UNIVERSITÉ PARIS-SUD 11

FACULTÉ DE MÉDECINE

Année : **2012**

N° attribué par la bibiothèque

|__|__|__|__|__|__|__|__|__|__|

THÈSE

en vue de l'obtention du diplôme de

DOCTEUR DE L'UNIVERSITÉ PARIS-SUD 11

Spécialité : **BIOSTATISTIQUE**

Présentée et soutenue publiquement par

Helena MARTÍ SOLER

Le 4 mai 2012

MODÉLISATION DES DONNÉES D'ENQUÊTES CAS-COHORTE PAR IMPUTATION MULTIPLE : APPLICATION EN ÉPIDÉMIOLOGIE CARDIO-VASCULAIRE

Directeur de thèse : M. Michel CHAVANCE

JURY

M. Denis HÉMON, DR.	Président
M^{me} Hélène JACQMIN-GADDA, DR.	Rapporteur
M^{me} Catherine QUANTIN, Pr.	Rapporteur
M. Norman BRESLOW, Pr.	Examinateur
M. Michel CHAVANCE, DR.	Examinateur
M. Roberto ELOSUA, DR.	Examinateur

Remerciements

En premier lieu, je tiens à remercier mon directeur de thèse Monsieur Michel Chavance de m'avoir accordé sa confiance et de m'avoir permis de réaliser une thèse au sein de l'équipe de biostatistique. Je le remercie pour sa disponibilité et son aide ainsi que de ses efforts pour me faire progresser dans la langue de Molière.

Je remercie Madame Hélène Jacqmin-Gadda et Madame Catherine Quantin d'avoir accepté d'être rapporteuses de ma thèse ainsi que Monsieur Norman Breslow, Monsieur Roberto Elosua et Monsieur Denis Hémon d'être membres du jury de ma thèse.

Je remercie également Monsieur Thierry Moreau, ex-directeur de l'U780 et Madame Pascale Tubert-Bitter, directrice de l'équipe de biostatistique du CESP INSERM 1018 pour leur accueil et leurs conseils, ainsi que l'ensemble des personnes de l'équipe Épidémiologie respiratoire et environnementale pour leur gentillesse.

J'en profite pour remercier la Région Île-de-France et l'Université Paris Descartes pour le financement de ma thèse grâce à l'Allocation de Recherche d'une part et au poste d'ATER d'autre part.

Un grand merci aux nombreux étudiants, doctorants et post-doctorants que j'ai eu le plaisir de rencontrer et qui ont rendu mes années de thèse très agréables. Très particulièrement, je remercie Dorota, Benedicte, Raphaëlle, Hélène et Ismaïl, pour leur accueil dès mon arrivée, leur amabilité et leur soutien tout au long de ces années.

J'accorde également un grand merci très chaleureux à mes chères amies Carmen et Olaya pour leur soutien constant, leurs encouragements, leur amitié et leur contribution à un séjour si agréable à Paris. Je remercie également les gens qui m'ont entouré pendant ces années.

Enfin, je souhaite remercier ma famille, et plus particulièrement ma mère.

Résumé

Les estimateurs pondérés généralement utilisés pour analyser les enquêtes cas-cohorte ne sont pas pleinement efficaces. Or, les enquêtes cas-cohorte sont un cas particulier de données incomplètes où le processus d'observation est contrôlé par les organisateurs de l'étude. Ainsi, des méthodes d'analyse pour données manquant au hasard (MA) peuvent être pertinentes, en particulier, l'imputation multiple, qui utilise toute l'information disponible et permet d'approcher l'estimateur du maximum de vraisemblance partielle.

Cette méthode est fondée sur la génération de plusieurs jeux plausibles de données complétées prenant en compte les différents niveaux d'incertitude sur les données manquantes. Elle permet d'adapter facilement n'importe quel outil statistique disponible pour les données de cohorte, par exemple, l'estimation de la capacité prédictive d'un modèle ou d'une variable additionnelle qui pose des problèmes spécifiques dans les enquêtes cas-cohorte. Nous avons montré que le modèle d'imputation doit être estimé à partir de tous les sujets complètement observés (cas et non-cas) en incluant l'indicatrice de statut parmi les variables explicatives.

Nous avons validé cette approche à l'aide de plusieurs séries de simulations : 1) données complètement simulées, où nous connaissions les vraies valeurs des paramètres, 2) enquêtes cas-cohorte simulées à partir de la cohorte PRIME, où nous ne disposions pas d'une variable de phase-1 (observée sur tous les sujets) fortement prédictive de la variable de phase-2 (incomplètement observée), 3) enquêtes cas-cohorte simulées à partir de la cohorte NWTS, où une variable de phase-1 fortement prédictive de la variable de phase-2 était disponible. Ces simulations ont montré que l'imputation multiple fournissait généralement des estimateurs sans biais des risques relatifs. Pour les variables de phase-1, ils approchaient la précision obtenue par l'analyse de la cohorte complète, ils étaient légèrement plus précis que l'estimateur calibré de Breslow *et al.* (2009b) et surtout que les estimateurs pondérés

classiques. Pour les variables de phase-2, l'estimateur de l'imputation multiple était généralement sans biais et d'une précision supérieure à celle des estimateurs pondérés classiques et analogue à celle de l'estimateur calibré de Breslow *et al.* (2009b). Les résultats des simulations réalisées à partir des données de la cohorte NWTS étaient cependant moins bons pour les effets impliquant la variable de phase-2 : les estimateurs de l'imputation multiple étaient légèrement biaisés et moins précis que les estimateurs pondérés. Cela s'explique par la présence de termes d'interaction impliquant la variable de phase-2 dans le modèle d'analyse, d'où la nécessité d'estimer des modèles d'imputation spécifiques à différentes strates de la cohorte incluant parfois trop peu de cas pour que les conditions asymptotiques soient réunies.

Nous recommandons d'utiliser l'imputation multiple pour obtenir des estimations plus précises des risques relatifs, tout en s'assurant qu'elles sont analogues à celles fournies par les analyses pondérées.

Nos simulations ont également montré que l'imputation multiple fournissait des estimations de la valeur prédictive d'un modèle (C de Harrell) ou d'une variable additionnelle (différence des indices C, NRI ou IDI) analogues à celles fournies par la cohorte complète.

Mots clés : *Enquêtes cas-cohorte, estimateurs pondérés, imputation multiple, capacité prédictive.*

Abstract

The weighted estimators generally used for analyzing case-cohort studies are not fully efficient. However, case-cohort surveys are a special type of incomplete data in which the observation process is controlled by the study organizers. So, methods for analyzing Missing At Random (MAR) data could be appropriate, in particular, multiple imputation, which uses all the available information and allows to approximate the partial maximum likelihood estimator.

This approach is based on the generation of several plausible complete data sets, taking into account all the uncertainty about the missing values. It allows adapting any statistical tool available for cohort data, for instance, estimators of the predictive ability of a model or of an additional variable, which meet specific problems with case-cohort data. We have shown that the imputation model must be estimated on all the completely observed subjects (cases and non-cases) including the case indicator among the explanatory variables.

We validated this approach with several sets of simulations : 1) completely simulated data where the true parameter values were known, 2) case-cohort data simulated from the PRIME cohort, without any phase-1 variable (completely observed) strongly predictive of the phase-2 variable (incompletely observed), 3) case-cohort data simulated from de NWTS cohort, where a phase-1 variable strongly predictive of the phase-2 variable was available. These simulations showed that multiple imputation generally provided unbiased estimates of the risk ratios. For the phase-1 variables, they were almost as precise as the estimates provided by the full cohort, slightly more precise than Breslow *et al.* (2009b) calibrated estimator and still more precise than classical weighted estimators. For the phase-2 variables, the multiple imputation estimator was generally unbiased, with a precision better than classical weighted estimators and similar to Breslow *et al.* (2009b) calibrated estimator.

The simulations performed with the NWTS cohort data provided less satisfactory results for the effects where the phase-2 variable was involved : the multiple imputation estimators were slightly biased and less precise than the weighted estimators. This can be explained by the interactions terms involving the phase-2 variable in the analysis model and the necessity of estimating specific imputation models in different strata not including sometimes enough cases to satisfy the asymptotic conditions.

We advocate the use of multiple imputation for improving the precision of the risk ratios estimates while making sure they are similar to the weighted estimates.

Our simulations also showed that multiple imputation provided estimates of a model predictive value (Harrell's C) or of an additional variable (difference of C indices, NRI or IDI) similar to those obtained from the full cohort.

Keywords : *Case-cohort surveys, weighted estimators, multiple imputation, predictive ability.*

Valorisation scientifique

Publications avec comité de lecture

- Marti H, Chavance M. Multiple imputation analysis of case-cohort studies. *Statistics in Medicine*. 2011 Jun 15 ;30(13) :1595-607. doi : 10.1002/sim.4130.

