt (Robins et al., 2000). Dans ce cas, les modèles structuraux marginaux devraient être utilisés pour ajuster sur les facteurs de confusion variant dans le temps (Pazzagli et al., 2018 ; Robins et al., 2000).

Des travaux ultérieurs seraient également nécessaires pour examiner s'il pourrait y avoir un avantage de l'analyse CTN appariée sur des covariables par rapport à l'analyse de cohorte ajustée sur les mêmes covariables tout en tenant compte des facteurs de confusion potentiels. D'autres travaux seraient également justifiés pour comparer les schémas d'étude de cohorte et CTN en présence d'évènements concurrents (Austin et al., 2012).

Bibliographie

Aalen, O. (1978). Non-parametric inference for a family of counting processes. *The Annals of Statistics*, *6*(4), 701-726.

Abrahamowicz, M., Bartlett, G., Tamblyn, R. & du Berger, R. (2006). Modeling cumulative dose and exposure duration provided insights regarding the associations between benzodiazepines and injuries. *Journal of Clinical Epidemiology*, *59*(4), 393-403.

Abrahamowicz, M., Beauchamp, M.-E. & Sylvestre, M.-P. (2012). Comparison of alternative models for linking drug exposure with adverse effects. *Statistics in Medicine*, *31*(11-12), 1014-1030.

Abrahamowicz, M., Mackenzie, T. & Esdaile, J. M. (1996). Time-dependent hazard ratio : modeling and hypothesis testing with application in lupus nephritis. *Journal of the American Statistical Association*, *91*(436), 1432-1439.

Allison, P. D. (2010). *Survival analysis using SAS : a practical guide* (2nd ed). Cary, NC : SAS Institute Inc.

Austin, P. C. (2012). Generating survival times to simulate Cox proportional hazards models with time-varying covariates. *Statistics in Medicine*, *31*(29), 3946-3958.

Austin, P. C. (2013). Correction : 'Generating survival times to simulate Cox proportional hazards models with time-varying covariates'. *Statistics in Medicine*, *32*(6), 1078-1078.

Austin, P. C., Anderson, G. M., Cigsar, C. & Gruneir, A. (2012). Comparing the cohort design and the nested case–control design in the presence of both time-invariant and time-dependent treatment and competing risks : bias and precision. *Pharmacoepidemiology and Drug Safety*, *21*(7), 714-724.

Bagdonavicius, V. & Nikulin, M. (2001). *Accelerated life models : modeling and statistical analysis* (1re éd.). Chapman and Hall/CRC Press, Boca Raton, Florida.

Bégaud, B. (1998). *Dictionnaire de pharmaco-épidémiologie.* Association pour la Recherche Méthodologique en Pharmacovigilance, Bordeaux.

Bender, R., Augustin, T. & Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, *24*(11), 1713-1723.

Beral, V. & Million Women Study Collaborators. (2003). Breast cancer and hormone-replacement therapy in the Million Women Study. *Lancet*, *362*(9382), 419-427.

Bertke, S., Hein, M., Schubauer-Berigan, M. & Deddens, J. (2013). A simulation study of relative efficiency and bias in the nested case–control study design. *Epidemiologic Methods*, *2*(1), 85-93.

Billioti de Gage, S., Begaud, B., Bazin, F., Verdoux, H., Dartigues, J.-F., Peres, K., Kurth, T. & Pariente, A. (2012). Benzodiazepine use and risk of dementia : prospective population based study. *BMJ*, *345*, e6231-e6231.

Brauer, R., Ruigómez, A., Klungel, O., Reynolds, R., Feudjo Tepie, M., Smeeth, L. & Douglas, I. (2016). The risk of acute liver injury among users of antibiotic medications : a comparison of case-only studies. *Pharmacoepidemiology and Drug Safety*, *25*(Suppl 1), 39-46.

Breslow, N. E. & Day, N. E. (1980). *The analysis of case-control studies*. International Agency for Research on Cancer, Lyon.

Breslow, N. E. & Day, N. E. (1987). *The design and analysis of cohort studies*. International Agency for Research on Cancer, Lyon.

Breslow, N. E., Day, N. E., Halvorsen, K. T., Prentice, R. L. & Sabai, C. (1978). Estimation of multiple relative risk functions in matched case-control studies. *American Journal of Epidemiology*, *108*(4), 299-307.

Breslow, N. E., Lubin, J. H., Marek, P. & Langholz, B. (1983). Multiplicative models and cohort analysis. *Journal of the American Statistical Association*, *78*(381), 1-12.

Breslow, N. (1972). Contribution to the discussion on the paper by DR Cox, Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, *34*(2), 216-217.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, *30*(1), 89-99.

Brickner, C. P. (2015). *Estimating the relationship between a transient effect and the onset of an acute event : a comparison of the case-crossover design and cohort design* (Thèse de doctorat). Rutgers, State University of New Jersey.

Buteau, S., Goldberg, M. S., Burnett, R. T., Gasparrini, A., Valois, M.-F., Brophy, J. M., Crouse, D. L. & Hatzopoulou, M. (2018). Associations between ambient air pollution and daily mortality in a cohort of congestive heart failure : Case-crossover and nested case-control analyses using a distributed lag nonlinear model. *Environment International*, *113*, 313-324.

Bykov, K., Wang, S. V., Hallas, J., Pottegard, A., Maclure, M. & Gagne, J. J. (2020). Bias in case-crossover studies of medications due to persistent use : A simulation study. *Pharmacoepidemiology and Drug Safety*, *29*(9), 1079-1085.

Cheung, Y. B., Ma, X., Lam, K. F., Li, J. & Milligan, P. (2019). Bias control in the analysis of case–control studies with incidence density sampling. *International Journal of Epidemiology*, *48*(6), 1981-1991.

Clavel-Chapelon, F. & E3N Study Group. (2015). Cohort profile : the French E3N cohort study. *International Journal of Epidemiology*, *44*(3), 801-809.

Collaborative Group on Hormonal Factors in Breast Cancer. (1997). Breast cancer and hormone replacement therapy : collaborative reanalysis of data from 51 epidemiological studies of 52,705 women with breast cancer and 108,411 women without breast cancer. *Lancet*, *350*(9084), 1047-1059.

Collaborative Group on Hormonal Factors in Breast Cancer. (2019). Type and timing of menopausal hormone therapy and breast cancer risk : individual participant meta-analysis of the worldwide epidemiological evidence. *Lancet*, *394*(10204), 1159-1168.

Commenges, D., Letenneur, L., Joly, P., Alioum, A. & Dartigues, J.-F. (1998). Modelling age-specific risk : application to dementia. *Statistics in Medicine*, *17*(17), 1973-1988.

Coveney, J. (2008). *FIRTHLOGIT : Stata module to calculate bias reduction in logistic regression*.

Cox, D. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, *34*(2), 187-202.

Cox, D. (1975). Partial likelihood. *Biometrika*, *62*(2), 269-276.

Crowther, M. J. & Lambert, P. C. (2013). Simulating biologically plausible complex survival data. *Statistics in Medicine*, *32*(23), 4118-4134.

D'Agostino, R. B., Lee, M. L., Belanger, A. J., Cupples, L. A., Anderson, K. & Kannel, W. B. (1990). Relation of pooled logistic regression to time dependent Cox regression analysis : the Framingham Heart Study. *Statistics in Medicine*, *9*(12), 1501-1515.

de Groot, M. C. H., Candore, G., Uddin, M. J., Souverein, P. C., Ali, M. S., Belitser, S. V., Huerta, C., Groenwold, R. H. H., Alvarez, Y., Slattery, J., Korevaar, J., Hoes, A. W., Roes, K. C. B., de Boer, A., Douglas, I. J., Schlienger, R. G., Reynolds, R., Klungel, O. H. & Gardarsdottir, H. (2016). Case-only designs for studying the association of antidepressants and hip or femur fracture. *Pharmacoepidemiology and Drug Safety*, *25*, 103-113.

Delaney, J. C. & Suissa, S. (2009). The case-crossover study design in pharmacoepidemiology. *Statistical Methods in Medical Research*, *18*(1), 53-65.

Desai, R., Glynn, R., Wang, S. & Gagne, J. (2016). Nested case control sampling macro guide.

Desquilbet, L. & Meyer, L. (2005). Variables dépendantes du temps dans le modèle de Cox Théorie et pratique. *Revue d'Épidémiologie et de Santé Publique*, *53*(1), 51-68.

Dixon, K. E. (1997). A comparison of case-crossover and case-control designs in a study of risk factors for hemorrhagic fever with renal syndrome. *Epidemiology*, *8*(3), 243-246.

Donnan, P. T. & Wang, J. (2001). The case-crossover and case-time-control designs in pharmacoepidemiology. *Pharmacoepidemiology and Drug Safety*, *10*(3), 259-262.

Edwards, I. R. & Aronson, J. K. (2000). Adverse drug reactions : definitions, diagnosis, and management. *Lancet*, *356*(9237), 1255-1259.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, *72*(359), 557-565.

Essebag, V., Platt, R. W., Abrahamowicz, M. & Pilote, L. (2005). Comparison of nested case-control and survival analysis methodologies for analysis of time-dependent exposure. *BMC Medical Research Methodology*, *5*(1), 5.

Etminan, M. (2004). Pharmacoepidemiology II : The nested case-control study—A novel approach in pharmacoepidemiologic research. *Pharmacotherapy*, *24*(9), 1105-1109.

Farewell, V. T. & Prentice, R. L. (1980). The approximation of partial likelihood with emphasis on case-control studies. *Biometrika*, *67*(2), 273-278.

Farrington, C. P. (1995). Relative incidence estimation from case series for vaccine safety evaluation. *Biometrics*, *51*(1), 228-235.

Farrington, C. P. (2004). Control without separate controls : evaluation of vaccine safety using case-only methods. *Vaccine*, *22*(15-16), 2064-2070.

Farrington, C. P., Anaya-Izquierdo, K., Whitaker, H. J., Hocine, M. N., Douglas, I. & Smeeth, L. (2011). Self-controlled case series analysis with event-dependent observation periods. *Journal of the American Statistical Association*, *106*(494), 417-426.

Farrington, C. P., Nash, J. & Miller, E. (1996). Case series analysis of adverse reactions to vaccines : a comparative evaluation. *American Journal of Epidemiology*, *143*(11), 1165-1173.

Farrington, C. P., Whitaker, H. J. & Hocine, M. N. (2008). Case series analysis for censored, perturbed, or curtailed post-event exposures. *Biostatistics*, *10*(1), 3-16.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, *80*(1), 27-38.

Fournier, A., Mesrine, S., Dossus, L., Boutron-Ruault, M.-C., Clavel-Chapelon, F. & Chabbert-Buffet, N. (2014). Risk of breast cancer after stopping menopausal hormone therapy in the E3N cohort. *Breast Cancer Research and Treatment*, *145*(2), 535-543.

Gail, M., Williams, R., Byar, D. P. & Brown, C. (1976). How many controls ? *Journal of Chronic Diseases*, *29*(11), 723-731.

Gasparrini, A. (2011). Distributed Lag Linear and Non-Linear Models in R : The Package dlnm. *Journal of Statistical Software*, *43*(8), 1-20.

Gasparrini, A. (2014). Modeling exposure-lag-response associations with distributed lag non-linear models. *Statistics in Medicine*, *33*(5), 881-899.

Gasparrini, A., Armstrong, B. & Kenward, M. G. (2010). Distributed lag non-linear models. *Statistics in Medicine*, *29*(21), 2224-2234.

Goldstein, L. & Langholz, B. (1992). Asymptotic theory for nested case-control sampling in the Cox regression model. *Annals of Statistics*, *20*(4), 1903-1928.

Greenland, S. & Mansournia, M. A. (2015). Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Statistics in Medicine*, *34*(23), 3133-3143.

Greenland, S., Mansournia, M. A. & Altman, D. G. (2016). Sparse data bias : a problem hiding in plain sight. *BMJ*, *353*, i1981.

Greenland, S., Schwartzbaum, J. A. & Finkle, W. D. (2000). Problems due to small samples and sparse data in conditional logistic regression analysis. *American Journal of Epidemiology*, *151*(5), 531-539.

Greenwood, M. (1926). A report on the Natural Duration of Cancer. [Publisher : London : H.M.S.O.]. *Reports on Public Health and Medical Subjects. Ministry of Health*, (33), 1-26.

Grosso, A., Douglas, I., MacAllister, R., Petersen, I., Smeeth, L. & Hingorani, A. D. (2011). Use of the self-controlled case series method in drug safety assessment. *Expert Opinion on Drug Safety*, *10*(3), 337-340.

Guess, H. A. (2006). Exposure-time-varying hazard function ratios in case-control studies of drug effects. *Pharmacoepidemiology and Drug Safety*, *15*(2), 81-92.

Hak, E., Wei, F., Grobbee, D. & Nichol, K. (2004). A nested case-control study of influenza vaccination was a cost-effective alternative to a full cohort analysis. *Journal of Clinical Epidemiology*, *57*(9), 875-880.

Hallas, J. (1996). Evidence of depression provoked by cardiovascular medication : a prescription sequence symmetry analysis. *Epidemiology*, *7*(5), 478-484.

Hauptmann, M., Wellmann, J., Lubin, J. H., Rosenberg, P. S. & Kreienbrock, L. (2000). Analysis of exposure-time-response relationships using a spline weight function. *Biometrics*, *56*(4), 1105-1108.

Heinze, G. & Ladner, T. (2012). *logistiX : Exact logistic regression including Firth correction* [R package version 1.0].

Heinze, G. & Ploner, M. (2002). SAS and SPLUS programs to perform Cox regression without convergence problems. *Computer Methods and Programs in Biomedicine*, (67), 217-223.

Heinze, G. & Ploner, M. (2003). Fixing the nonconvergence bug in logistic regression with SPLUS and SAS. *Computer Methods and Programs in Biomedicine*, (71), 181-187.

Heinze, G., Ploner, M., Dunkler, D., Southworth, H. & Jiricka, L. (2022). *logistf : Firth's Bias-Reduced Logistic Regression* [R package version 1.24.1].

Heinze, G., Ploner, M. & Jiricka, L. (2020). *coxphf : Cox Regression with Firth's Penalized Likelihood* [R package version 1.13.1].

Heinze, G. & Puhr, R. (2010). Bias-reduced and separation-proof conditional logistic regression with small or sparse data sets. *Statistics in Medicine*, *29*(7-8), 770-777.

Heinze, G. & Schemper, M. (2001). A solution to the problem of monotone likelihood in Cox regression. *Biometrics*, *57*(1), 114-119.

Hendry, D. J. (2014). Data generation for the Cox proportional hazards model with time-dependent covariates : a method for medical researchers. *Statistics in Medicine*, *33*(3), 436-454.

Hernán, M. A., Jick, S. S., Olek, M. J. & Jick, H. (2004). Recombinant hepatitis B vaccine and the risk of multiple sclerosis : a prospective study. *Neurology*, *63*(5), 838-842.

Hernandez-Diaz, S. (2003). Case-crossover and case-time-control designs in birth defects epidemiology. *American Journal of Epidemiology*, *158*(4), 385-391.

Hertz-Picciotto, I. & Rockhill, B. (1997). Validity and efficiency of approximation methods for tied survival times in Cox regression. *Biometrics*, *53*(3), 1151-1156.

Hocine, M. N. & Chavance, M. (2010). La méthode de la série de cas. *Revue d'Épidémiologie et de Santé Publique*, *58*(6), 435-440.

Kalbfleisch, J. D. & Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, Hoboken, New Jersey.

Kaplan, E. L. & Meier, P. (1958). Non-parametric estimation from incomplete observations. *Journal of the American Statistical Association*, *53*(282), 457-481.

Kim, R. S. (2015). A new comparison of nested case–control and case–cohort designs and methods. *European Journal of Epidemiology*, *30*(3), 197-207.

Kim, R. (2015). A new comparison of nested case–control and case–cohort designs and methods. *European Journal of Epidemiology*, *30*, 197-207.

Klein, J. P., van Houwelingen, H. C., Ibrahim, J. G. & Scheike, T. H. (Éd.). (2014). *Handbook of Survival Analysis*. CRC Press, Taylor & Francis Group, Boca Raton.

Korn, E., Graubard, B. I. & Midthune, D. (1997). Time-to-event analysis of longitudinal follow-up of a survey : choice of the time-scale. *American Journal of Epidemiology*, *145*(1), 72-80.

Kosmidis, I. (2013). *brglm : Bias reduction in binary-response Generalized Linear Models* [R package version 0.5-9].

Lai, E. C.-C., Pratt, N., Hsieh, C.-Y., Lin, S.-J., Pottegård, A., Roughead, E. E., Kao Yang, Y.-H. & Hallas, J. (2017). Sequence symmetry analysis in pharmacovigilance and pharmacoepidemiologic studies. *European Journal of Epidemiology*, *32*(7), 567-582.

Langholz, B. & Richardson, D. B. (2010). Fitting general relative risk models for survival time and matched case-control analysis. *American Journal of Epidemiology*, *171*(3), 377-383.

Langholz, B. & Thomas, D. C. (1990). Nested case-control and case-cohort methods of sampling from a cohort : a critical comparison. *American Journal of Epidemiology*, *131*(1), 169-176.

Le Vu, S., Bertrand, M., Jabagi, M.-J., Botton, J., Drouin, J., Baricault, B., Bouillon, K., Semenzato, L., Weill, A., Dray-Spira, R. & Zureik, M. (2021). Association entre les vaccins COVID-19 à ARN messager et la survenue de myocardite et péricardite chez les personnes de 12 à 50 ans en France : Etude à partir des

données du Système National des Données de Santé (SNDS) (rapp. tech.). EPI-PHARE - Groupement d'intérêt scientifique (GIS) ANSM-CNAM.

Leemis, L. M. (1987). Variate generation for accelerated life and proportional hazards models. *Operations Research*, *35*(6), 892-894.

Leffondré, K., Abrahamowicz, M. & Siemiatycki, J. (2003). Evaluation of Cox's model and logistic regression for matched case-control data with time-dependent covariates : a simulation study : Cox's model for case-control data with time-dependent covariates. *Statistics in Medicine*, *22*(24), 3781-3794.

Leffondré, K., Wynant, W., Cao, Z., Abrahamowicz, M., Heinze, G. & Siemiatycki, J. (2010). A weighted Cox model for modelling time-dependent exposures in the analysis of case-control studies. *Statistics in Medicine*, *29*(7-8), 839-850.

Liddell, F. D. K., McDonald, J. C., Thomas, D. C. & Cunliffe, S. V. (1977). Methods of Cohort analysis : appraisal by application to asbestos mining. *Journal of the Royal Statistical Society. Series A (General)*, *140*(4), 469.

Lin, D. & Ying, Z. (1993). Cox regression with incomplete covariate measurements. *Journal of the American Statistical Association*, *88*(424), 1341-1349.

Lubin, J. H. & Gail, M. H. (1984). Biased selection of controls for case-control analyses of cohort studies. *Biometrics*, *40*(1), 63-75.

Mackenzie, T. & Abrahamowicz, M. (2002). Marginal and hazard ratio specific random data generation : Applications to semi-parametric bootstrapping. *Statistics and Computing*, *12*(3), 245-252.

Maclure, M. (1991). The case-crossover design : a method for studying transient effects on the risk of acute events. *American Journal of Epidemiology*, *133*(2), 144-153.

Maclure, M. & Mittleman, M. A. (2000). Should we use a case-crossover design? *Annual Review of Public Health*, *21*, 193-221.

Marti, H. & Chavance, M. (2013). Les enquêtes cas–cohorte. *Revue d'Épidémiologie et de Santé Publique*, *61*(1), 67-74.

Mittleman, M. A., Maclure, M. & Robins, J. M. (1995). Control sampling strategies for case-crossover studies : An assessment of relative efficiency. *American Journal of Epidemiology*, *142*(1), 91-98.

Montez-Rath, M. E., Kapphahn, K., Mathur, M. B., Mitani, A. A., Hendry, D. J. & Desai, M. (2017). Guidelines for generating right-censored outcomes from a Cox model extended to accommodate time-varying covariates. *Journal of Modern Applied Statistical Methods*, *16*(1), 86-106.

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, *14*(4), 945-966.

Nicholas, J. M., Grieve, A. P. & Gulliford, M. C. (2012). Within-person study designs had lower precision and greater susceptibility to bias because of trends in exposure than cohort and nested case–control designs. *Journal of Clinical Epidemiology*, *65*(4), 384-393.

Pang, D. (1999). A relative power table for nested matched case-control studies. *Occupational and Environmental Medicine*, *56*(1), 67-69.

Pazzagli, L., Linder, M., Zhang, M., Vago, E., Stang, P., Myers, D., Andersen, M. & Bahmanyar, S. (2018). Methods for time-varying exposure related problems in pharmacoepidemiology : An overview. *Pharmacoepidemiology and Drug Safety*, *27*(2), 148-160.

Petersen, I., Douglas, I. & Whitaker, H. J. (2016). Self controlled case series methods : an alternative to standard epidemiological study designs. *BMJ*, *354*, i4515.

Peto, R. (1972). Contribution to discussion of paper : Regression models and life-tables. by D.R. Cox. *Journal of the Royal Statistical Society. Series B (Methodological)*, *34*(2), 205-207.

Platt, R. W. (2004). A proportional hazards model with time-dependent covariates and time-varying effects for analysis of fetal and infant death. *American Journal of Epidemiology*, *160*(3), 199-206.

Prentice, R. L. & Breslow, N. E. (1978). Retrospective studies and failure time models. *Biometrika*, *65*(1), 153-158.

Prentice, R. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, *73*(1), 1-11.

Preston, D. (2000). Poisson regression for survival data in epidemiology. In M. H. Gail & J. Benichou (Éd.), *Encyclopedia of Epidemiologic Methods* (p. 723-727). John Wiley & Sons, Chichester.

Rahman, M. S. & Sultana, M. (2017). Performance of Firth-and logF-type penalized methods in risk prediction for small or sparse binary data. *BMC Medical Research Methodology*, *17*(1), 33.

Robins, J. M., Hernán, M. Á. & Brumback, B. (2000). Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology*, *11*(5), 550.

Rossouw, J. E., Anderson, G. L., Prentice, R. L., LaCroix, A. Z., Kooperberg, C., Stefanick, M. L. & Jackson, R. D. (2002). Risks and benefits of estrogen plus progestin in healthy postmenopausal women : principal results From the Women's Health Initiative randomized controlled trial. *JAMA*, *288*(3), 321-333.

Ryan, T. P. & Woodall, W. H. (2005). The most-cited statistical papers. *Journal of Applied Statistics*, *32*(5), 461-474.

Schneeweiss, S., Stürmer, T. & Maclure, M. (1997). Case-crossover and case-time-control designs as alternatives in pharmacoepidemiologic research. *Pharmacoepidemiology and Drug Safety*, *6 Suppl 3*, S51-59.

Strom, B. L., Kimmel, S. E. & Hennessy, S. (Éd.). (2013). *Textbook of pharmacoepidemiology* (2nd ed). Wiley Blackwell, Chichester.

Suissa, S. (1995). The case-time-control design. *Epidemiology (Cambridge, Mass.)*, *6*(3), 248-253.

Sun, J. X., Sinha, S., Wang, S. & Maiti, T. (2011). Bias reduction in conditional logistic regression. *Statistics in Medicine*, *30*(4), 348-355.

Sylvestre, M.-P. & Abrahamowicz, M. (2008). Comparison of algorithms to generate event times conditional on time-dependent covariates. *Statistics in Medicine*, *27*(14), 2618-2634.

Sylvestre, M.-P. & Abrahamowicz, M. (2009). Flexible modeling of the cumulative effects of time-dependent exposures on the hazard. *Statistics in Medicine*, *28*(27), 3437-3453.

Sylvestre, M.-P., Beauchamp, M.-E., Kyle, R. P. & Abrahamowicz, M. (2018). *Package 'WCE'* [R package version 1.0.2].

Sylvestre, M.-P., Edens, T., MacKenzie, T. & Abrahamowicz, M. (2015). *PermAlgo : Permutational Algorithm to Simulate Survival Data* [R package version 1.1].

Takeuchi, Y., Shinozaki, T. & Matsuyama, Y. (2018). A comparison of estimators from self-controlled case series, case-crossover design, and sequence symmetry analysis for pharmacoepidemiological studies. *BMC Medical Research Methodology*, *18*(1).

Taylor, J. M. G. (1986). Choosing the number of controls in a matched case-control study, some sample size, power and efficiency considerations. *Statistics in Medicine*, *5*(1), 29-36.

Thiébaut, A. C. M. & Bénichou, J. (2004). Choice of time-scale in Cox's model analysis of epidemiologic cohort data : a simulation study. *Statistics in Medicine*, *23*(24), 3803-3820.

Tsiropoulos, I., Andersen, M. & Hallas, J. (2009). Adverse events with use of antiepileptic drugs : a prescription and event symmetry analysis. *Pharmacoepidemiol Drug Saf*, *18*(6), 483-491.

Vacek, P. M. (1997). Asessing the effect of intensity when exposure varies over time. *Statistics in Medicine*, *16*(5), 505-513.

Viel, J. F. (2004). La régression de Poisson en épidémiologie. *Revue d'Epidémiologie et de Santé Publique*, *42*(1), 79-87.

Vinogradova, Y., Coupland, C. & Hippisley-Cox, J. (2020). Use of hormone replacement therapy and risk of breast cancer : nested case-control studies using the QResearch and CPRD databases. *BMJ*, *371*, m3873.