- Marti H, Carcaillon L, Chavance M, Multiple imputation for estimating hazard ratios and predictive capability in case-cohort surveys, *BMC Medical research methodology*. 2012, accepté.

- Marti H, Chavance M, Les enquêtes cas-cohorte. Revue d'Épidémiologie et de Santé Publique, en révision.

- Straczek C, Ducimetière P, Marti H, Helmer C, Ritchie K, Tzourio C, Empana JP. Clinical utility of standard lipids and apolipoproteins in relation to coronary heart disease events in the elderly. The Three City Study, en préparation.

Communications orales

- Marti H, Chavance M. *Multiple imputation for estimating relative risks and predictive capability in case-cohort surveys. International Biometrics Society (IBS) Channel Network*, 11-13 avril, 2011, Bordeaux, France.

- Marti H, Chavance M. Estimation par imputation multiple du rapport des risques et de la capacitée prédictive dans les enquêtes cas-cohorte. 43èmes Journées de Statistique. Société française de Statistique, 23-27 mai, 2011, Tunisia.

- Marti H, Chavance M. Case-cohort analysis by multiple imputation. Journée de la recherche de la Faculté de Medicine, Paris Sud. 2 avril, 2010, Kremlin-Bicêtre, France (prix poster).

- Marti H, Chavance M. *Case-cohort analysis by multiple imputation. 6th International Meeting. Recent Advances Trends in Statistics Applied to Clinical Trials.*

Statistical methods in biopharmacy. 21-22 septembre 2009, Paris, France.

- Marti H, Chavance M. *Case-cohort analysis by multiple imputation*. 41èmes Journées de Statistique de la Société Française de Statistique, 24 - 29 mai, 2009. Bordeaux, France.

- Marti H, Chavance M. *Case-cohort analysis by multiple imputation*. Journée des Jeunes Chercheurs de la Société française de biométrie, 5 décembre, 2008. Villejuif, France.

- Marti H, Chavance M. *Case-cohort analysis by multiple imputation*. *Epidemiology and Biometry*, (réunion satellite de l'*International Biometric Conference*), 10-11 juillet, 2008, Paris, France.

Intervenante invitée

- Marti H. *Analyzing case-cohort studies : application to cardio-vascular epidemiology*. *Recent study designs in epidemiology*, INSERM Workshop 198, 30 septembre - 2 octobre, 2009. Saint Raphaël, France.

Table des matières

Remerciements i

Résumé iii

Abstract v

Valorisation scientifique vi

Publications avec comité de lecture vi

Communications orales vi

1 Introduction 1

1.1 Contexte 1

1.2 Objectif de la thèse 3

2 Enquêtes cas-cohorte et estimateurs pondérés 5

2.1 Échantillonnage 5

2.2 Estimateurs pondérés 6

2.2.1 Approche pondérée classique 6

2.2.2 Estimation de la variance de l'estimateur pondéré 11

2.2.3 Prise en compte de l'ensemble des données de phase-1 12

2.3 Calcul du nombre de sujets nécessaire dans les études cas-cohorte . . 14

2.4 Mise en œuvre 17

3 Données incomplètes et imputation multiple 24

3.1 Typologie des données incomplètes 24

3.2 Stratégies d'analyse 26

3.2.1 Modélisation des cas complets 26

3.2.2 Indicatrice de données manquantes 27

3.2.3 Pondération par l'inverse de la probabilité 28

3.2.4 Imputation simple . . . 29
3.2.5 Imputation multiple . . . 30

4 Mise en œuvre de l'imputation multiple dans les enquêtes cas-cohorte 38
4.1 Modèle d'imputation . . . 41
4.2 Mise en œuvre avec R . . . 45

5 Capacité prédictive d'un modèle ou d'une variable additionnelle 46
5.1 C de Harrell . . . 47
5.2 NRI et IDI . . . 51

6 Validation par simulations 56
6.1 Données complètement simulées . . . 57
6.1.1 Conditions générales des simulations . . . 57
6.1.2 Résultats . . . 59
6.1.2.1 Estimation des log de risques relatifs . . . 59
6.1.2.2 Estimation de la capacité prédictive . . . 63
6.1.2.3 Robustesse de l'imputation multiple quand le modèle d'imputation est mal spécifié . . . 65
6.2 Données cas-cohorte simulées depuis la cohorte PRIME . . . 70
6.2.1 Description des données . . . 70
6.2.2 Mise en œuvre . . . 70
6.2.3 Résultats . . . 75
6.3 Données NWTS . . . 79
6.4 Discussion . . . 85

7 Application : Étude des Trois Cités 90
7.1 D-dimère et risque d'événement cardiovasculaire et démence vasculaire 92
7.2 Apolipoprotéines A-I (ApoA) et B-100 (ApoB) et risque d'événement cardiovasculaire . . . 98
7.3 Discussion . . . 100

8 Conclusion et perspectives 102

Annexes 111

Liste des tableaux

2.1 Tableau des pondérations sous un échantillonnage simple 10
2.2 Estimation des risques relatifs (RR), intervalles de confiance à 95% (IC 95%) et efficacité relative (ER) de l'analyse cas-cohorte par rapport à la cohorte entière. Échantillonnage simple d'une sous-cohorte de taille 1925. 22
2.3 Estimation des risques relatifs (RR), intervalles de confiance à 95% (IC 95%) et efficacité relative (ER) de l'analyse cas-cohorte par rapport à la cohorte entière. Échantillonnage simple d'une sous-cohorte de taille 700. 23

3.1 Efficacité relative en pourcentage d'un nombre fini M plutôt qu'infini d'imputations en fonction du taux d'information manquante, γ. 32

4.1 Revue non-exhaustive de 20 études cas-cohorte publiées 40

6.1 Distribution des sujets de la sous-cohorte par strate. 58
6.2 Paramètres estimés : échantillonnage simple 60
6.3 Paramètres estimés : échantillonnage stratifié, $n_{sc} = 300$ 61
6.4 Paramètres estimés : échantillonnage stratifié $n_{sc} = 1000$ 62
6.5 Estimation moyenne de la capacité prédictive (Est), écart-type moyen estimé ($\widehat{ET}$) et écart-type observé (ET). Résultats des 1000 simulations. 64
6.6 Paramètres estimés à partir de l'échantillon des non-cas. Échantillonnage stratifié. $n_{sc} = 1000$ 66
6.7 Mauvaise spécification de la distribution de la variable de phase-2 dans le modèle d'imputation 68
6.8 Capacité prédictive des modèles sans et avec la variable de phase-2. Résultats des 1000 simulations. 69
6.9 Distribution des cas par strate. 73

6.10 Estimation des risques relatifs (RR), intervalles de confiance à 95% (IC 95%) et efficacité relative (ER) de l'analyse cas-cohorte par rapport à la cohorte entière. Échantillonnage stratifié. Sous-cohortes de taille 2100. Résultats des 1000 simulations. 76
6.11 Estimation des risques relatifs (RR), intervalles de confiance à 95% (IC 95%) et efficacité relative (ER) de l'analyse cas-cohorte par rapport à la cohorte entière. Échantillonnage stratifié. Sous-cohortes de taille 700. Résultats des 1000 simulations. 77
6.12 Distribution des cas par strate. 78
6.13 Stratification et échantillonnage stratifié des données NWTS. 81
6.14 NWTS : Resultats moyens des 1000 échantillons cas-cohorte simulés. 84

7.1 Estimation des risques relatifs (RR) et intervalles de confiance à 95% (IC 95%) associés aux tertiles de D-dimère. 95
7.2 Estimation des risques relatifs (RR) et intervalles de confiance à 95% (IC 95%) associés aux variables de phase-1. 96
7.3 Capacité prédictive et intervalle de confiance à 95% (IC 95%) des tertiles de D-dimère sur le risque d'événement cardiovasculaire et sur le risque de démence vasculaire. 97
7.4 Estimation par imputation multiple des risques relatifs (RR) et intervalles de confiance à 95% (IC 95%) associés à différents marqueurs lipidiques. 99
7.5 Estimation de la capacité prédictive (C de Harrel, Δ) et NRI. 99

Table des figures

1.1 Plan d'échantillonnage d'une enquête cas-cohorte 3

3.1 Complétion séquentielle de données manquantes à structure monotone 33
3.2 Étapes principales de l'imputation multiple 35

4.1 Distribution observée de Z_2 : a) échantillon de non-cas, b) échantillon cas-cohorte. 44

6.1 Relation entre les niveaux de fibrinogène et la consommation de tabac 72
6.2 Distribution du fibrinogène par strate et statut 73
6.3 Probabilité de survie sans rechute et son intervalle de confiance à 95% en fonction de l'évaluation histologique centrale de la tumeur 79

CHAPITRE

1 Introduction

1.1 Contexte

Avec des cohortes de taille importante, le coût du recueil des covariables sur tous les sujets peut être prohibitif. Les enquêtes en deux phases, cas-cohorte ou cas-témoins nichée dans une cohorte, permettent de réduire ce coût au prix d'une perte minime d'efficacité (Langholz, 1998), car la précision des estimateurs des risques relatifs est limitée essentiellement par le nombre de cas et tous sont complètement observés dans les deux types d'enquêtes. Ces deux schémas d'étude diffèrent par le mode d'échantillonnage. Dans les études cas-témoins nichées, les témoins sont sélectionnés de façon rétrospective : on connaît les sujets qui ont présenté l'événement d'intérêt et on sélectionne des témoins appariés aux cas, ce qui implique que l'on n'étudie qu'un type d'événement. Dans les études cas-cohorte, la sous-cohorte est constituée, en général, de façon prospective : à partir de la cohorte initiale, on sélectionne une sous-cohorte par tirage au sort, simple ou stratifié, et cet échantillon peut être utilisé comme base de comparaison pour plusieurs événements différents.