Wacholder, S. (1991). Practical considerations in choosing between the case-cohort and nested case-control designs. *Epidemiology*, *2*(2), 155-158.

Wacholder, S. (2009). Bias in full cohort and nested case-control studies? *Epidemiology*, *20*(3), 339-340.

Wahab, I. A., Pratt, N. L., Ellett, L. K. & Roughead, E. E. (2016). Sequence symmetry analysis as a signal detection tool for potential heart failure adverse events in an administrative claims database. *Drug Safety*, *39*(4), 347-354.

Wang, S., Linkletter, C., Maclure, M., Dore, D., Mor, V., Buka, S. & Wellenius, G. A. (2011). Future-cases as present controls to adjust for exposure-trend bias in case-only studies. *Epidemiology*, *22*(4), 568-574.

Whitaker, H. J., Farrington, C. P., Spiessens, B. & Musonda, P. (2006). Tutorial in biostatistics : the self-controlled case series method. *Statistics in Medicine*, *25*(10), 1768-1797.

Whitaker, H. J., Hocine, M. N. & Farrington, C. P. (2009). The methodology of self-controlled case series studies. *Statistical Methods in Medical Research*, *18*(1), 7-26.

Zhou, M. (2001). Understanding the Cox regression models with time-change co-variates. *American Statistician*, *55*(2), 153-155.

A - Travaux comparant les différents schémas d'étude

Tableau A.1 – *Présentation de quelques travaux comparant les différents schémas d'étude à partir de données simulées*

Référence des travaux	Schéma d'étude								Type d'exposition	Approche de génération des données	Méthode de gestion des ex-aequos	Méthodes d'analyse des données	Critères de comparaison	Scénarios
	COH	CTN	CCH	SCCS	CCO	CTC	CCTC	SSA						
(Bykov et al., 2020)					■	■			Exposition(binaire) dépendante du temps avec au plus deux changements			Régression logistique conditionnelle	Biais	4 scénarios de durée d'exposition (tous les patients exposés pendant 30 jours ; 180 jours ; 2 ans et certains patients exposés pendant 30 jours, les autres restant exposés jusqu'à la fin du suivi (utilisateurs persistants)); 5 proportions d' utilisateurs persistants dans leur traitement (10 %, 30 %, 50 %, 70 % et 90 %); OR=1 pour tous les scénarios et OR= 0,25 ; 0,5; 0,8 ; 1,0; 1,25; 2,0 et 4,0 pour le scénario avec 30 % de persistance
(Takeuchi et al., 2018)				■	■			■	Exposition(binaire) dépendante du temps avec changements multiples			Régression de Poisson conditionnelle, Régression logistique conditionnelle	Biais; Ecart type empirique; Erreur quadratique moyenne; Ecart type moyen; Probabilité de couverture des intervalles de confiance	Facteurs de confusion invariables dans le temps avec interactions entre eux ; Facteurs de confusion fixes et dépendants du temps, sans interaction entre eux, avec la durée d'exposition qui varie (5, 10, 15, 20 ou 30 jours); Facteurs de confusion fixes et dépendants du temps, avec interaction entre eux, avec la durée d'exposition qui varie (5, 10, 15, 20 ou 30 jrs); Tendances temporelles des expositions et des évènements ; Durée de suivi limitée en fonction de la survenue d'un évènement ; Exclusion des patients qui ont connu le premier évènement avant leur première exposition ; Durées des périodes à risque mal spécifiées
(Kim, 2015)	■	■	■						Exposition(continue ou binaire) fixe			Modèle de Cox à risque proportionnels; Régression logistique conditionnelle; Pondération par la probabilité inverse	Biais; écart type empirique; erreur de type 1; puissance	Le type de prédicteur (continu ou binaire); L'ampleur l'effet (0 ; 0,1 ; 0,2 ; … ; 0,8); La taille de la cohorte (500, 1000, 1500); Le nombre de témoins (1, 2 ou 5)
(Brickner, 2015)	■				■				Exposition (binaire) dépendante du temps avec au plus deux changements	(Austin, 2012)		Modèle de Cox à risque proportionnels; Régression logistique conditionnelle	Biais; Ecart type moyen; Puissance; Couverture	L'ampleur l'effet (0 ; 0,8); La taille de la cohorte (2500, 5000, 10000, 25000, 50000 ou 100000); La proportion des sujets ayant connu l'évènement d'intérêt (5%, 10%); Le nombre de périodes témoins (1, 2, 3, 5, 10 ou 25)
(Bertke et al., 2013)	■	■							Exposition(continue) fixe		Approximation de Breslow	Régression logistique conditionnelle	Efficacité relative; Biais	Le nombre de cas (30, 100 ou 300); L'intensité de la relation exposition-réponse (HR = 1 ; 1,005 ; 1,010 ou 1,015); Le nombre de témoins par cas (1, 5, 10, 15 ou 20); L'asymétrie de la distribution de l'exposition (distribution symmétrique, légèrement inclinée à droite ou très à droite)
(Austin et al., 2012)	■	■							Exposition(binaire) fixe; exposition (binaire) dépendante du temps avec changement unique	(Austin, 2012 ; Bender et al., 2005)		Modèle de Cox à risque proportionnels; Régression logistique conditionnelle;	Biais; Ecart type moyen; Couverture ; Erreur quadratique moyenne	L'ampleur de l'effet (HR=1.25 / 2); La proportion des sujets traités/exposés (10%, 25%, 50%); La proportion des sujets ayant connu l'évènement d'intérêt (5%, 10%, 25%); Le nombre de témoins (1 ou 5)
(Wang et al., 2011)					■	■	■		Exposition (binaire) dépendante du temps avec changements multiples			Régression logistique conditionnelle	Biais	Trois scénarios de confusion (pas de facteur de confusion non mesuré ; présence d'un facteur de confusion binaire et invariable dans le temps ; présence d' un facteur de confusion invariable dans le temps et augmentation de la probabilité d'exposition dans le temps); Taux d'évènements fixés pour être 1 ; 2 ; ou 3 fois plus élevés pendant la période "exposée" que pendant la période "non exposée".
(Leffondré et al., 2010)	■	■							Exposition(continue) dépendante du temps avec changements multiples	(Sylvestre & Abrahamowicz, 2008)		Régression logistique; Régression logistique conditionnelle; Modèle de Cox à risques proportionnels, Modèle de Cox pondéré	Biais relatif; Erreur quadratique moyenne; Ecart-type moyen; Couverture; Ecart type empirique	Trois scénarios de changement pour l'intensité d'exposition : constante, croissante ou décroissante, avec des probabilités (q1, q2 et q3) d'appartenir à chacune de ces trois catégories ; Deux modèles : l'un avec une exposition cumulée et l'autre avec l'intensité actuelle de l'exposition et la durée d'exposition
(Leffondré et al., 2003)	■	■							Exposition (binaire ou continue) dépendante du temps avec au plus deux changements	(Sylvestre & Abrahamowicz, 2008)		Régression logistique; Régression logistique conditionnelle; Modèle de Cox à risques proportionnels (naif, adapté)	Biais relatif; Ecart type empirique; Puissance	Deux représentations de l'exposition dépendante du temps, chacune ayant deux valeurs de l'effet sur le risque : le statut d'exposition actuelle (intensité de l'effet= 0,4 ou 1,4); la durée de l'exposition passée (intensité de l'effet= 0,006 ou 0,03)

COH : cohorte ; CTN : cas-témoins niché ; CCH : cas-cohorte ; SCCS : série de cas ; CCO : cas-croisé ; CTC : *case-time-control* ; CCTC : *case-case-time-control* ; SSA : *sequence symmetry analysis*

Tableau A.2 – *Présentation de quelques travaux comparant les différents schémas d'étude à partir de données réelles*

Référence des travaux	Schéma d'étude								Type d'exposition	Méthodes d'analyse de données	Critères de comparaison
	COH	CTN	CCH	SCCS	CCO	CTC	CCTC	SSA			
(Buteau et al., 2018)		■			■				exposition dépendante du temps à changement multiple	modèle de regression logistique conditionnelle	
(de Groot et al., 2016)				■	■				exposition dépendante du temps à changement multiple	modèle de regression logistique conditionnelle; modèle de régression de Poisson conditionnelle (SCCS)	Risque estimé, significativité
(Brauer et al., 2016)				■	■				exposition dépendante du temps à changement multiple	modèle de regression l ogistique conditionnelle (CCO); modèle de régression de Poisson conditionnelle (SCCS)	Risque estimé, significativité
(Billioti de Gage et al., 2012)	■	■							exposition (binaire) dépendante du temps à changement multiple	Modèle de Cox à risques proportionnels (cohorte); modèle de regression logistique conditionnelle (CTN(1 :1) à CTN (4 :1))	HR et OR
(Nicholas et al., 2012)	■	■		■	■	■			exposition dépendante du temps	Modèle de Cox à risques proportionnels (Cohorte); modèle de regression logistique conditionnelle (CTN (4 :1), CCO et CTC); modèle de régression de Poisson conditionnelle (SCCS)	Risque estimé (HR/OR/IRR), écart type des coefficients de régression
(Essebag et al., 2005)	■	■							exposition (binaire) dépendante du temps à changement unique	Modèle de Cox à risques proportionnels (Cohorte); Régression logistique conditionnelle (CTN avec 4, 8, 16, 32 et 64 témoins par cas) repétée 100 fois	Risque estimé (HR et HR moyen) et temps de calcul
(Hak et al., 2004)	■	■							exposition dépendante du temps à changement multiple	modèle de Cox à risques proportionnels (cohorte); modèle de regression logistique ctn(1 :1) à ctn(4 :1)	Incidence relative estimée et intervalle de confiance
(Hernandez-Diaz, 2003)					■	■			exposition dépendante du temps à changement multiple	Modèle de regression logistique	OR estimé et intervalle de confiance
(Dixon, 1997)		■			■				exposition dépendante du temps à changement multiple	Modèle de regression logistique	OR estimé
(Farrington et al., 1996)	■	■		■					exposition dépendante du temps à changement multiple	Modèle de Cox à risques proportionnels (Cohorte); modèle de regression logistique conditionnelle (CTN); modèle de régression de Poisson conditionnelle (SCCS)	Incidence relative estimée et intervalle de confiance

COH : cohorte; CTN : cas-témoins niché; CCH : cas-cohorte; SCCS : série de cas; CCO : cas-croisé; CTC : *case-time-control*; CCTC : *case-case-time-control*; SSA : *sequence symmetry analysis*

B - Article soumis à *Biometrical Journal*

Biometrical Journal

Comparison of cohort and nested case-control designs for estimating the effect of time-varying drug exposure on the risk of adverse event in the presence of ties

Journal:	*Biometrical Journal*
Manuscript ID	Draft
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	n/a
Complete List of Authors:	MANITCHOKO, Liliane Abrahamowicz, Michal; McGill University, Epidemiology, Biostatistics and Occupational Health Thiébaut, Anne; INSERM, UMR 1181 B2PHI; INSERM, UMR1181 PhEMI Benichou, Jacques Tubert-Bitter, Pascale
Keywords:	Breast cancer, Cohort design, Nested case control design, , Pharmacoepidemiology, Simulation study, Tied events, Time-varying exposure

Comparison of cohort and nested case-control designs for estimating the effect of time-varying drug exposure on the risk of adverse event in the presence of ties

Liliane Manitchoko[1], Michal Abrahamowicz[2], Pascale Tubert-Bitter[1], Jacques Benichou[1,3*], Anne C.M. Thiébaut[1*]

[1] Université Paris-Saclay, UVSQ, Inserm, CESP, High Dimensional Biostatistics for Drug Safety and Genomics, 94807, Villejuif, France

[2] Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada

[3] Department of Biostatistics, Rouen University Hospital, Rouen, France; University of Rouen-Normandie, France

*These authors contributed equally to this work

Correspondence

Liliane Manitchoko, Université Paris-Saclay, UVSQ, Inserm, CESP, High Dimensional Biostatistics Team, 16 avenue Paul Vaillant Couturier, 94807, Villejuif, France

Email: liliane.manitchoko@inserm.fr

Acknowledgements

This research was supported by the French Medicines Agency (*Agence Nationale de Sécurité du Médicament et des produits de santé*, ANSM) and the French Institute for Research in Public Health (*Institut pour la Recherche en Santé Publique*, GIS-IReSP). MA is a James McGill Professor of Biostatistics at McGill University.

The authors wish to thank Agnès Fournier and Gianluca Severi 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 National*' (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*.

Conflict of interest

The authors declare that they have no competing interests.

Data availability statement

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.