On recueille d'abord, pour tous les sujets, les variables dites de phase-1, en général des variables faciles à mesurer, comme des informations socio-démographiques, cliniques, etc. La cohorte entière est suivie de manière à identifier la date de survenue du ou des événements. Puis, dans un second temps, on recueille les variables de phase-2, plus coûteuses à mesurer, comme certaines variables biologiques, sur les non-cas de la sous-cohorte ainsi que sur tous les cas. Il est évidemment nécessaire d'effectuer ce recueil dans des conditions strictement identiques pour les cas et les non-cas.

Soient C l'ensemble des sujets de la cohorte, de taille N, SC l'ensemble des sujets appartenant à la sous-cohorte, de taille n_{sc}, et E l'ensemble des sujets de la cohorte qui présentent l'événement d'intérêt pendant la période de suivi, de taille n_e. L'échantillon cas-cohorte est définit par $CC = SC \cup E$. Étant donné que la sous-cohorte est sélectionnée au départ, en général $SC \cap E \neq \emptyset$. Une illustration correspondant au schéma d'échantillonnage cas-cohorte, sans respecter toutefois les proportions de cas dans les différentes sous-populations, est donnée par la figure 1.1.

Dans les études cas-cohorte, la perte d'efficacité provient de deux sources : de l'observation incomplète, mais aussi des méthodes d'analyse classiquement utilisées, les estimateurs pondérés (Borgan *et al.*, 2000; Kalbfleisch et Lawless, 1988; Prentice, 1986; Self et Prentice, 1988). Par la suite, nous allons utiliser le terme estimateurs pondérés "classiques" au sens des plus utilisés en pratique car ils sont implémentés dans la majorité de logiciels permettant l'analyse des enquêtes cas-cohorte. Les estimateurs pondérés classiques, limitent d'une part l'efficacité en n'atteignant pas toujours l'efficacité de l'estimateur du maximum de vraisemblance, et d'autre part en ignorant l'information apportée par les non-cas en dehors de la sous-cohorte.

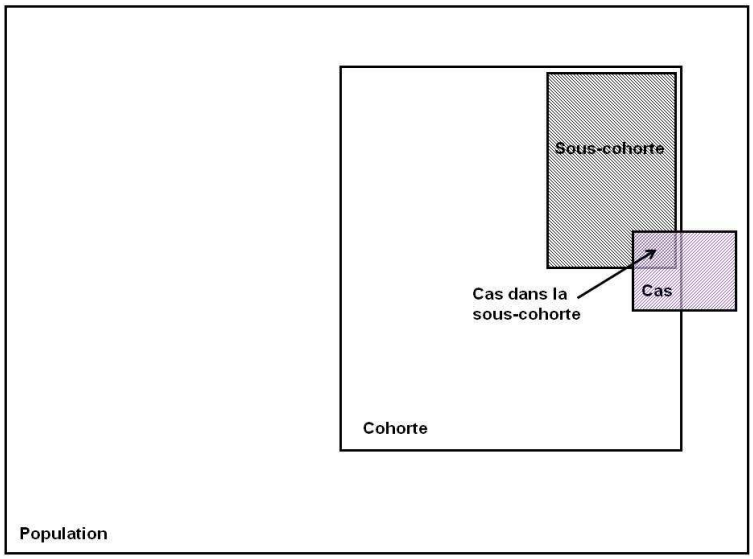


FIGURE 1.1 – Plan d'échantillonnage d'une enquête cas-cohorte

Malheureusement, l'estimateur du maximum de vraisemblance partielle est difficile à mettre en œuvre puisqu'il implique une intégration par rapport à la distribution des données manquantes.

1.2 Objectif de la thèse

Les enquêtes cas-cohorte peuvent aussi être vues comme un cas particulier de données incomplètes où le processus d'observation est contrôlé par les organisateurs de l'étude. Ainsi, des méthodes d'analyse pour données incomplètes peuvent être pertinentes (Rubin, 1987). En particulier, l'imputation multiple, qui permet d'approcher l'estimateur du maximum de vraisemblance partielle (Chavance et Manfredi, 2000; Cottrell *et al.*, 2009) en remplaçant l'intégrale sur la distribution des données manquantes par une moyenne arithmétique sur un petit nombre d'imputations tirées de cette distribution. Elle utilise toutes les données disponibles en phase-1 : un

gain en précision sur les estimations des effets des variables de phase-1, par rapport aux estimations fournies par les analyses pondérées est attendu.

Un autre intérêt de l'imputation multiple est qu'elle permet d'adapter facilement au cadre des enquêtes cas-cohorte n'importe quel outil statistique disponible pour les données de cohorte, par exemple, l'estimation de la capacité prédictive d'un modèle ou d'une variable additionnelle dont l'estimation dans les enquêtes cas-cohorte pose des problèmes spécifiques.

L'objectif de ce travail est de mettre en œuvre l'imputation multiple pour analyser les études cas-cohorte. Le chapitre 2 porte sur les méthodes d'échantillonnage des enquêtes cas-cohorte, la description des estimateurs pondérés et du calcul du nombre de sujets nécessaire, puis présente une illustration de leur mise en œuvre. Le chapitre 3 est consacré aux différentes stratégies d'analyse en présence de données incomplètes, dont l'imputation multiple. Le chapitre 4 détaille la mise en œuvre de l'imputation multiple dans les enquêtes cas-cohorte. Le chapitre 5 porte sur l'étude de la capacité prédictive d'un modèle ou d'une variable supplémentaire dans ce contexte. Le chapitre 6 inclut la validation par simulations de l'imputation multiple pour l'estimation des risques relatifs et de la capacité prédictive dans les enquêtes cas-cohorte : 1) sur des données entièrement simulées, 2) sur les données de l'étude PRospective sur l'Infarctus de MyocardE (PRIME) (Yarnell, 1998) à partir des quelles nous avons simulé des enquêtes cas-cohorte, 3) sur les données de *The National Wilms Tumor Study Group* (NWTS, http ://www.nwtsg.org) analysées par Breslow *et al.* (2009b). Le chapitre 7 est consacré à l'application dans une étude cas-cohorte issue de l'étude des 3 cités (cohorte 3C). Enfin, le chapitre 8 conclut sur le travail réalisé.

CHAPITRE

2 Enquêtes cas-cohorte et estimateurs pondérés

Nous allons d'abord décrire les deux types d'échantillonnage possibles de la sous-cohorte ; puis nous aborderons les méthodes d'analyse utilisées dans les enquêtes cas-cohorte pour estimer les risques relatifs et leur variance ; ensuite, nous présenterons les outils permettant de calculer la puissance d'un test du log-rank ou la taille de la sous-cohorte pour atteindre une puissance souhaitée ; enfin, nous illustrerons leur mise en œuvre sur un exemple.

2.1 Échantillonnage de la sous-cohorte

Initialement, Prentice (1986) avait proposé un échantillonnage simple de la sous-cohorte, ignorant toute l'information disponible en phase-1. Wacholder et Boivin (1987) avaient souligné qu'un échantillonnage stratifié de la sous-cohorte pourrait être avantageux. Ce n'est qu'une douzaine d'années plus tard, que Borgan *et al.*

(2000) ont développé cette idée. S'il existe une variable de phase-1 prédictive de la variable de phase-2, une stratification de la cohorte utilisant cette variable et un échantillonnage simple à l'intérieur de chaque strate peuvent permettre de gagner en efficacité par rapport à l'échantillonnage simple de la sous-cohorte.
En pratique, les enquêtes cas-cohorte ont rarement comme objectif d'étudier un seul événement et une seule variable de phase-2, ce qui peut rendre difficile une stratification de la cohorte de départ prenant en compte les variables prédictives de chacune des variables de phase-2. Un échantillonnage simple est alors conseillé.

2.2 Estimateurs pondérés

2.2.1 Approche pondérée classique

Les méthodes d'analyse les plus couramment utilisées ignorent l'information de phase-1 disponible sur les non-cas en dehors de la sous-cohorte. Les poids attribués aux sujets de l'ensemble cas-cohorte visent à reconstituer approximativement la cohorte entière à partir du seul échantillon cas-cohorte. Pour simplifier la présentation de cette démarche, nous utiliserons dans cette section la terminologie des études de survie.

Soit T_i le délai de survie d'un sujet i, $i = 1, \ldots, N$ et C_i le délai de censure. On observe $X_i = min(T_i, C_i)$. Soit $\Delta_i = I(T_i < C_i)$ l'indicatrice de décès. Le modèle à risques proportionnels de Cox suppose que le risque instantané de décès d'un sujet, sachant le vecteur des variables explicatives au temps t, $Z(t)$, est :

$$\lambda(t; Z(t)) = \lambda_0(t) exp\{\beta' Z(t)\} \tag{2.1}$$

où $\lambda_0(t)$ représente le risque de base pour un sujet ayant toutes les valeurs de covariables $Z(t)$ égales à zéro et où β' est la transposée du vecteur des coefficients inconnus.

En supposant qu'il n'y a pas d'ex-aequo, on dénote $t_1 < t_2 < \dots$ les temps auxquels les décès se produisent, et j l'indice du sujet pour lequel le décès se produit au temps t_j. Si l'on dispose de la cohorte entière, la fonction de vraisemblance partielle pour l'ensemble de la cohorte de taille N s'écrit :

$$VP(\beta) = \prod_{t_j} \left(\frac{exp\{\beta' Z_j(t_j)\}}{\sum_{k=1}^{N} Y_k(t_j) exp\{\beta' Z_k(t_j)\}} \right) \tag{2.2}$$

où $Y_k(t_j)$ indique si le sujet k est à risque au temps t_j.
La log-vraisemblance partielle s'écrit alors :

$$LogVP(\beta) = \sum_{t_j} \left(\beta' Z_j(t_j) - log \sum_{k=1}^{N} Y_k(t_j) exp\{\beta' Z_k(t_j)\} \right) \tag{2.3}$$

et le vecteur du score :

$$\begin{aligned} U(\beta) &= \frac{\delta LogVP(\beta)}{\delta \beta} \\ &= \sum_{t_j} \left(Z_j(t_j) - \frac{\sum_{k=1}^{N} Y_k(t_j) exp\{\beta' Z_k(t_j)\} Z_k(t_j)}{\sum_{k=1}^{N} Y_k(t_j) exp\{\beta' Z_k(t_j)\}} \right) \\ &= \sum_{t_j} (Z_j(t_j) - \bar{Z}(\beta, t_j)) \end{aligned} \tag{2.4}$$

À la date t_j, le terme $Z_j(t_j) - \bar{Z}(\beta, t_j)$ compare le vecteur des covariables du sujet j décédé à la moyenne pondérée des covariables dans la population à risque à cet instant. L'estimateur du maximum de vraisemblance partielle est celui qui maximise $VP(\beta)$ et qui est donc la solution de l'équation $U(\beta) = 0$.

Dans les enquêtes cas-cohorte la maximisation de cette expression limitée aux sujets pour lesquels elle est définie, l'échantillon cas-cohorte, fournirait de mauvais estimateurs car on rencontre deux difficultés : d'une part, un problème de biais lié à la différente probabilité d'inclusion des cas et des non-cas ; d'autre part, un problème de précision lié a la non indépendance des contributions à la vraisemblance. En effet, au dénominateur ce sont les sujets à risque de l'échantillon cas-cohorte qui interviennent au lieu des sujets à risque de la cohorte entière. Le problème de biais est résolu à l'aide d'une fonction de pseudo-vraisemblance pondérée :

$$\widetilde{V}(\beta) = \prod_{t_j} \left(\frac{exp\{\beta' Z_j(t_j)\} w_j(t_j)}{\sum\limits_{k \epsilon SC \cup E} Y_k(t_j) exp\{\beta' Z_k(t_j)\} w_k(t_j)} \right) \tag{2.5}$$

où $w_k(t_j)$ est le poids du sujet k au temps t_j. Le vecteur du score de la fonction de pseudo-vraisemblance pondérée est :

$$\begin{aligned} \widetilde{U}(\beta) &= \frac{\delta Log \widetilde{V}(\beta)}{\delta \beta} \\ &= \sum_{t_j} \left(Z_j(t_j) - \frac{\sum\limits_{k \epsilon SC \cup E} Y_k(t_j) exp\{\beta' Z_k(t_j)\} Z_k(t_j) w_k(t_j)}{\sum\limits_{k \epsilon SC \cup E} Y_k(t_j) exp\{\beta' Z_k(t_j) w_k(t_j)\}} \right) \end{aligned} \tag{2.6}$$

L'estimateur du maximum de la fonction de pseudo-vraisemblance pondérée est solution de l'équation $\widetilde{U}(\beta) = 0$.

Différentes pondérations ont été proposées au long des années. Dans le cadre d'un échantillonnage simple, Prentice (1986) avait initialement attribué un poids de 1 aux sujets appartenant à la sous-cohorte (cas et non-cas) et un poids 0 aux non-cas n'appartenant pas à l'ensemble cas-cohorte. Les cas n'appartenant pas à la

sous-cohorte avaient un poids de 0 jusqu'à l'instant précis du décès où leur poids était 1. Donc, dans cette pondération, les cas en dehors de la sous-cohorte contribuent au numérateur de la fonction de quasi-vraisemblance pondérée alors que seuls les sujets de la sous-cohorte contribuent au dénominateur. Les cas en dehors de la sous-cohorte ne sont jamais considérés comme étant à risque sauf au moment précis où le décès se produit. Dans une étude sur l'efficacité de cette approche, Self et Prentice (1988) avaient envisagé une pondération différente de celle proposée par Prentice : les cas et les non-cas n'appartenant pas à la sous-cohorte avaient toujours un poids nul. L'analyse était donc limitée au sujets de la sous-cohorte. La justification de ce choix était uniquement d'illustrer ce que l'on gagne à prendre tous les cas. Lin et Ying (1993), envisageant les enquêtes cas-cohorte comme un cas particulier de données incomplètes, avaient proposé une autre pondération : l'inverse de la probabilité d'observer complètement un sujet, soit 1 pour les cas et $(1/q)$ pour les non-cas, où q est la probabilité d'échantillonnage de la sous-cohorte. Barlow (1994) avait envisagé une pondération identique à celle proposée par Prentice pour les cas, mais un poids correspondant à l'inverse de la probabilité d'échantillonnage pour les non-cas de la sous-cohorte au lieu de 1. Therneau et Li (1999) ont proposé pondérer tous les sujets de la sous-cohorte (cas et non-cas) par l'inverse de la probabilité d'échantillonnage. Seuls les cas n'appartenant pas à la sous-cohorte avaient un poids de 1. Un résumé des pondérations proposées dans le cadre d'un échantillonnage simple est présenté au tableau 2.1.

Borgan *et al.* (2000) ont adapté les pondérations à l'échantillonnage stratifié, en proposant différents estimateurs. L'estimateur I, analogue à celui proposé par Self et Prentice (1988), tient seulement compte des sujets appartenant à la sous-cohorte mais avec des poids définis comme l'inverse de la fréquence d'observation, n_{sc_l}/N_l,

Tableau 2.1 – Tableau des pondérations sous un échantillonnage simple

	Prentice	Self et Prentice	Lin et Ying	Barlow	Therneau et Li
Cas en dehors de la SC avant le décès	0	0	0	0	0
Cas en dehors de la SC au temps du décès	1	0	1	1	1
Cas dans la SC avant le décès	1	1	$1/q$	$1/q$	$1/q$
Cas dans la SC au temps du décès	1	1	1	1	$1/q$
Non-cas dans la SC	1	1	$1/q$	$1/q$	$1/q$

SC : sous-cohorte, q : probabilité d'échantillonnage de la sous-cohorte.

avec $l = 1, \ldots, L$ où L est le nombre de strates, et n_{sc_l} et N_l le nombre de sujets, respectivement de la sous-cohorte et de la cohorte, dans la strate l. L'estimateur II attribue aux cas un poids de 1, dans l'ensemble de sujets à risque, et aux non-cas, des poids $n^0_{sc_l}/N^0_l$, définis en fonction du nombre de non-cas dans chacune des l strates dans la sous-cohorte, $n^0_{sc_l}$, et dans la cohorte, N^0_l.