Abstract

Cohort and nested case control (NCC) designs are frequently used in pharmacoepidemiology to assess the associations of drug exposure that can vary over time with the risk of an adverse event. Although it is typically expected that estimates from NCC analyses are similar to those from the full cohort analysis, with moderate loss of precision, only few studies have actually compared their respective performance for estimating the effects of time-varying exposures (TVE). We used simulations to compare the properties of the resulting estimators of these designs in terms of bias, precision, coverage probability, and root mean square error for both time-invariant exposure and TVE. We varied exposure prevalence, proportion of subjects experiencing the event, hazard ratio and control to case ratio. Using both designs, we also estimated the real-world associations of time-invariant ever use of menopausal hormone therapy (MHT) at baseline and updated, time-varying MHT use with breast cancer incidence. In all simulated scenarios, the cohort-based estimates had small relative bias and greater precision than the NCC design. NCC estimates displayed bias to the null that decreased with greater number of controls per case. This bias markedly increased with higher proportion of events. Bias was seen with Breslow's and Efron's approximations for handling tied event times but was greatly reduced with the exact method. When analyzing the MHT-breast cancer association, differences between the two designs were consistent with simulated data. Once ties were taken correctly into account, NCC estimates were very similar to those of the full cohort analysis.

Keywords

Breast cancer, Cohort design, Nested case control design, Pharmacoepidemiology, Simulation study, Tied events, Time-varying exposure

Abstract word count: 249

Text word count: 4678

1 Introduction

In pharmacoepidemiology, assessing the effect of drug exposure on the risk of an adverse event is a real challenge because exposure can vary over time and its effect can be complex (Abrahamowicz et al., 2012). To account for within-subject variation in drug exposure, longitudinal data need to be collected for a cohort of individuals and time-to-event analyses are generally used to model the time-varying exposure (TVE) to drug (Pazzagli et al., 2018).

The analysis of the entire cohort using the Cox proportional hazards regression model (Cox, 1972) is the classical approach to analyze such data (Ryan & Woodall, 2005). Cox model allows to account for change of exposure during the follow-up using time-dependent covariates (O'Quigley, 2008). Estimation is based on partial likelihood, which, at each observed event, compares the values of the covariates of the individual who experienced the event with those of all individuals at risk at that event time (Cox, 1975).

The Nested Case-Control (NCC) study design is another widely used approach to explore the association between drug exposure and the event of interest. NCC analysis is conducted as a case-control study based on all cases (subjects who experienced the event of interest during follow-up) and a random sample of control subjects individually matched to cases on time at risk, using incidence density sampling (Lubin & Gail, 1984; Prentice & Breslow, 1978), and possibly some covariates (Breslow & Day 1987; Klein et al., 2014). The sampled controls may be selected as controls for more than one case and may later become cases. The TVE of interest, for both the case and his/her matched controls, is evaluated at the case's time of event, and then compared between the case and his/her matched controls. Therefore, for each study subject, a truly TVE is assessed only at a single time point, avoiding the need for complex statistical techniques (Etminan, 2004). The NCC is an efficient sampling design that requires exposure assessment on a far smaller sample that the full cohort design (Liddell et al., 1977), which is advantageous when exposure measurements (*e.g.*, biomarkers) are difficult to obtain for logistical and/or financial reasons. Thus, the NCC design has been used for its convenience, cost-efficiency and analytical flexibility (Essebag et al., 2005; Langholz, 2014). One key advantage of the NCC design is, indeed, that it allows the use of matching variables other than time and therefore is effective in reducing the risk of residual confounding due to possibly mis-modeling of some important confounders (Etminan, 2004). Estimation in the NCC design is based on a stratified version of the partial likelihood, where for each stratum the risk set, created at the time of case's event, is limited to a case and his/her matched controls (Prentice & Breslow, 1978).

The partial likelihood used to estimate the Cox proportional hazards model depends on the ordering of events in time. Yet, due to imprecise measurement of time (*e.g.*, in days, weeks or months), datasets may contain several *tied* event times, *i.e.*, events that occur at the same time, which poses special problems for partial likelihood estimation (Hertz-Picciotto & Rockhill, 1997). In the presence of ties, exact expression of the partial likelihood considers all possible permutations at each time when more than one event occurs (Peto, 1972). However, this is time-consuming and there are situations in which this approach may become computationally impossible due to the large number of ties in the dataset. To overcome these computational difficulties, some approximations have been developed. Both Breslow's (1974) and Efron's (1977) approximations are much faster than the exact method for handling the ties, and both work well when ties are relatively few. However, both approximations can yield biased estimates when the average proportion of cases in the risk sets used in the analyses becomes relatively large (Allison, 2010; Borucka, 2014; Hertz-Picciotto & Rockhill, 1997), which is likely to happen in NCC studies (Farewell & Prentice, 1980).

It is generally admitted that estimates obtained from NCC analyses are similar to those obtained from analysis of the entire cohort, with only a moderate loss of precision (Breslow et al., 1983; Essebag et al., 2005; Liddell et al., 1977). However, to our knowledge, only one study relied on simulations to systematically evaluate and compare their respective performance for estimating the effect of TVE (Austin et al., 2012). In contrast with the previous findings of NCC and whole cohort analyses yielding similar estimates, the authors reported potentially important biases in NCC analyses with time-varying exposure that increased with the proportion of un-censored events (Austin et al., 2012).

The aim of this study was to further investigate the performances of cohort and NCC approaches for estimating the association between exposure and the risk of adverse event and the potential for bias in the latter analyses, in particular for a TVE. Special attention was paid to the presence, frequency and handling of ties, as they can greatly influence the results (Allison, 2010; Farewel & Prentice, 1980).

This article is organized as follows: first, we describe the simulation study that was conducted to assess and compare the properties of the NCC *versus* cohort-based estimators for both a time-invariant exposure and a simple TVE. Then, we present the results of a real-world case study, where we apply these designs to estimate the risk of invasive breast cancer associated with ever use of menopausal hormone therapy (MHT) in a large cohort of French

women with long follow-up (Fournier et al., 2014). Finally, we conclude with a discussion and recommendations.

2 Simulation study

2.1 Methods

We performed two simulation studies, one focusing on a time-invariant binary exposure and another on a simple binary change-of-state TVE, such as treatment initiation during the follow-up. For each simulation, we considered 18 alternative scenarios, that differed with respect to (*i*) value of 'true' hazard ratio (HR), (*ii*) proportion of exposed subjects and (*iii*) incidence rate for the outcome of interest. For each scenario, we generated 1,000 independent random samples, each representing a hypothetical cohort of N= 5,000 subjects followed for up to 2 years (730 days).

2.1.1 Data generation

For simulations with time-invariant exposure, individual exposure status $A(i)$, i=1,…,N was generated as a binary variable (1=exposed *versus* 0=un-exposed) from a Bernoulli distribution of a pre-specified probability of exposure, and kept unchanged throughout the follow-up period. For TVE, the current exposure status $A(i,t)$ at time t was assumed to change from "un-exposed" to "exposed" at most once during follow-up. Thus, for each exposed subject *i,* we generated his/her individual time of change $\tau(i)$ from a uniform distribution [0;T] where T=maximum follow-up duration for the entire cohort. Then, we set $A(i,t)$=0 for $0<t<\tau(i)$, and $A(i,t)$=1 for $t\geq\tau(i)$. For the sake of simplicity, no covariates were generated and, thus, no adjusting or additional matching was considered in the analyses.

We used the permutational algorithm (Sylvestre & Abrahamowicz, 2008), specifically designed and validated to simulate event times conditional on time-varying effects and/or exposure/covariates, and the corresponding *permalgo* function in R (Sylvestre et al., 2015) to generate event times, conditional on exposure, as well as random times of right censoring. We assumed an exponential distribution, with a constant hazard rate, for generating event times, and a uniform distribution to generate censoring times, assumed to reflect random losses to follow-up, with additional administrative censoring of all subjects who remained at risk at two years of follow-up. Our unit time for analysis was one day so that event and censoring times were rounded up to the nearest day. For each of the two exposure types, in the data-generating model we assumed two values of 'true' HR: 2 and 1.25 (exposed *versus* un-exposed). Across

the 18 simulated scenarios, we then varied combinations of the expected proportions of subjects who: (*i*) were exposed (0.10, 0.25, and 0.50), and (*ii*) experienced an un-censored event (0.05, 0.10, and 0.25) during the two years of follow-up. The latter incidence was approximately controlled by adjusting the hazard rate λ of the exponential distribution used for generating event times.

2.1.2 Analyses of simulated data

Simulated data were analyzed using two alternative approaches: one corresponding to a full cohort design and another assuming an NCC study was embedded in the same cohort.

Data from the entire cohort were analyzed using conventional univariate un-adjusted Cox proportional hazards model with either time-invariant $A(i)$ or time-varying $A(i,t)$ exposure, depending on the structure of the simulated data. Log-HR estimates for exposure, together with 95% confidence interval (CI), were obtained using the *coxph* function from the *survival* package in R, with Efron's approximation for handling ties (Efron, 1977). This method is recommended because it is faster than the exact method (Borucka, 2014) and more accurate than Breslow's approximation (Hertz-Picciotto & Rockhill, 1997).

To perform analyses according to the NCC design, first all 'cases' (subjects who had the event during the follow-up) and their corresponding event times were identified in the cohort generated for a given simulated sample. We then used either 1:1 or 5:1 control-to-case matching. For 5:1 matching, the controls were selected in a nested manner, such that the control selected in 1:1 matching was among the five selected controls. We implemented incidence density sampling by randomly selecting controls from each case's risk set (subjects in the cohort who remain at risk, *i.e.*, event-free, until the time at which this particular case experienced the event) using the *ccwc* function in *Epi* package in R. This function groups all cases who experienced the event of interest at the same time, and their respective controls in the same stratum, therefore generating ties in those strata where two or more cases had the same event time.

To handle ties within the strata, in three alternative analyses of the same dataset, we used the exact method (Peto, 1972), as well as the Breslow (1974) and the Efron (1977) approximations. As the fourth analytical method, we modified the *ccwc* function so that cases who experienced the event at the same time were allocated to separate strata, ensuring that each stratum included only a *single* case and, thus, avoiding ties. This approach, referred to as 'modified *ccwc* approach' in the remainder of the paper, is described in Appendix 1. The NCC

samples with either 1 or 5 controls per case were analyzed using conditional logistic regression to estimate the association between exposure and the occurrence of the event of interest, by conditioning on the matched sets. Log-Odds Ratio estimates for conditional logistic regression, equivalent to log-HR were obtained using the *clogit* function from the *survival* package in R.

We compared the performance of each analytical approach with respect to estimated log-HR for exposure, across the 1,000 samples simulated for a given scenario, using: (*i*) relative bias, (*ii*) average standard error estimator (SEE), *i.e.*, the observed standard deviation of the 1,000 estimates, (*iii*) root mean square error (RMSE), and (*iv*) coverage probability of the 95% CI.

2.2 Results for a time-invariant exposure

The results of simulations for time-invariant exposure and HR=2 are presented in Figure 1 and Table 1. Figure 1 compares the relative bias of log-HR estimates obtained from the entire cohort and NCC analyses with Efron's approximation for handling ties. The frequency of ties was correlated to the proportion of un-censored events: from 15% to 53% in median for respectively 5% to 25% of subjects who experienced an event regardless of exposure prevalence (see Table in Appendix 2).

In all scenarios, the log-HR estimates were unbiased (relative bias <1%) for the cohort design. In contrast, the log-HR estimates for the NCC design displayed systematic, often important, bias to the null with Efron's approximation and even more so with Breslow's approximation. This bias was much more pronounced with 1:1 matching than 5:1 matching (*e.g.*, maximal relative bias of 27% for 1 control *versus* 6% for 5 controls with Efron's approximation, Figure 1). This bias increased markedly with higher proportion of events (*e.g.*, for 10% exposed subjects, relative bias was 10%, 19% and 27% for respectively 5%, 10% and 25% subjects who experienced an event, Figure 1). Moreover, bias decreased slightly with increasing proportion of exposed subjects *(e.g.*, for 25% of subjects who experienced an event, relative bias was 27%, 24% and 21% for respectively 10%, 25% and 50% of exposed subjects, Figure 1). These biases were almost entirely eliminated when using the exact method for handling ties (Peto, 1972) or the modified *ccwc* approach (maximal relative bias of only 1% for both methods *versus* 41% and 27% for Breslow's and Efron's approximations, respectively, Table 1).

In all scenarios, the full cohort-based estimates displayed systematically smaller variance than those of the NCC design (SEE's in Table 1). As expected, precision generally

improved in both designs as the proportion of events, exposure prevalence and (for NCC analyses) the number of controls per case increased. Precision from the NCC analysis with the exact method or modified *ccwc* approach was systematically lower than that from the NCC analysis with Breslow's or Efron's approximation. Consequently, the bias-variance trade-off was in favor of the exact method or modified *ccwc* approach only for greater proportion of events, *i.e.*, 25% and perhaps 10% (Table 1). In all scenarios, coverage probability of the cohort design maintained nominal levels (0.95) while that of NCC design was lower than nominal, especially for Breslow's and Efron's approximations with higher proportions of events (scenarios for which considerable biases were observed).