Le choix des probabilités de tirage optimales reste encore une question délicate. Cependant, une sélection de la sous-cohorte de façon à obtenir des effectifs approximativement proportionnels dans les différentes strates chez les cas et chez les non-cas est conseillée.

Barlow (1994) a montré, pour un échantillonnage simple, qu'il pouvait être préférable de pondérer les observations en fonction des fréquences observées, n_{sc}/N, plutôt que des probabilités d'échantillonnage q. Borgan *et al.* (2000) ont également montré qu'il est théoriquement préférable d'utiliser des poids variables en fonction du temps, en actualisant les fréquences d'observation pour les sujets encore à risque à chaque temps de décès. Chen (2001), pour un échantillonnage simple, et Samuel-

sen *et al.* (2007) pour un échantillonnage stratifié, ont proposé une mise à jour des poids moins fréquente. En pratique, le gain est cependant négligeable pour les événements rares étudiés généralement dans les enquêtes cas-cohorte. Des études par simulation ont montré que pour une grande taille de sous-cohorte, environ 15% de la cohorte, les différentes pondérations fournissent des estimations similaires (Onland-Moret *et al.*, 2007).

2.2.2 Estimation de la variance de l'estimateur pondéré

Dans les enquêtes cas-cohorte, l'utilisation de la formule naïve d'estimation de la variance surestime la précision du paramètre d'intérêt, car on néglige le caractère partiel de l'observation des données : à la composante correspondant à la cohorte entière il faut ajouter une composante due au tirage au sort de la sous-cohorte, qui reflète l'incertitude additionnelle provenant de l'observation partielle des informations en phase-2. Self et Prentice (1988) ont proposé un estimateur qui tient compte de cette variabilité additionnelle.

$$\widehat{Var}(\hat{\beta}) = \widehat{I}^{-1} + \frac{n_{sc}(N - n_{sc})}{N} CovD_C \tag{2.7}$$

où I est la matrice d'information de Fisher pour la cohorte entière et $CovD_C$ est la matrice des covariances empiriques de la fonction d'influence pour les sujets appartenant à la sous-cohorte, estimées à partir des résidus *dfbeta* (Therneau et Grambsch, 2000)

$$dfbeta_{ji} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{s_{j(i)}\sqrt{(Z'Z)^{-1}_{jj}}} \tag{2.8}$$

où $\hat{\beta}_{j(i)}$ est l'estimation du paramètre β_j et $s^2_{j(i)}$ sa variance obtenues après suppression de la $i^{ème}$ observation.

Cette approche avait été critiquée à l'époque pour la complexité de sa mise en œuvre (Lin et Ying, 1993). Ultérieurement, Lin et Ying (1993) et Barlow (1994) ont fourni une estimation robuste de la variance, utilisant le jackknife. Ces différents estimateurs de la variance présentent des performances similaires, comme l'ont montré Therneau et Li (1999).

2.2.3 Prise en compte de l'ensemble des données de phase-1

Dans l'intention d'augmenter la précision des estimateurs pondérés classiques, Breslow *et al.* (2009b) ont minimisé la composante de la variance liée au tirage au sort de la sous-cohorte. Ils se sont appuyés sur la méthode de calibration proposée par Deville et Sarndal (1992). L'idée était d'utiliser des informations disponibles sur tous les sujets pour mieux refléter la cohorte de départ.

Pour estimer une somme ou une moyenne, disons la valeur produite dans un secteur économique donné, on sait qu'un échantillonnage stratifié sur-représentant les entreprises de taille importante permet d'augmenter la précision. Si un échantillonnage simple a été utilisé mais que la taille de toutes les entreprises est connue, il reste possible d'améliorer la précision en post-stratifiant sur cette taille : on stratifie les entreprises en fonction de leur taille, puis on choisit des poids tels que, dans chaque strate, la somme pondérée des tailles d'entreprises observées corresponde exactement à la somme totale.

Dans les enquêtes cas-cohorte, on ne s'intéresse pas à une somme mais à des coefficients de régression, or améliorer la précision sur la somme de la variable de phase-2 n'améliore pas la précision de l'estimateur du coefficient associé. Breslow *et al.* (2009a) ont utilisé le résultat que la somme des résidus sur la cohorte entière est nulle et ils ont imposé la contrainte que la somme pondérée des résidus sur l'échantillon cas-cohorte soit nulle. Comme il existe une infinité de pondérations respectant cette condition, on a choisit celle qui minimise la distance entre les nouveaux poids calibrés w_k et les poids de départ $d_k = 1/q_k$, où q_k est la probabilité d'observer complètement le sujet k. Une distance couramment utilisée est la distance $\sum_k (w_k - d_k)^2/2d_k$ (Breslow *et al.*, 2009a), connu comme l'estimateur GREG (Särndal *et al.*, 1989). En pratique, on construit un modèle linéaire généralisé pour prédire la variable de phase-2, en fonction des prédicteurs qui lui sont liés. Les valeurs attendues selon ce modèle de prédiction sont utilisées pour tous les sujets, qu'ils appartiennent ou non a l'échantillon cas-cohorte, et on ajuste un modèle de Cox aux données ainsi complétées. Les résidus *dfbeta* (Belsley *et al.*, 1980) des variables incluses dans ce modèle sont alors calculés. La calibration des poids peut alors être effectuée sous la contrainte que les sommes pondérées des résidus *dfbeta* associés à certaines ou à toutes les variables du modèle ajusté aux données de l'ensemble cas-cohorte soient nulles. On attend un gain en précision pour les coefficients sur lesquels les résidus *dfbeta* ont été calibrés. Cependant, l'approche par calibration des poids est susceptible de rencontrer des problèmes numériques quand le nombre de variables utilisées dans le modèle de prédiction est élevé : l'algorithme de calibration risque de ne pas converger et des estimations biaisées peuvent être obtenues. Breslow *et al.* (2009a) ont proposé d'utiliser alors l'approche par re-estimation, dont les poids correspondent à l'inverse de la probabilité d'inclusion prédite par un modèle logistique ajusté sur la cohorte entière.

2.3 Calcul du nombre de sujets nécessaire dans les études cas-cohorte

Cai et Zeng (2004) ont montré comment adapter aux études cas-cohorte les calculs de la puissance et du nombre de sujets nécessaire quand on utilise un test du log-rank pour comparer deux groupes. Ils se sont placés dans le cadre d'un échantillonnage simple de la sous-cohorte, un facteur de risque binaire et une incidence faible de la maladie. Les calculs font intervenir : la taille de la cohorte (N), le log du risque relatif jugé épidémiologiquement intéressant (θ), la puissance (1-β) et le risque de première espèce (α) souhaités, le taux attendu d'événements dans la cohorte (p_D), et la fréquence de l'exposition $p_1 = P(Z = 1)$, où Z est le facteur de risque binaire.

Cas bilatéral

Les hypothèses du test bilatéral d'une absence d'effet, fondé sur une généralisation du test du log-rank, sont :

$H_0 : \Lambda_1(t) = \Lambda_2(t)$

$H_1 : \Lambda_1(t) = \Lambda_2(t)e^{\theta}$, avec $\theta \neq 0$

où Λ_i est la fonction de risque cumulé pour le groupe i, $i = 1, 2$.

Sous les hypothèses suivantes : (a) la distribution des délais de censure est la même dans les deux groupes, (b) l'incidence totale dans la cohorte est très petite mais le nombre d'événements attendus est largement supérieur à 1, (c) la distribution des délais de survie observés est supposée continue (absence d'ex-aequo), la puissance du test du log-rank comparant deux groupes de sujets, avec un risque de première espèce égal à α, une sous-cohorte de taille n_{sc} et un risque relatif $exp(\theta)$ est donnée

par :

$$\Phi\left(z_{\alpha/2}+\theta\sqrt{n_{sc}\frac{p_1(1-p_1)p_D}{q+(1-q)p_D}}\right) \tag{2.9}$$

où z_c correspond au quantile c de la distribution normale standard et q est la fraction d'échantillonnage de la sous-cohorte.

La taille de la sous-cohorte nécessaire pour détecter un risque relatif $exp(\theta)$ avec une puissance égale à 1-β et un risque de première espèce égal à α est donnée par :

$$n_{sc}=\frac{NBp_D}{N-B(1-p_D)} \tag{2.10}$$

sous la contrainte $N > B(1-p_D)$ et avec

$$B=\frac{(z_{1-\alpha/2}+z_{1-\beta})^2}{\theta^2 p_1(1-p_1)p_D}. \tag{2.11}$$

Cas unilatéral

Supposons que l'on attende une augmentation du risque dans le groupe 2. Les hypothèse du test unilatéral d'une absence d'effet, fondé sur une généralisation du test du log-rank, sont :

$H_0 : \Lambda_1(t) = \Lambda_2(t)$

$H_1 : \Lambda_1(t) = \Lambda_2(t)e^{\theta}$, avec $\theta > 0$

où Λ_i est la fonction de risque cumulé pour le groupe i, $i = 1, 2$.