Simulation results for HR=1.25 were globally similar to those for HR=2: although somewhat reduced in magnitude, under-estimation bias increasing with higher proportion of events was again seen in NCC analyses using Efron's or Breslow's approximation (data not shown).

2.3 Results for a time-varying exposure

Results for a TVE were similar to those obtained for a fixed exposure, whether HR=2 or HR=1.25. The results of the simulations for TVE and HR=2 are shown in Figure 2 and Table 2. The relative biases of estimates obtained from full cohort and NCC analyses with Efron's approximation for handling ties are reported in Figure 2. Frequency of ties was comparable to that observed with a fixed exposure across all nine scenarios (Table in Appendix 3). The cohort design yielded unbiased estimates (relative bias <2.5%) while NCC estimates with Breslow's and Efron's approximations were again biased to the null, sometimes greatly (maximal relative bias from 27% for 1 control to 7% for 5 controls). The same trend of an increasing bias as the proportion of events increased (e.g., for 10% of exposed subjects, the bias increased from 12% to 19% and then 27% for respectively 5%, 10% and 25% of subjects who experienced an event) was apparent. There was no clear trend, however, with the proportion of exposed subjects. Breslow's approximation for handling ties resulted in larger bias in the estimates from the NCC design than Efron's approximation in all scenarios (Table 2). These biases were greatly reduced when using the exact method or modified *ccwc* approach with only one case in each stratum (maximal relative bias from 41% and 27% for Breslow's and Efron's approximations respectively to 2% for the exact method and modified *ccwc* approach).

As expected, Table 2 shows smaller SEE's and RMSE's for the full cohort than for the NCC design in all scenarios. Predictably, precision of estimates improved for both designs as

the proportion of events, exposure prevalence and, for NCC analyses, number of controls per case increased. Similar to time-invariant exposure, for NCC analysis with Breslow's and Efron's approximations where estimates produced larger bias, RMSE was greater for higher proportion of events and 1 control per case and SEE's were systematically smaller than those from the NCC analysis with the exact method or modifed *ccwc* approach.

Similarly to what we observed for a fixed exposure, coverage probability of the cohort design was close to 95% for all scenarios while that of NCC design fell below nominal especially for Breslow's and Efron's approximations with higher proportions of events and subjects exposed (scenarios for which most important biases were observed).

3 Real data application

3.1 Population and methods

To compare the estimates obtained from the two study designs using real-world data, we examined the risk of breast cancer associated with ever use of menopausal hormone therapy (MHT) in the French E3N women cohort (*Étude Épidémiologique auprès de Femmes de la Mutuelle Générale de l'Éducation Nationale*) (Fournier et al., 2014). 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) and all participants in the study provided informed consent.

Our study population was composed of 38,091 postmenopausal women who were free of any cancer when they completed a detailed questionnaire on their past use of any MHT in 1992. The participants were actively followed-up until 2008 through self-administered questionnaires every 2-3 years, that included updated information on recent MHT use. Among them, 17,194 (45.1%) had ever used MHT at baseline while 7,805 (20.5%) started using MHT during follow-up. A total of 2,261 (5.9%) invasive breast cancers were diagnosed during 532,925 person-years of follow-up (incidence of 424 cases per 10^5 person-years).

In the full cohort, the associations of each of the two MHT exposure metrics: (*i*) fixed (assessed at baseline: any prior use before cohort entry) and (*ii*) time-varying (any previous MHT use before a given time during follow-up) with breast cancer incidence were estimated using the univariate Cox proportional hazards regression model with Efron's approximation for ties. Although these two exposure metrics resemble those in simulations, it is important to note the following differences. The first metric used fixed MHT exposure based on baseline status, thus ignoring possible exposure changes during follow-up, whereas exposure remained fixed

throughout follow-up in the simulations. Regarding the second metric, a sizeable proportion of women were already exposed at baseline in the E3N data whereas all subjects were considered un-exposed at the beginning of their follow-up in the simulations.

The NCC analyses with the conditional logistic regression model were repeated 100 times for each of 1 and 5 controls per case, using four different approaches to handle ties: the exact method, Breslow's and Efron's approximations, as well as the modified *ccwc* approach. In secondary analyses, to increase the number of ties, the time unit of analysis was changed from day to month. We compared the estimates obtained from each approach using un-adjusted HR with empirical 95% CI, relative difference of log-HR (difference between mean log-HR of NCC and log-HR of cohort design over the latter) and relative width of 95% CI (ratio of the width of the 95% CI from each of NCC analyses over the width of 95% CI from full cohort design).

3.2 Results

With daily time units, there were 1,984 distinct times at which at least one event occurred. Among them, tied events occurred at 253 event times but the number of ties was relatively low with at most 4 cases occurring at the same time (see Table 1 in Appendix 4). Table 3 compares the estimates obtained with full cohort (top row) *versus* the mean of 100 NCC-based estimates, observed using different NCC analytical strategies. All full cohort and NCC analyses indicated that both time-invariant and time-varying MHT exposures were associated with a statistically significant albeit moderate increase of breast cancer risk (Table 3). Estimated HR for full cohort was 1.23 (95% CI, 1.13 to 1.34) and 1.49 (95% CI, 1.36 to 1.64) for, respectively, time-invariant and time-varying exposures. Mean NCC estimates for 1 control per case using the exact method or the modified *ccwc* approach were closer to those obtained from the full cohort analysis than mean NCC estimates obtained using Breslow's or Efron's approximation (relative difference in log-HR of about 1% *versus* 5% to 10%, and <1% *versus* >4% for respectively time-invariant and time-varying ever use of MHT, Table 3). As in simulations (sections 2.2 and 2.3), the Breslow and Efron methods suggested somewhat weaker associations than the full cohort analyses. Consistently with simulation results, differences were much smaller for all NCC approaches with 5 controls.

When the time unit of analysis was changed to months, the number of ties increased markedly. There were 363 distinct times at which at least one event occurred. Among them, tied events occurred at 309 event times and the number of ties was high, with up to 20 cases

occurring at the same event time (Table 2 in Appendix 4). NCC estimates with Breslow's and Efron's approximations differed considerably from those obtained from the full cohort analysis (relative difference in log-HR up to, respectively, 45% and 23% for 1 control per case and time-invariant ever use of MHT, Table 4). Consistently with the simulation results, using the Breslow and Efron methods for handling ties resulted in substantial biases toward the null, relative to the estimates yielded by full cohort analyses (Table 4). The use of the exact method or modified *ccwc* approach for NCC analyses greatly reduced this difference (relative difference in log-HR fell to 2% and 4% for 1 control per case and time-invariant ever use of MHT, Table 4). Overall, although exposure features were not quite identical to those in simulations (see above), these results were consistent with those from simulations both for fixed exposure and TVE.

4 Discussion

In this paper, we have performed a simulation study to investigate and compare the performances of full cohort *versus* different NCC analyses for estimating associations between either fixed exposure or TVE and the risk of an adverse event. Our initial findings suggested that NCC estimates could be substantially biased, especially when only a single control was matched to each case. Bias of NCC estimates increased with higher proportion of events, while estimates from the full cohort analysis remained unbiased across all simulated scenarios. However, it appeared that such bias mostly resulted from tied events so that, once appropriately taken into account with the exact method or modified *ccwc* approach, NCC estimates were almost free of bias and quite close to those from the full cohort analysis.

To the best of our knowledge, most studies comparing the relative performances of full cohort and NCC analyses have been limited to fixed exposures (Bertke et al., 2013; Essebag et al., 2005). In our simulations, the performances of cohort and NCC analyses were comparable for fixed exposure and TVE. While our study was limited to simple TVE with a single change from unexposed to exposed during the follow-up, further work should investigate more complex TVEs with multiple changes (Desquilbet & Meyer, 2005), decreasing hazard ratios (Brickner, 2015) or with TVE representing cumulative effects of past exposures (Abrahamowicz et al., 2012; Pazzagli et al., 2018; Sylvestre & Abrahamowicz, 2009). Subsequent work is also needed to examine if there could be an advantage of matched NCC analysis over adjusted cohort analysis while considering potential confounders. Further work is also warranted to compare cohort and NCC design in the presence of competing events (Austin et al., 2012).

In all the scenarios we investigated, the NCC design with Breslow's (1974) or Efron's (1977) approximations for handling ties in event times resulted in greater bias and lower precision than the full cohort design. The loss of precision of NCC estimates compared to those of the full cohort analysis was expected because in the cohort design all data are used whereas in the NCC only data on cases and selected controls are used. The bias still apparent with 5 controls per case, however, was less expected and contradictory to the general idea that NCC analyses with a sufficient number of controls per case provide estimates close to those of full cohort analyses (Breslow & Day 1987; Goldstein & Langholz, 1992; Pang 1999, Essebag et al., 2005; Bertke et al., 2013). Moreover, counter-intuitively, bias tended to increase with higher proportion of events, a similar finding to that reported by Austin et al. (2012) whereas one could expect more accurate estimation as the wealth of information increases. In our case, however, the observed biases most likely resulted from tied events and the way they were handled. Those biases were indeed markedly reduced (at the expense of slightly lower precision for lower proportion of events) when the exact method (Peto, 1972) or modified *ccwc* approach were used instead of Breslow's or Efron's approximation.

The observed trend of increasing biases with greater proportion of events in Breslow's and Efron's approximations is consistent with the previous finding that these approximations deteriorate as the proportion of tied events increases (Farewell and Prentice 1980). In our simulation setting, as the proportion of events, *i.e.*, the number of cases in a fixed-size cohort of 5,000 individuals, increases, so does the probability of having two or more events occurring at the same time within a limited timeframe of 730 days. Increasing the proportion of tied events had strong detrimental impact on NCC estimates while estimates from the full cohort analysis were hardly affected and remained virtually unbiased. Indeed, in full cohort analyses, the proportion of ties within risk sets at each event time is considerably diluted among all subjects belonging to the risk set at that time whereas, in NCC analyses, the risk set is much smaller, constrained to the cases occurring at this specific event time and their matched controls. Formerly, Farewell and Prentice (1980) had shown through simulations of case-control data with 5, 20 and 50 cases (*i.e.*, tied events) per stratum that NCC estimates with Efron's approximation, and to an even greater extent with Breslow's approximation, are affected by systematic under-estimation bias. They, therefore, advised not to use these two approximation methods for case-control or cohort studies in which the number of ties per event time is large. Other authors have advocated the use of the exact method (Peto, 1972) to analyze case-control studies when the proportion of cases in each case-control stratum is not small enough (Langholz

& Richardson, 2010). This problem can also be avoided in case-control studies by arranging the case-control sets in such a way as to obtain a single case per stratum and, thus, not to have tied events in the strata. This solution requires, however, that each of the cases occurring at the same event time may be allocated a sufficient number of matched controls. This motivated our modification of the *ccwc* function in R, in accordance with the sampling procedure performed in SAS macro %nCCsampling (Desai et al., 2016).

We are concerned that researchers may not be aware of the potential deterioration of NCC estimation in the presence of numerous tied events and that this problem may be insufficiently taken into account in practice. In particular, researchers should be aware of the possible limitations of the default options used in the available statistical software for implementing conditional logistic regression. In R, while the *clogit* function uses the exact method by default, it calls the *coxph* function which uses Efron's approximation by default. In SAS, conditional logistic regression can be performed using either the PHREG procedure with a STRATA statement, which employs Breslow's approximation as the default, or the LOGISTIC procedure with a STRATA statement. The latter procedure relies upon the exact method for discrete data assuming that events really occurred at exactly the same time (Allison, 2010). This procedure is called, e.g., by the %nCCsampling macro which performs incidence density sampling for NCC analysis (Desai et al., 2016). In our simulated and example data, especially when a month was taken as the time unit, ties resulted from imprecise measurement of continuous time. Consequently, the exact method assuming that there is a true but unknown ordering for the tied event times should be used. The computation of the exact likelihood, that considers all possible orderings of tied events, "can be a daunting task" (Allison, 2010) for full cohort analyses but, in our experience, resulted in a reasonable computation time for NCC analyses.

In conclusion, our simulation study showed that NCC analyses with Breslow's or Efron's approximations could lead to substantial bias when there is a large number of tied events in the data. However, once ties were taken correctly into account, NCC estimates were almost free of bias and close to those from the full cohort analysis. We strongly recommend using the exact or modified *ccwc* method in NCC analyses when there are many tied event times.

References

Abrahamowicz, M., Beauchamp, M.-E., & Sylvestre, M.-P. (2012). Comparison of alternative models for linking drug exposure with adverse effects. *Statistics in Medicine*, *31*(11–12), 1014–1030.

Allison, P. D. (2010). *Survival analysis using SAS: A practical guide* (2nd ed). SAS Press.