Sous les mêmes hypothèses que pour le cas bilatéral, la puissance du test du log-rank comparant deux groupes de sujets, avec un risque de première espèce égal à

α, une sous-cohorte de taille n_{sc} et un risque relatif $exp(\theta)$ est donnée par :

$$\Phi\left(z_\alpha + \theta\sqrt{n_{sc}\frac{p_1(1-p_1)p_D}{q+(1-q)p_D}}\right) \tag{2.12}$$

La taille de la sous-cohorte nécessaire pour détecter un risque relatif $exp(\theta)$ avec une puissance égale à 1-β et un risque de première espèce égal à α est donnée par l'équation 2.10 sous la contrainte $N > B(1-p_D)$, avec

$$B = \frac{(z_{1-\alpha} + z_{1-\beta})^2}{\theta^2 p_1(1-p_1)p_D}. \tag{2.13}$$

Sous l'hypothèse de maladie rare, $\frac{p_D}{1-p_D} \approx p_D$, le nombre de sujets nécessaire, pour le cas bilatéral comme unilatéral, est donné par le nombre de cas attendus multiplié par un facteur ($B/N - B$)

$$n_{sc} \approx N p_D \frac{B}{N-B} \tag{2.14}$$

Une autre formule adaptant aux études cas-cohorte le calcul du nombre de sujets nécessaire dans une cohorte entière a été proposée par Kubota et Wakana (2011). Par ailleurs, Cai et Zeng (2007) ont développé le calcul de la puissance et du nombre de sujets nécessaire pour une étude cas-cohorte dans le cadre d'événements non rares.

En outre, Self et Prentice (1988) ont calculé l'efficacité relative asymptotique (ERA), comparant la précision fournie par une enquête cas-cohorte à celle qu'aurait procurée la cohorte entière. Elle dépend du taux d'échantillonnage de la sous-

cohorte, n_{sc}/N, et du taux d'événements, p_D.

$$\text{ERA} \approx \left\{1 + 2\frac{1 - n_{sc}/N}{n_{sc}/N}\left[1 + \frac{1 - p_D}{p_D}\log(1 - p_D)\right]\right\}^{-1} \tag{2.15}$$

2.4 Mise en œuvre

Des outils pour le calcul du nombre de sujets nécessaire, l'échantillonnage et l'analyse des enquêtes cas-cohorte sont disponibles sous différents logiciels. Avec R, la fonction *ccsize* de la bibliothèque *gap* effectue le calcul de la taille de la sous-cohorte ; dans la bibliothèque *survey*, la fonction *twophase* permet de sélectionner la sous-cohorte et *svycoxph* fournit plusieurs estimations pondérées, dont celles de Breslow *et al.* (2009b) ainsi qu'une estimation correcte de leur variances ; la bibliothèque *survival* dispose de la fonction *cch*, qui permet d'analyser des données issues d'une enquête cas-cohorte, mais sans prendre en compte l'ensemble des données de phase-1. Avec SAS, les procédures à utiliser sont : *proc surveyselect* pour sélectionner la sous-cohorte et *proc phreg* pour l'estimation pondérée des paramètres ; dans l'instruction d'appel à la procédure *phreg*, l'option *covsandwich(aggregate)* permet d'obtenir une estimation de la variance robuste des estimations, car par défaut, la variance n'est pas corrigée. Avec Stata, il faut utiliser les procédures *stcascoh* pour sélectionner la sous-cohorte et *stselpre* pour estimer les paramètres et leur variance.

À l'aide du logiciel R, et à partir de la cohorte PRIME (Yarnell, 1998), nous avons, d'une part, calculé la taille de sous-cohorte nécessaire pour atteindre une puissance donnée. D'autre part, nous avons fixé une fraction d'échantillonnage et nous avons calculé la puissance attendue. Puis nous avons simulé et analysé les deux études cas-cohorte correspondantes.

L'étude PRIME (*Prospective Epidemiological Study of Myocardial Infarction*) est une étude de cohorte multicentrique. La cohorte était constituée de N=9510 sujets, dont 642 (p_D = 0,068) ont présenté un événement cardiovasculaire pendant les 10 ans de suivi. Le fibrinogène est une protéine dont la concentration plasmatique augmente dans les états inflammatoires. Elle joue également un rôle important dans la formation de caillots. Les niveaux de fibrinogène élevés sont associés à une augmentation du risque d'événement cardiovasculaire (Scarabin *et al.*, 2003). Nous avons supposé que le fibrinogène était une variable de phase-2, donc non disponible pour les non-cas en dehors de la sous-cohorte et nous avons choisi d'estimer l'effet du fibrinogène sur la survenue d'un événement cardiovasculaire, après ajustement sur le centre de recrutement et les caractéristiques du sujet à l'entrée dans l'étude : âge (années), cholestérol total (g/l), cholestérol HDL (g/l), pression artérielle systolique (mmHg) et consommation de tabac (g/j). Une concentration plasmatique de fibrinogène inférieure à 4g/l est généralement jugée normale (Mackie *et al.*, 2003), nous avons dichotomisé la concentration plasmatique de fibrinogène en utilisant ce seuil. Ainsi, la fréquence d'exposition à une concentration élevée était p_1 =0,16. Nous avons fixé le risque de première espèce à 5%. Deux sous-cohortes ont été sélectionnées et analysées. D'une part, nous avons calculé le nombre de sujets nécessaire afin d'atteindre une puissance de 70%. D'autre part, nous avons fixé la probabilité d'échantillonnage $q = 0,07$ et nous avons calculé la puissance attendue.

Supposons que nous souhaitions mettre en évidence un risque relatif d'au moins 1,35, soit $\theta = 0,3$. Le nombre de sujets nécessaire dans la sous-cohorte pour détecter un risque relatif de $exp(\theta) = 1,35$ ou $exp(-\theta) = 0,74$ avec une puissance égale à

70% et un risque de première espèce égal à 5% pour un test bilatéral était :

$$n_{sc} = 9510 * 0,068 * \frac{7500}{9510 - 7500 * (1 - 0,068)} \approx 1925 \tag{2.16}$$

car

$$B = \frac{(1,96 + 0,52)^2}{0,3^2 * 0,16 * (1 - 0,16) * 0,068} \approx 7500 \tag{2.17}$$

L'échantillon cas-cohorte devait être de l'ordre de n_{sc}+($N - n_{sc}$)$p_D \approx$ 2440, sachant que certains sujets appartenant à la sous-cohorte deviendront des cas. Dans ce cadre, ERA = 0,79 et la perte d'efficacité aurait été de 21% bien que pour 75% des sujets le fibrinogène n'aurait pas été mesuré.

Puisque nous ne disposons pas d'une variable de phase-1 fortement prédictive du fibrinogène, nous avons, d'abord, sélectionné par échantillonnage simple 1925 sujets de la cohorte. L'échantillon cas-cohorte était constitué de 2444 sujets au total : 1925 sujets de la sous-cohorte (1803 non-cas et 122 cas) plus les 520 cas en dehors de la sous-cohorte. Puis, nous avons fixé la probabilité d'échantillonnage à $0,07$ afin d'obtenir une sous-cohorte d'approximativement 700 sujets. La puissance correspondant à un risque α égal à 5%, était donnée par :

$$\Phi\left(-1,96 + 0,3 * \sqrt{700 * \frac{0,16 * (1 - 0,16) * 0,068}{0,07 + (1 - 0,07) * 0,068}}\right) = \Phi(0,12) \approx 0,55 \tag{2.18}$$

Pour un test unilatéral, le nombre de sujets nécessaire dans la sous-cohorte pour détecter un risque relatif de $exp(\theta) = 1,35$ avec une puissance égale à 70% et un risque de première espèce égal à 5% aurait été

$$n_{sc} = 9510 * 0,068 * \frac{5672}{9510 - 5672 * (1 - 0,068)} \approx 869 \tag{2.19}$$

car

$$B = \frac{(1,64 + 0,52)^2}{0,3^2 * 0,16 * (1 - 0,16) * 0,068} \approx 5672 \tag{2.20}$$

Et la puissance pour une sous-cohorte d'approximativement 700 sujets aurait été de

$$\Phi\left(-1,64 + 0,3 * \sqrt{700 * \frac{0,16 * (1 - 0,16) * 0,068}{0,07 + (1 - 0,07) * 0,068}}\right) = \Phi(0,43) \approx 0,67 \tag{2.21}$$

Nous avons estimé les risques relatifs sur l'ensemble de la cohorte PRIME. Puis, pour analyser les enquêtes cas-cohorte simulées, nous avons utilisé la pondération proposée par Lin et Ying (1993) et la calibration proposée par Breslow *et al.* (2009b) en calibrant, soit sur les seuls résidus *dfbeta* du seul fibrinogène, soit sur les résidus de toutes les variables incluses dans le modèle d'analyse.