Austin, P. C., Anderson, G. M., Cigsar, C., & Gruneir, A. (2012). Comparing the cohort design and the nested case–control design in the presence of both time-invariant and time-dependent treatment and competing risks: Bias and precision. *Pharmacoepidemiology and Drug Safety*, *21*(7), 714–724. https://doi.org/10.1002/pds.3299

Bertke, S., Hein, M., Schubauer-Berigan, M., & Deddens, J. (2013). A simulation study of relative efficiency and bias in the nested case–control study design. *Epidemiologic Methods*, *2*(1).

Borucka, J. (2014). Methods of handling tied events in the Cox proportional hazard model. *Studia Oeconomica Posnaniensia*, *2*(2), 263.

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

Breslow, N. E., Lubin, J. H., Marek, P., & Langholz, B. (1983). Multiplicative models and cohort analysis. *Journal of the American Statistical Association*, *78*(381), 1–12.

Brickner, C. P. (2015). *Estimating the relationship between a transient effect and the onset of an acute event: a comparison of the case-crossover design and cohort design*. Rutgers, The State University of New Jersey.

Cox, D.R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, *34*(2), 187-202.

Cox DR (1975). Partial likelihood. *Biometrika*, *62*(2), 269–276.

Desai, R. J., Glynn, R. J., Wang, S., & Gagne, J. J. (2016). Performance of Disease Risk Score Matching in Nested Case-Control Studies: A Simulation Study. *American Journal of Epidemiology*, *183*(10), 949–957.

Desquilbet, L., & Meyer, L. (2005). Variables dépendantes du temps dans le modèle de Cox Théorie et pratique. *Revue d'Épidémiologie et de Santé Publique*, *53*(1), 51–68.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, *72*(359), 557–565.

Essebag, V., Platt, R. W., Abrahamowicz, M., & Pilote, L. (2005). Comparison of nested case-control and survival analysis methodologies for analysis of time-dependent exposure. *BMC Medical Research Methodology*, *5*(1), 5.

Etminan, M. (2004). Pharmacoepidemiology II: The Nested Case-Control Study—A novel approach in pharmacoepidemiologic research. *Pharmacotherapy*, *24*(9), 1105–1109.

Farewel, V. T., & Prentice, R. L. (1980). The approximation of partial likelihood with emphasis on case-control studies. *Biometrika*, *67*(2), 273–278.

Fournier, A., Mesrine, S., Dossus, L., Boutron-Ruault, M.-C., Clavel-Chapelon, F., & Chabbert-Buffet, N. (2014). Risk of breast cancer after stopping menopausal hormone therapy in the E3N cohort. *Breast Cancer Research and Treatment*, *145*(2), 535–543.

Goldstein, L., & Langholz, B. (1992). Asymptotic theory for nested case-control sampling in the Cox regression model. *Annals of Statistics*, *20*(4), 1903–1928.

Hertz-Picciotto, I., & Rockhill, B. (1997). Validity and efficiency of approximation methods for tied survival times in Cox regression. *Biometrics*, *53*(3), 1151–1156.

Klein, J. P., van Houwelingen, H. C., Ibrahim, J. G., & Scheike, T. H. (Eds.). (2014). *Handbook of Survival Analysis*. CRC Press, Taylor & Francis Group.

Langholz, B. (2014). Case-Control Study, Nested. In N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri, & J. L. Teugels (Eds.), *Wiley StatsRef: Statistics Reference Online*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118445112.stat05121

Langholz, B., & Richardson, D. B. (2010). Fitting general relative risk models for survival time and matched case-control analysis. *American Journal of Epidemiology*, *171*(3), 377–383.

Liddell, F. D. K., McDonald, J. C., Thomas, D. C., & Cunliffe, S. V. (1977). Methods of cohort analysis: appraisal by application to asbestos mining. *Journal of the Royal Statistical Society. Series A*, *140*(4), 469.

Lubin, J. H., & Gail, M. H. (1984). Biased selection of controls for case-control analyses of cohort studies. *Biometrics*, *40*(1), 63–75.

Mackenzie, T., & Abrahamowicz, M. (2002). Marginal and hazard ratio specific random data generation: Applications to semi-parametric bootstrapping. *Statistics and Computing*, *12*(3), 245–252.

O'Quigley, J. (2008). *Proportional Hazards Regression*. Springer New York.

Pazzagli, L., Linder, M., Zhang, M., Vago, E., Stang, P., Myers, D., Andersen, M., & Bahmanyar, S. (2018). Methods for time-varying exposure related problems in pharmacoepidemiology: An overview. *Pharmacoepidemiology and Drug Safety*, *27*(2), 148–160.

Peto, R. (1972). Discussion of: Regression models and life tables, by D.R. Cox. *Journal of the Royal Statistical Society, Series B*, *34*(2), 205-207.

Prentice, R. L., & Breslow, N. E. (1978). Retrospective studies and failure time models. *Biometrika*, *65*(1), 153–158.

Ryan, T. P., & Woodall, W. H. (2005). The most-cited statistical papers. *Journal of Applied Statistics*, *32*(5), 461–474.

Sylvestre, M.-P., & Abrahamowicz, M. (2008). Comparison of algorithms to generate event times conditional on time-dependent covariates. *Statistics in Medicine*, *27*(14), 2618–2634.

Sylvestre, M.-P., & Abrahamowicz, M. (2009). Flexible modeling of the cumulative effects of time-dependent exposures on the hazard. *Statistics in Medicine*, *28*(27), 3437–3453.

Sylvestre, M.-P., Edens, T., MacKenzie, T., & Abrahamowicz, M. (2015). *PermAlgo: Permutational Algorithm to Simulate Survival Data* (1.1) [Computer software]. https://CRAN.R-project.org/package=PermAlgo

TABLE 1 Simulations results of the full cohort (with Efron's method) and nested case-control (for 1 and 5 controls per case with the exact method, Breslow's and Efron's approximations and modified *ccwc* approach) analyses for time-invariant exposure and true hazard ratio of 2

10% subjects exposed												
Statistical methods	5% events				10% events				25% events			
	Relative bias (%)	SEE	RMSE	CP (%)	Relative bias (%)	SEE	RMSE	CP (%)	Relative bias (%)	SEE	RMSE	CP (%)
Full cohort	**-0.636**	0.167	0.166	95.3	**0.200**	0.116	0.113	95.5	**0.406**	0.076	0.076	95.4
NCC (1:1) B	-16.227	0.242	0.249	93.1	-28.039	0.153	0.231	80.5	-41.119	0.090	0.292	4.6
NCC (1:1) Ef	**-10.256**	0.244	0.251	93.8	**-18.677**	0.155	0.194	87.6	**-27.263**	0.091	0.205	45.7
NCC (1:1) Ex	3.326	0.278	0.293	94.7	1.309	0.191	0.194	95.2	0.878	0.125	0.124	94.6
NCC (1:1) *	3.059	0.279	0.284	95.3	1.105	0.192	0.188	96.2	1.135	0.127	0.129	94.0
NCC (5:1) B	-4.120	0.188	0.181	95.1	-6.987	0.127	0.129	95.5	-11.896	0.081	0.111	84.2
NCC (5:1) Ef	**-2.311**	0.188	0.184	95.0	**-3.673**	0.128	0.128	95.4	**-6.190**	0.081	0.092	91.6
NCC (5:1) Ex	-0.022	0.193	0.189	95.1	0.274	0.134	0.131	95.3	0.610	0.088	0.089	95.4
NCC (5:1) *	-0.591	0.193	0.188	95.9	0.614	0.134	0.126	96.0	1.062	0.088	0.087	95.2
25% subjects exposed												
Statistical methods	5% events				10% events				25% events			
	Relative bias (%)	SEE	RMSE	CP (%)	Relative bias (%)	SEE	RMSE	CP (%)	Relative bias (%)	SEE	RMSE	CP (%)
Full cohort	**-0.498**	0.130	0.129	95.4	**-0.076**	0.090	0.089	95.8	**0.137**	0.058	0.057	96.0
NCC (1:1) B	-15.659	0.181	0.195	93.2	-24.881	0.116	0.197	72.8	-38.109	0.068	0.269	0.2
NCC (1:1) Ef	**-10.193**	0.182	0.187	94.0	**-15.829**	0.117	0.154	86.1	**-24.480**	0.068	0.181	28.4
NCC (1:1) Ex	0.481	0.202	0.201	95.3	1.209	0.139	0.136	95.9	0.550	0.090	0.088	94.8
NCC (1:1) *	1.640	0.203	0.200	96.1	0.731	0.140	0.139	95.1	0.887	0.091	0.091	95.7
NCC (5:1) B	-3.948	0.143	0.144	95.4	-6.259	0.098	0.102	94.0	-10.972	0.061	0.095	77.8
NCC (5:1) Ef	**-2.361**	0.143	0.145	95.4	**-3.328**	0.098	0.098	95.0	**-5.812**	0.061	0.073	91.0
NCC (5:1) Ex	-0.472	0.146	0.147	94.9	0.046	0.101	0.099	96.1	0.100	0.066	0.065	94.3
NCC (5:1) *	-0.007	0.146	0.146	95.5	0.319	0.101	0.101	95.2	0.188	0.066	0.065	96.2

50% subjects exposed												
Statistical methods	5% events				10% events				25% events			
	Relative bias (%)	SEE	RMSE	CP (%)	Relative bias (%)	SEE	RMSE	CP (%)	Relative bias (%)	SEE	RMSE	CP (%)
Full cohort	**0.207**	0.135	0.137	94.8	**0.235**	0.093	0.091	94.8	**0.089**	0.059	0.059	95.5
NCC (1:1) B	-13.234	0.177	0.192	92.6	-21.738	0.115	0.180	76.0	-33.994	0.067	0.242	2.8
NCC (1:1) Ef	**-8.408**	0.177	0.187	93.9	**-13.388**	0.116	0.143	88.8	**-20.890**	0.068	0.158	42.1
NCC (1:1) Ex	-0.345	0.190	0.194	94.9	-0.110	0.130	0.126	95.9	-0.227	0.082	0.081	95.1
NCC (1:1) *	0.557	0.191	0.198	93.9	-0.523	0.131	0.130	95.6	0.277	0.084	0.083	95.8
NCC (5:1) B	-2.741	0.144	0.146	94.9	-4.922	0.098	0.099	94.5	-9.192	0.061	0.087	82.5
NCC (5:1) Ef	**-1.397**	0.144	0.147	94.5	**-2.446**	0.098	0.097	95.4	**-4.752**	0.061	0.070	91.5
NCC (5:1) Ex	0.093	0.146	0.149	94.0	0.292	0.101	0.098	96.0	0.119	0.064	0.065	94.9
NCC (5:1) *	0.342	0.146	0.147	94.9	0.221	0.101	0.100	95.3	-0.046	0.064	0.065	94.7

SEE Average standard error estimator; RMSE Root mean square error; CP Coverage probability; NCC Nested case-control; B Breslow; Ef Efron, Ex Exact; * Modified *ccwc* approach. Results in bold are those displayed in Figure 1

TABLE 2 Simulations results of the full cohort (with Efron method) and nested case-control (for 1 and 5 controls per case with exact method, Breslow and Efron's approximations and modified *ccwc* approach) analyses for time-varying exposure and true hazard ratio of 2

10% subjects exposed												
Statistical methods	5% events				10% events				25% events			
	Relative bias (%)	SEE	RMSE	CP (%)	Relative bias (%)	SEE	RMSE	CP (%)	Relative bias (%)	SEE	RMSE	CP (%)
Full cohort	-2.328	0.226	0.226	96.1	-0.611	0.157	0.146	96.9	0.506	0.104	0.100	96.7
NCC (1 :1) B	-18.475	0.331	0.318	95.9	-28.404	0.210	0.258	89.5	-41.108	0.123	0.297	29.2
NCC (1 :1) Ef	**-12.283**	0.335	0.328	95.2	**-19.010**	0.213	0.232	92.4	**-27.395**	0.125	0.218	70.6
NCC (1 :1) Ex	2.200	0.385	0.390	96.0	2.265	0.265	0.264	96.0	1.689	0.174	0.168	96.0
NCC (1 :1) *	2.188	0.385	0.383	96.1	2.061	0.266	0.265	95.5	1.602	0.175	0.182	94.7
NCC (5 :1) B	-6.503	0.256	0.259	94.7	-7.464	0.174	0.164	96.7	-12.415	0.111	0.130	92.0
NCC (5 :1) Ef	**-4.591**	0.256	0.264	94.1	**-4.080**	0.174	0.165	96.5	**-6.716**	0.111	0.116	95.3
NCC (5 :1) Ex	-2.263	0.263	0.271	93.7	-0.003	0.183	0.172	96.1	0.127	0.121	0.116	96.5
NCC (5 :1) *	-1.924	0.263	0.259	95.8	-0.526	0.183	0.173	96.8	0.481	0.121	0.123	95.7
25% subjects exposed												
Statistical methods	5% events				10% events				25% events			
	Relative bias (%)	SEE	RMSE	CP (%)	Relative bias (%)	SEE	RMSE	CP (%)	Relative bias (%)	SEE	RMSE	CP (%)
Full cohort	-1.616	0.163	0.170	94.2	0.183	0.113	0.113	95.1	-0.820	0.074	0.074	94.5
NCC (1 :1) B	-18.136	0.232	0.252	92.4	-26.770	0.148	0.223	79.6	-39.979	0.087	0.284	4.0
NCC (1 :1) Ef	**-12.551**	0.234	0.251	93.0	**-17.568**	0.150	0.186	88.3	**-26.514**	0.088	0.200	45.9
NCC (1 :1) Ex	-0.552	0.263	0.279	93.8	1.085	0.182	0.185	94.9	-0.093	0.119	0.116	95.6
NCC (1 :1) *	0.992	0.265	0.273	94.2	2.183	0.183	0.187	95.4	-0.482	0.120	0.119	94.7
NCC (5 :1) B	-5.158	0.182	0.189	94.3	-6.632	0.123	0.127	94.0	-12.451	0.078	0.112	84.5
NCC (5 :1) Ef	**-3.458**	0.182	0.191	94.3	**-3.471**	0.124	0.125	94.9	**-7.119**	0.079	0.091	90.8
NCC (5 :1) Ex	-1.344	0.186	0.195	94.3	0.226	0.129	0.129	94.5	-0.887	0.085	0.083	94.4
NCC (5 :1) *	-1.313	0.186	0.192	94.3	0.074	0.129	0.132	94.6	-1.081	0.085	0.087	94.1

50% subjects exposed												
Statistical methods	5% events				10% events				25% events			
	Relative bias (%)	SEE	RMSE	CP (%)	Relative bias (%)	SEE	RMSE	CP (%)	Relative bias (%)	SEE	RMSE	CP (%)
Full cohort	-1.766	0.143	0.147	94.0	-0.938	0.099	0.099	94.3	-0.044	0.064	0.066	94.6
NCC (1 :1) B	-17.164	0.196	0.217	91.7	-25.875	0.127	0.210	72.6	-37.784	0.074	0.268	2.3
NCC (1 :1) Ef	**-11.772**	0.198	0.211	92.6	**-17.178**	0.128	0.171	85.7	**-24.454**	0.075	0.184	37.3
NCC (1 :1) Ex	-1.807	0.218	0.221	94.7	-1.127	0.150	0.151	94.2	-0.543	0.097	0.099	94.4
NCC (1 :1) *	-0.228	0.220	0.220	95.7	-0.444	0.152	0.151	94.8	-0.233	0.098	0.097	95.7
NCC (5 :1) B	-5.204	0.157	0.159	93.9	-7.162	0.107	0.115	92.9	-10.977	0.067	0.100	79.4
NCC (5 :1) Ef	**-3.611**	0.157	0.160	93.8	**-4.305**	0.107	0.111	94.1	**-5.983**	0.067	0.081	89.2
NCC (5 :1) Ex	-1.843	0.160	0.162	93.7	-1.079	0.111	0.112	95.0	-0.287	0.072	0.074	94.4
NCC (5 :1) *	-2.668	0.160	0.165	94.2	-1.209	0.111	0.108	95.8	-0.235	0.072	0.074	94.6

Average standard error estimator; RMSE Root mean square error; CP Coverage probability; NCC Nested case-control; B Breslow, Ef Efron, Ex Exact; * Modified *ccwc* approach. Resu
ld are those displayed in Figure 2

TABLE 3 Results of the E3N cohort analyses with day as the time unit and 100 nested case-control analyses repeating the random sampling of controls

Statistical method	Time-invariant ever use of MHT					Time-varying ever use of MHT				
	HR† (95%CI)	Min HR	Max HR	RD of log-HR	RW of 95%CI	HR† (95%CI)	Min HR	Max HR	RD of log-HR	RW of 95%CI
Full cohort	1.229 (1.131 – 1.335)	-	-	Ref.	Ref.	1.494 (1.364 – 1.637)	-	-	Ref.	Ref.
NCC (1 :1) B	1.204 (1.077 – 1.346)	1.109	1.327	-10.12%	1.315	1.439 (1.276 – 1.623)	1.264	1.598	-9.61%	1.270
NCC (1 :1) Ef	1.217 (1.088 – 1.361)	1.113	1.344	-5.07%	1.333	1.470 (1.303 – 1.658)	1.282	1.636	-4.36%	1.300
NCC (1 :1) Ex	1.232 (1.095 – 1.387)	1.124	1.375	1.06%	1.428	1.498 (1.320 – 1.700)	1.296	1.686	0.45%	1.391
NCC (1 :1) *	1.227 (1.091 – 1.381)	1.079	1.353	-0.96%	1.418	1.499 (1.321 – 1.702)	1.340	1.669	0.64%	1.396
NCC (5 :1) B	1.222 (1.117 – 1.337)	1.170	1.292	-2.83%	1.078	1.488 (1.349 – 1.642)	1.421	1.562	-1.06%	1.072
NCC (5 :1) Ef	1.225 (1.120 – 1.341)	1.173	1.297	-1.39%	1.081	1.497 (1.356 – 1.651)	1.429	1.572	0.38%	1.078
NCC (5 :1) Ex	1.228 (1.121 – 1.335)	1.175	1.300	-0.42%	1.096	1.501 (1.359 – 1.658)	1.433	1.578	1.18%	1.092
NCC (5 :1) *	1.231 (1.124 – 1.348)	1.171	1.304	0.87%	1.098	1.503 (1.361 – 1.660)	1.435	1.575	1.54%	1.094

HR (95%CI) Hazard ratio (with 95% confidence interval); HR† Mean HR for NCC analysis; RD Relative difference; RW Relative width; NCC Nested case-control; B Breslow; Ef Efron; Ex Exact; *Modified ccwc approach

TABLE 4: Results of the E3N cohort analyses with month as the time unit and 100 nested case-control analyses repeating the random sampling of controls

Statistical method	Time-invariant ever use of MHT					Time-varying ever use of MHT				
	HR† (95%CI)	Min HR	Max HR	RD of log-HR	RW of 95%CI	HR† (95%CI)	Min HR	Max HR	RD of log-HR	RW of 95%CI
Full cohort	1.227 (1.129 – 1.333)	-	-	Ref.	Ref.	1.491 (1.361 – 1.634)	-	-	Ref.	Ref.
NCC (1 :1) B	1.119 (1.027 – 1.221)	1.063	1.178	-44.93%	0.951	1.260 (1.146 – 1.386)	1.178	1.338	-42.28%	0.880
NCC (1 :1) Ef	1.164 (1.067 – 1.270)	1.088	1.243	-25.77%	0.994	1.360 (1.236 – 1.496)	1.242	1.466	-23.20%	0.955
NCC (1 :1) Ex	1.232 (1.095 – 1.386)	1.119	1.354	1.65%	1.425	1.510 (1.332 – 1.711)	1.341	1.677	2.87%	1.391
NCC (1 :1) *	1.217 (1.082 – 1.370)	1.063	1.331	-4.35%	1.409	1.500 (1.321 – 1.703)	1.340	1.676	1.09%	1.398
NCC (5 :1) B	1.190 (1.094 – 1.295)	1.126	1.241	-14.86%	0.983	1.414 (1.289 – 1.551)	1.365	1.472	-13.45%	0.960
NCC (5 :1) Ef	1.208 (1.111 – 1.315)	1.139	1.264	-7.55%	0.999	1.453 (1.324 – 1.594)	1.399	1.518	-6.61%	0.988
NCC (5 :1) Ex	1.226 (1.119 – 1.343)	1.149	1.287	-0.46%	1.095	1.493 (1.352 – 1.648)	1.434	1.564	0.16%	1.086
NCC (5 :1) *	1.231 (1.124 – 1.348)	1.170	1.284	1.40%	1.100	1.495 (1.354 – 1.651)	1.419	1.557	0.52%	1.089

HR (95%CI) Hazard ratio (with 95% confidence interval); HR† Mean HR for NCC analysis; RD Relative difference; RW Relative width; NCC Nested case-control; B Breslow; Ef Efron; Ex Exact; *Modified ccwc approach

FIGURE 1 Relative bias of Log-HR from full cohort and NCC (for 1 and 5 controls per case) analyses with Efron's approximation for handling tied events: time-invariant exposure and true hazard ratio of 2

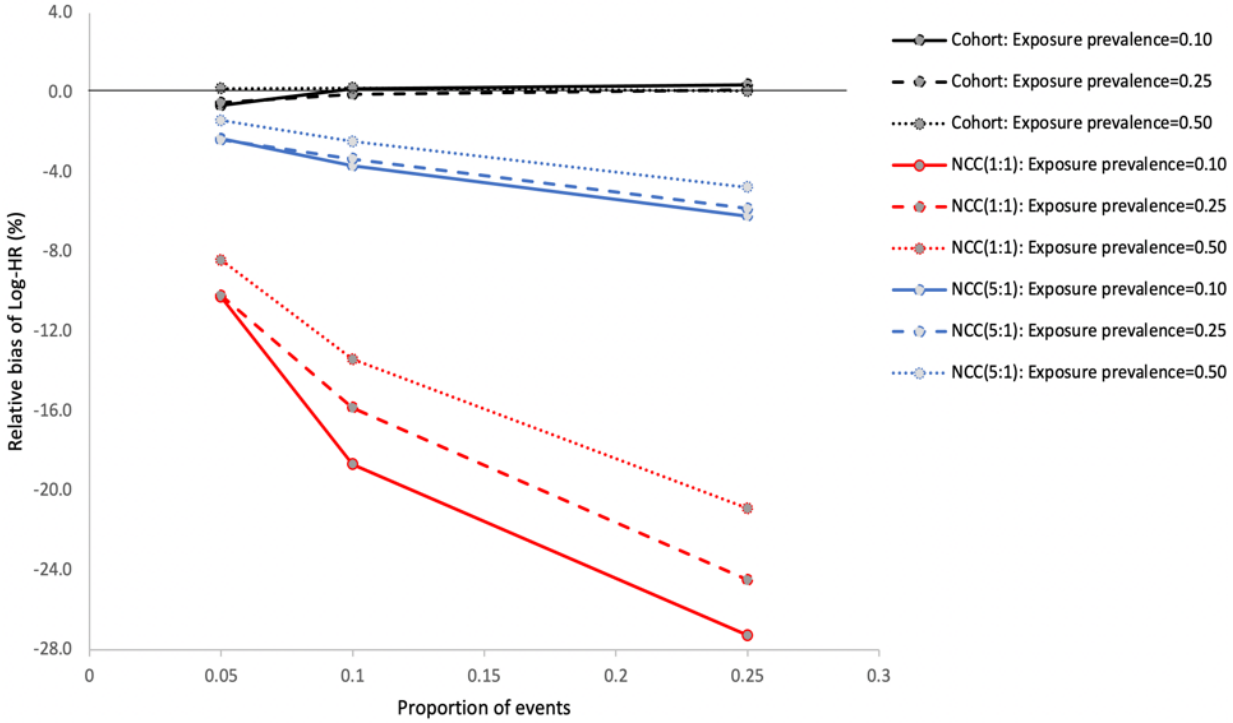
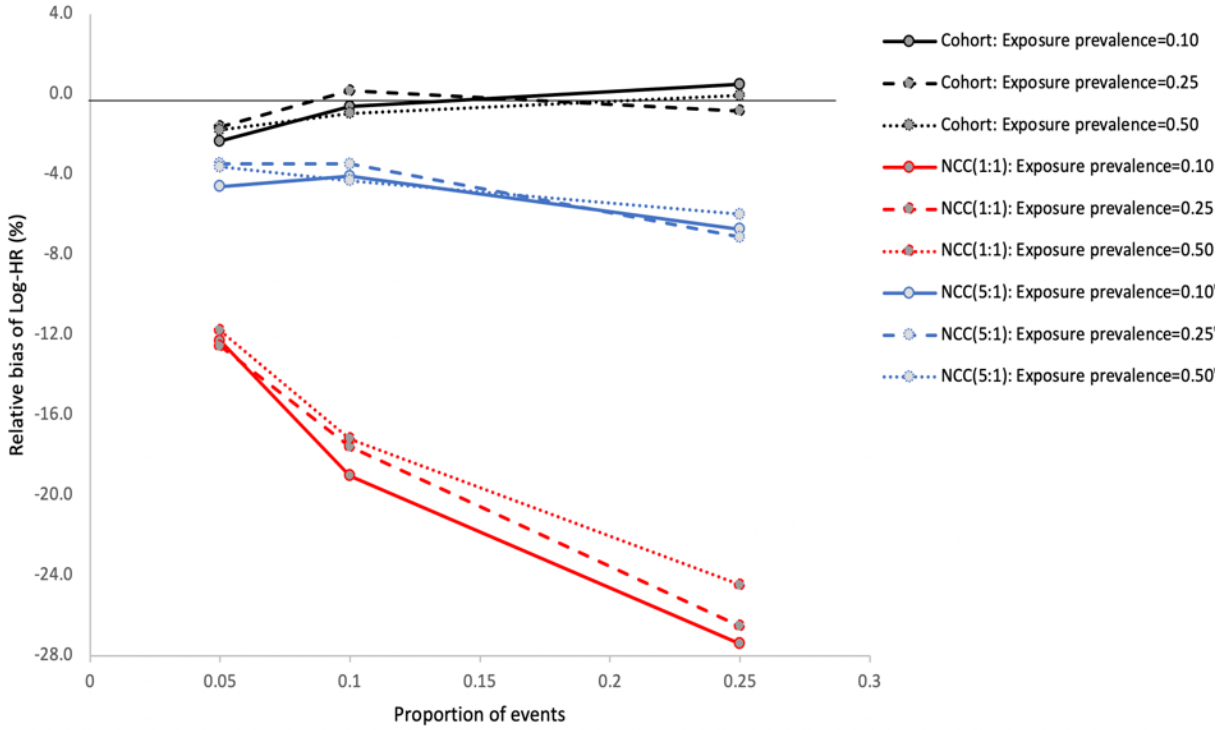