Globalement, les estimations des risques relatifs fournies par les analyses pondérées classique et calibrée étaient proches de celles fournies par l'analyse de la cohorte entière, pour la sous-cohorte de taille 1925, $q \approx 0,2$ (Tableau 2.2), comme pour la sous-cohorte de taille 700, $q = 0,07$ (Tableau 2.3). Concernant la précision des estimations, pour les variables de phase-1, c'est-à-dire, complètement observées, l'analyse pondérée classique fournissait, comme attendu, des estimations moins précises que l'analyse de la cohorte entière. L'analyse pondérée calibrée fournissait des estimations plus précises que l'analyse pondérée classique pour les variables dont

les résidus avaient été utilisés pour la calibration. L'analyse calibrée sur les résidus *dfbeta* de toutes les variables d'ajustement, fournissait des estimations plus précises que l'analyse pondérée classique pour l'ensemble de variables sans atteindre la précision des estimations fournies par l'analyse de la cohorte entière. L'efficacité relative, mesurée comme le rapport des variances fournies par la cohorte complète sur celles des analyses pondérées pour une petite sous-cohorte ($q = 0,07$) était faible, entre 0,33 et 0,47 pour les estimateurs pondérés classiques, et entre 0,39 et 0,48 pour les estimateurs pondérés calibrés (Tableau 2.3). Pour le fibrinogène, variable de phase-2, et une grande sous-cohorte, l'efficacité relative était de 0,69 pour l'estimation fournie par l'analyse pondérée classique contre 0,73 et 0,77 pour les estimations calibrées (Tableau 2.2). Comme attendu, pour une petite sous-cohorte, cette efficacité relative était sensiblement plus faible que pour une grande sous-cohorte, 0,41 et 0,42, pour les estimations de l'analyse pondérée classique et calibrée respectivement (Tableau 2.3). Notons que ces résultats ne correspondent qu'au tirage d'une seule sous-cohorte de chaque taille et ne permettent pas d'apprécier la distribution des efficacités relatives.

Tableau 2.2 – Estimation des risques relatifs (RR), intervalles de confiance à 95% (IC 95%) et efficacité relative (ER) de l'analyse cas-cohorte par rapport à la cohorte entière. Échantillonnage simple d'une sous-cohorte de taille 1925.

	Cohorte		Cas-cohorte								
	PRIME		Lin-Ying			Cal1[a]			Cal2[b]		
	RR	IC 95%	RR	IC 95%	ER	RR	IC 95%	ER	RR	IC 95%	ER
Fibrinogène	1,14	(1,06 ; 1,22)	1,12	(1,03 ; 1,22)	0,69	1,12	(1,04 ; 1,22)	0,73	1,12	(1,04 ; 1,21)	0,77
Tabac[d]	1,24	(1,16 ; 1,34)	1,23	(1,13 ; 1,35)	0,61	1,24	(1,14 ; 1,34)	0,77	1,24	(1,14 ; 1,34)	0,81
Âge	1,06	(1,03 ; 1,09)	1,06	(1,03 ; 1,10)	0,68	1,06	(1,03 ; 1,10)	0,68	1,06	(1,03 ; 1,10)	0,77
Centre	1,32	(1,12 ; 1,56)	1,26	(1,03 ; 1,54)	0,70	1,27	(1,03 ; 1,55)	0,70	1,27	(1,04 ; 1,54)	0,74
Cholestérol total[d]	1,08	(1,06 ; 1,10)	1,09	(1,06 ; 1,11)	0,69	1,09	(1,06 ; 1,11)	0,69	1,08	(1,06 ; 1,11)	0,83
Cholestérol HDL[d]	0,59	(0,52 ; 0,68)	0,60	(0,52 ; 0,70)	0,81	0,60	(0,51 ; 0,71)	0,83	0,60	(0,52 ; 0,70)	0,85
PAS[d,e]	1,11	(1,08 ; 1,14)	1,10	(1,06 ; 1,14)	0,64	1,10	(1,06 ; 1,14)	0,71	1,10	(1,06 ; 1,14)	0,79

[a] Cal1 : Poids calibrés sur les résidus *dfbeta* de la variable fibrinogène

[b] Cal2 : Poids calibrés sur les résidus *dfbeta* de toutes les variables d'ajustement

[d] coefficients multipliés par 10

[e] PAS, pression artérielle systolique

Tableau 2.3 – Estimation des risques relatifs (RR), intervalles de confiance à 95% (IC 95%) et efficacité relative (ER) de l'analyse cas-cohorte par rapport à la cohorte entière. Échantillonnage simple d'une sous-cohorte de taille 700.

	Cohorte		Cas-cohorte								
	PRIME		Lin-Ying			Cal1[a]			Cal2[b]		
	RR	IC 95%	RR	IC 95%	ER	RR	IC 95%	ER	RR	IC 95%	ER
Fibrinogène	1,14	(1,06 ; 1,22)	1,12	(1,01 ; 1,25)	0,41	1,12	(1,01 ; 1,25)	0,42	1,12	(1,01 ; 1,25)	0,41
Tabac[d]	1,24	(1,16 ; 1,34)	1,27	(1,12 ; 1,43)	0,33	1,27	(1,12 ; 1,43)	0,38	1,27	(1,12 ; 1,42)	0,39
Âge	1,06	(1,03 ; 1,09)	1,07	(1,02 ; 1,11)	0,41	1,07	(1,02 ; 1,11)	0,41	1,07	(1,02 ; 1,11)	0,45
Centre	1,32	(1,12 ; 1,56)	1,24	(0,95 ; 1,61)	0,40	1,24	(0,95 ; 1,61)	0,40	1,24	(0,96 ; 1,61)	0,43
Cholestérol total[d]	1,08	(1,06 ; 1,10)	1,09	(1,06 ; 1,12)	0,42	1,09	(1,06 ; 1,12)	0,42	1,09	(1,06 ; 1,12)	0,48
Cholestérol HDL[d]	0,59	(0,52 ; 0,68)	0,60	(0,49 ; 0,73)	0,47	0,60	(0,49 ; 0,63)	0,47	0,60	(0,50 ; 0,72)	0,55
PAS[d,e]	1,11	(1,08 ; 1,14)	1,09	(1,03 ; 1,15)	0,36	1,09	(1,03 ; 1,14)	0,39	1,09	(1,03 ; 1,14)	0,39

[a]Cal1 : Poids calibrés sur les résidus *dfbeta* de la variable fibrinogène

[b]Cal2 : Poids calibrés sur les résidus *dfbeta* de toutes les variables d'ajustement

[d] coefficients multipliés par 10

[e]PAS, pression artérielle systolique

CHAPITRE

3 Données incomplètes et imputation multiple

3.1 Typologie des données incomplètes

Les données incomplètes, très fréquentes dans la plupart des études quel que soit le domaine d'application, posent des problèmes de biais et de précision. L'ampleur de ces problèmes dépend d'abord du processus de génération des données manquantes mais il dépend aussi des méthodes d'analyse mises en œuvre. Cependant, si la perte de précision est inévitable, le biais dépend du processus d'observation et du modèle d'analyse.

Little et Rubin (1987) ont proposé de distinguer trois type de données manquantes : 1) données *manquant complètement aléatoirement* (MCA) lorsque la probabilité qu'une observation soit incomplète est une constante. Ce processus entraîne une perte de précision mais aucun biais ; 2) données *manquant aléatoirement* (MA) lorsque cette probabilité ne dépend que de valeurs observées. Ce processus entraîne

une perte de précision mais aucun biais avec des méthodes statistiques appropriées ; 3) données *manquant non aléatoirement* (MNA) lorsque cette probabilité dépend de valeurs non observées. Ce processus de données manquantes entraîne une perte de précision et le biais de sélection éventuel ne peut pas être redressé. Il est donc important d'effectuer la distinction entre données MA ou MNA car avec des données MA et une méthode d'analyse pertinente, il est possible d'effectuer des inférences correctes sur les paramètres d'intérêt, en prenant en compte l'ensemble des données observées, tandis qu'avec des données MNA il faut modéliser simultanément le processus de réponse et le processus d'observation.