FIGURE 2 Relative bias of Log-HR from full cohort and NCC (for 1 and 5 controls per case) analyses with Efron's approximation for handling tied events: time-varying exposure and true hazard ratio of 2



Appendix 1

R code of the modified *ccwc* function to obtain a single case in each stratum by using a unique identifier for each case (changes to the original *ccwc* function in bold).

```
modified_ccwc <- function (entry = 0, exit, fail, origin = 0, controls = 1, id, match = list(),
                  include = list(), data = NULL, silent = FALSE) # id (identifier variable) added
{
 entry <- eval(substitute(entry), data)
 exit <- eval(substitute(exit), data)
 fail <- eval(substitute(fail), data)
 origin <- eval(substitute(origin), data)
 id <- eval(substitute(id), data) # added
 n <- length(fail)
 if (length(exit) != n)
  stop("All vectors must have same length")
 if (length(entry) != 1 && length(entry) != n)
  stop("All vectors must have same length")
 if (length(id) != n)    # added
  stop("All vectors must have same length") # added
 if (length(origin) == 1) {
  origin <- rep(origin, n)
 }
 else {
  if (length(origin) != n)
   stop("All vectors must have same length")
 }
 t.entry <- as.numeric(entry - origin)
 t.exit <- as.numeric(exit - origin)
 marg <- substitute(match)
 if (mode(marg) == "name") {
  match <- list(eval(marg, data))
  names(match) <- as.character(marg)
 }
 else if (mode(marg) == "call" && marg[[1]] == "list") {
  mnames <- names(marg)
  nm <- length(marg)
  if (!is.null(mnames)) {
   if (nm > 1) {
    for (i in 2:nm) {
     if (mode(marg[[i]]) == "name")
      mnames[i] <- as.character(marg[[i]])
     else stop("illegal argument (match)")
```

```
    }
   }
   else {
    for (i in 2:nm) {
     if (mode(marg[[i]]) == "name")
      mnames[i] <- as.character(marg[[i]])
     else stop("illegal argument (match)")
    }
    mnames[1] <= ""
   }
  }
  names(marg) <- mnames
  match <- eval(marg, data)
 }
 else {
  stop("illegal argument (match)")
 }
 m <- length(match)
 mnames <- names(match)
 if (m > 0) {
  for (i in 1:m) {
   if (length(match[[i]]) != n) {
    stop("incorrect length for matching variable")
   }
  }
 }
 iarg <- substitute(include)
 if (mode(iarg) == "name") {
  include <- list(eval(iarg, data))
  names(include) <- as.character(iarg)
 }
 else if (mode(iarg) == "call" && iarg[[1]] == "list") {
  ni <- length(iarg)
  inames <- names(iarg)
  if (ni > 1) {
   if (!is.null(inames)) {
    for (i in 2:ni) {
     if (mode(iarg[[i]]) == "name")
      inames[i] <- as.character(iarg[[i]])
     else stop("illegal argument (include)")
    }
   }
```

```
    else {
      for (i in 2:ni) {
        if (mode(iarg[[i]]) == "name")
          inames[i] <- as.character(iarg[[i]])
        else stop("illegal argument (include)")
      }
      inames[1] <= ""
    }
  }
  names(iarg) <- inames
  include <- eval(iarg, data)
}
else {
  stop("illegal argument (include)")
}
ni <- length(include)
inames <- names(include)
if (ni > 0) {
  for (i in 1:ni) {
    if (length(include[[i]]) != n) {
      stop("incorrect length for included variable")
    }
  }
}
grp <- rep(1, n)
pd <- 1
if (m > 0) {
  for (im in 1:m) {
    v <- match[[im]]
    if (length(v) != n)
      stop("All vectors must have same length")
    if (!is.factor(v))
      v <- factor(v)
    grp <- grp + pd * (as.numeric(v) - 1)
    pd <- pd * length(levels(v))
  }
}
nn <- (1 + controls) * sum(fail != 0)
pr <- numeric(nn)
sr <- numeric(nn)
tr <- vector("numeric", nn)
fr <- numeric(nn)
```

```
  nn <- 0
  if (!silent) {
    cat("\nSampling risk sets: ")
  }
  set <- 0
  tf <- 0 # added
  nomatch <- 0
  incomplete <- 0
  ties <- FALSE
  fg <- unique(grp[fail != 0])
  for (g in fg) {
    ft <- t.exit[(grp == g) & (fail != 0)] # modified
    ident <- id[(grp == g) & (fail != 0)] # added
    nft <- length(ft) # added
    if (length(ident) != nft) # added
      stop("All vectors must have same length") # added
    for (a in ident) { # added
      if (!silent) {
        cat(".")
      }
      set <- set + 1
      tf <- tf + 1 # added
      case <- (grp == g) & (t.exit == ft[tf]) & (fail != 0) & (id == a) # modified
      ncase <- sum(case)
      if (ncase > 0)
        ties <- TRUE
      noncase <- (grp == g) & (t.entry <= ft[tf]) & ((t.exit > ft[tf]) | ((t.exit == ft[tf]) & fail == 0)) & !case
#modified
      ncont <- controls * ncase
      if (ncont > sum(noncase)) {
        ncont <- sum(noncase)
        if (ncont > 0)
          incomplete <- incomplete + 1
      }
      if (ncont > 0) {
        newnn <- nn + ncase + ncont
        sr[(nn + 1):newnn] <- set
        tr[(nn + 1):newnn] <- ft[tf]   # modified
        fr[(nn + 1):(nn + ncase)] <- 1
        fr[(nn + ncase + 1):newnn] <- 0
        pr[(nn + 1):(nn + ncase)] <- (1:n)[case]
        pr[(nn + ncase + 1):(newnn)] <- sample((1:n)[noncase],
                                               size = ncont)
```

```
      nn <- newnn
    }
    else {
      nomatch <- nomatch + ncase
    }
  }
}
if (!silent) {
  cat("\n")
}
res <- vector("list", 4 + m + ni)
if (nn > 0) {
  res[[1]] <- sr[1:nn]
  res[[2]] <- map <- pr[1:nn]
  res[[3]] <- tr[1:nn] + origin[map]
  res[[4]] <- fr[1:nn]
}
if (m > 0) {
  for (i in 1:m) {
    res[[4 + i]] <- match[[i]][map]
  }
}
if (ni > 0) {
  for (i in 1:ni) {
    res[[4 + m + i]] <- include[[i]][map]
  }
}
names(res) <- c("Set", "Map", "Time", "Fail", mnames, inames)
if (incomplete > 0)
  warning(paste(incomplete, "case-control sets are incomplete"))
if (nomatch > 0)
  warning(paste(nomatch, "cases could not be matched"))
if (ties)
  warning("there were tied failure times")
data.frame(res)
}
```

Appendix 2

TABLE 1 Distribution of event times for 1000 simulated datasets of 5000 individuals each according to proportions of events and exposed subjects for time-invariant exposure and true hazard ratio of 2

10% subjects exposed									
	5% events			10% events			25% events		
	Median	Min	Max	Median	Min	Max	Median	Min	Max
Total number of events, M	249	202	313	521	456	593	1,265	1,173	1,363
Number of distinct event times, K	211	174	259	373	331	413	600	567	636
Number of single event times, M_1	177	149	213	255	211	296	223	178	263
Proportion of tied events, $(M\text{-}K)/M$	15%	8%	23%	28%	23%	33%	53%	50%	56%
25% subjects exposed									
	5% events			10% events			25% events		
	Median	Min	Max	Median	Min	Max	Median	Min	Max
Total number of events, M	249	202	313	521	456	593	1,265	1,173	1,363
Number of distinct event times, K	211	174	259	373	331	413	600	567	636
Number of single event times, M_1	177	149	213	255	211	296	223	178	263
Proportion of tied events, $(M\text{-}K)/M$	15%	8%	23%	28%	23%	33%	53%	50%	56%
50% subjects exposed									
	5% events			10% events			25% events		
	Median	Min	Max	Median	Min	Max	Median	Min	Max
Total number of events, M	249	202	313	521	456	593	1,265	1,173	1,363
Number of distinct event times, K	211	174	259	373	331	413	600	567	636
Number of single event times, M_1	177	149	213	255	211	296	223	178	263
Proportion of tied events, $(M\text{-}K)/M$	15%	8%	23%	28%	23%	33%	53%	50%	56%

Min minimum, Max maximum

$M = \sum_{j\geq 1} jM_j$ is the total number of events, where M_j is the number of event times present j times, M_1 the number of event times present only once

$K = \sum_{j\geq 1} M_j$, is the number of distinct event times

$M - K = \sum_{j\geq 1}(j-1)M_j$ is the number of ties

Appendix 3

TABLE 1 Distribution of event times for 1000 simulated datasets of 5000 individuals each according to proportions of events and exposed subjects for time-varying exposure and true hazard ratio of 2

	5% events			10% events			25% events		
10% subjects exposed									
	Median	Min	Max	Median	Min	Max	Median	Min	Max
Total number of events, M	249	202	314	521	462	592	1,265	1,172	1,354
Number of distinct event times, K	211	174	259	373	334	420	600	569	630
Number of single event times, M_1	178	147	212	255	218	298	223	182	259
Proportion of tied events, $(M\text{-}K)/M$	15%	8%	22%	28%	23%	33%	53%	49%	56%
25% subjects exposed									
	5% events			10% events			25% events		
	Median	Min	Max	Median	Min	Max	Median	Min	Max
Total number of events, M	249	202	302	522	453	585	1,265	1,179	1,366
Number of distinct event times, K	211	175	251	372	330	416	600	571	631
Number of single event times, M_1	177	147	216	255	212	300	223	184	263
Proportion of tied events, $(M\text{-}K)/M$	15%	8%	23%	28%	22%	35%	53%	49%	56%
50% subjects exposed									
	5% events			10% events			25% events		
	Median	Min	Max	Median	Min	Max	Median	Min	Max
Total number of events, M	249	204	300	522	451	582	1,266	1,171	1,367
Number of distinct event times, K	211	174	248	372	332	407	600	566	628
Number of single event times, M_1	177	137	213	255	215	293	223	184	264
Proportion of tied events, $(M\text{-}K)/M$	15%	8%	23%	29%	23%	35%	53%	50%	56%

Min minimum, Max maximum

$M = \sum_{j\geq 1} jM_j$ is the total number of events, where M_j is the number of event times present j times, M_1 the number of event times present only once

$K = \sum_{j\geq 1} M_j$, is the number of distinct event times

$M - K = \sum_{j\geq 1}(j-1)M_j$ is the number of ties

Appendix 4

TABLE 1 Distribution of the 1984 distinct ages (in daily time units) of women at diagnosis of breast cancer according to their number of ties

Number of ties for each age	Number of distinct ages	Total number of tied ages at diagnosis of breast cancer
1 (no ties)	1731	-
2	232	464
3	18	54
4	3	12
Total	1984	530

TABLE 2 Distribution of the 363 distinct ages (in months) of women at diagnosis of breast cancer according to their number of ties

Number of ties for each age	Number of distinct ages	Total number of tied ages at diagnosis of breast cancer
1 (no ties)	54	-
2	43	86
3	33	99
4	20	80
5	24	120
6	33	198
7	20	140
8	26	208
9	25	225
10	22	220
11	22	242
12	6	72
13	13	169
14	7	98
15	5	75
16	4	64
17	2	34
18	1	18
19	1	19
20	2	40
Total	363	2207