Supposons que Y soit la variable réponse complètement observée, X un vecteur incomplètement observée, $X = (X^{obs}, X^{mis})$ où X^{obs} est la partie observée et X^{mis} est la partie non observée , et R une variable indicatrice permettant d'identifier les observations complètes et incomplètes :

$$R = \begin{cases} 1 & \text{observation complète} \\ 0 & \text{observation incomplète} \end{cases}$$

La vraisemblance jointe s'obtient à partir de la densité conjointe de Y et de R, conditionnellement à X, qui peut s'écrire :

$$\begin{aligned} V(\theta, \varphi | Y, R, X^{obs}, X^{mis}) &= f_{\theta\varphi}(Y, R | X^{obs}, X^{mis}) \\ &= f_{\theta}(Y | X^{obs}, X^{mis}) P_{\varphi}[R | Y, X^{obs}, X^{mis}] \end{aligned} \tag{3.1}$$

où f_θ et P_φ représentent respectivement la densité marginale (par rapport à R) de Y et la probabilité conditionnelle de R sachant Y qui dépendent respectivement de paramètres θ et φ supposés distincts.

La définition des données MCA ou MA implique $f_\theta(Y|R,X) = f_\theta(Y|X^{obs})$, donc, sous l'hypothèse de données MCA ou MA, la vraisemblance observée de θ et φ est :

$$\begin{aligned} V(\theta,\varphi|Y,R,X^{obs}) &= \int f_{\theta\varphi}(Y,R|X^{obs},X^{mis})dX^{mis} \\ &= \int f_\theta(Y|X^{obs},X^{mis})P_\varphi[R|Y,X^{obs}]dX^{mis} \\ &= P_\varphi[R|Y,X^{obs}]\int f_\theta(Y|X^{obs},X^{mis})dX^{mis} \\ &= P_\varphi[R|Y,X^{obs}]f_\theta(Y|X^{obs}) \end{aligned} \tag{3.2}$$

Si l'on suppose que θ et φ sont distincts, seul le second facteur contribue à la vraisemblance de θ et l'estimateur du maximum de vraisemblance obtenu à partir des observations disponibles est sans biais. C'est en ce sens que le processus d'observation MCA ou MA sont dits ignorables (Little et Rubin, 1987). Cet estimateur du maximum de vraisemblance n'est cependant pas toujours facile à obtenir car il implique des intégrales multiples.

3.2 Stratégies d'analyse

En présence de données incomplètes, différentes stratégies d'analyse sont possibles (Chavance et Manfredi, 2000).

3.2.1 Modélisation des cas complets

La modélisation des cas complets consiste à supposer que les observations sont MCA et à limiter l'analyse au sous-échantillon de sujets pour lesquels les observations sont complètes. Si les données sont MCA on obtient des estimations non biaisées, mais on s'expose souvent à une diminution considérable de la précision

et de la puissance. Dans une analyse multivariée, plusieurs variables peuvent être observées incomplètement. Même si pour chaque variable il n'y a que peu de données manquantes, la proportion d'observations complètes risque d'être faible. En outre, l'hypothèse de données MCA est généralement peu réaliste, il est préférable de prendre en compte le risque d'un biais de sélection.

3.2.2 Indicatrice de données manquantes

Cette méthode consiste à associer à chaque variable explicative incomplètement observée, X_j a) une variable indicatrice R_j permettant d'identifier les observations complètes et incomplètes :

$$R_j = \begin{cases} 1 & \text{observation complète} \\ 0 & \text{observation incomplète} \end{cases}$$

et b) une variable X_j^*

$$X_j^* = \begin{cases} X_j & \text{observation complète} \\ k & \text{observation incomplète} \end{cases}$$

ou k est une constante arbitraire. Il est alors possible d'analyser la totalité de l'échantillon observé en remplaçant chaque variable incomplète X_j par le couple (X_j^*, R_j). Pour une variable catégorielle, cela revient à ajouter une modalité à la variable catégorielle incomplètement observée. Par apport à l'analyse des données complètes, cette méthode permet d'améliorer la précision de certains estimateurs puisque les informations disponibles sur les sujets incomplètement observés sont intégralement prises en compte. Elle permet aussi d'apprécier le risque de biais. Supposons que X_1 soit complètement observée et X_2 incomplètement observée. Tant l'effet simple de

l'indicatrice de donnée manquante, R_2, qu'une interaction dans le prédicteur linéaire η de la réponse Y entre indicatrice de donnée manquante et variable explicative X_1 signale l'existence d'un biais de sélection dans l'analyse des cas complets.

$$\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2^* + \beta_3 R_2 + \beta_4 R_2 X_1 \tag{3.3}$$

La dernière signifie que l'effet de la variable X_1 n'est pas le même chez les sujets complètement et incomplètement observés. Mais cette stratégie ne protège qu'imparfaitement contre le risque de biais (Greenland et Finkle, 1995; Vach et Blettner, 1991), car l'estimation de l'effet de X_1 sur la totalité de l'échantillon tout en éliminant le biais de sélection est maintenant une moyenne pondérée entre un effet ajusté sur X_2, estimé sur les sujets complets, et un effet non ajusté sur X_2, donc affecté d'un biais de confusion, estimé sur les sujets incomplets.

3.2.3 Pondération par l'inverse de la probabilité

La pondération par l'inverse de la probabilité d'observation ou IPW (*Inverse Probability Weighting*) est fondée sur le même principe que l'estimateur de Horvitz-Thompson. Elle vise à reconstituer approximativement la cohorte entière à partir des données complètement observées en attribuant à chaque sujet complet un poids égal à l'inverse de sa probabilité d'être observé. S'il n'existe qu'un petit nombre de profils incomplets, les probabilités sont estimées par les fréquences observées. Dans les situation plus complexes, cette probabilité peut être estimée par un modèle logistique (Robins *et al.*, 1994, 1995).

3.2.4 Imputation simple

Cette méthode consiste à remplacer chaque valeur manquante par une prédiction obtenue à partir des autres observations. L'imputation requiert la modélisation correcte des relations entre les variables incomplètes et les variables qui leur sont liées, ce qui suppose des données MA. Dans le modèle d'imputation la variable réponse est la variable incomplètement observée. Les variables explicatives sont les variables qui lui sont liées et qui vont permettre de refléter sans biais de confusion la relation entre la variable incomplète et la variable réponse dans le modèle d'analyse. Cette dernière doit donc figurer impérativement dans le modèle d'imputation, ainsi que les variables de confusion potentielles, que sont les autres variables explicatives du modèle d'analyse. Omettre la variable réponse reviendrait à imputer les données manquantes sous l'hypothèse nulle d'absence de relation entre celles-ci et celle-là. Notons que si le modèle d'analyse inclut des termes d'interaction entre la variable à imputer et d'autres variables, le modèle d'imputation doit inclure des termes d'interaction entre ces autres variables et la variable réponse du modèle d'analyse. En outre, si l'on dispose d'une ou plusieurs variables prédictives de la variable à imputer qui ne figurent pas dans le modèle d'analyse, il convient évidemment de les prendre en compte. Pour chaque enregistrement où une variable est incomplète, on impute la valeur attendue sous le modèle considéré.

Dans la mesure où le processus d'observation est MA et où l'on dispose d'un modèle d'imputation correct, on conçoit que cette méthode résolve les problèmes de biais. En revanche, de nouvelles difficultés apparaissent en ce qui concerne la précision. Ignorer l'erreur résiduelle dans le modèle d'imputation implique que l'on impute systématiquement la même valeur à tous les sujets présentant le même profil de variables explicatives ; ignorer l'incertitude sur les paramètres du modèle de

prédiction signifie que l'on impute une valeur excessivement ressemblante aux observations complètes. Une conséquence globale est la sous-estimation de la variance des effets fixes.

3.2.5 Imputation multiple

L'imputation multiple, développée par Little et Rubin (1987), permet de résoudre le problème de précision issu de l'utilisation de l'imputation simple. Elle fournit une approximation de l'estimateur du maximum de vraisemblance, en remplaçant l'intégrale sur la distribution des données manquantes par une moyenne arithmétique sur un petit nombre d'imputations tirées de cette distribution, et permet donc de corriger le biais de sélection, s'il existe. Cette méthode consiste à générer plusieurs jeux plausibles de données complétées ($M \geq 2$) en prenant en compte tous les niveaux d'incertitude concernant les valeurs manquantes. Les données manquantes ne sont pas remplacées par leur espérance selon le modèle d'imputation mais par une valeur tirée dans la loi fournie par ce modèle. Pour tenir compte de l'incertitude sur les paramètres du modèle d'imputation on effectue plusieurs imputations avec des paramètres tirés dans la loi asymptotique de leur estimateur.

On obtient un estimateur du paramètre d'intérêt, $\hat{\theta}_m$, $m = \{1, \ldots, M\}$ et une estimation de la variance de cet estimateur $\widehat{Var}(\hat{\theta}_m)$ pour chaque jeu de données complété. Si le modèle d'imputation est correct, les estimations $\hat{\theta}_m$ sont asymptotiquement non-biaisées. L'estimateur de l'imputation multiple est la moyenne arithmétique de ces M estimations, qui est également asymptotiquement sans biais :

$$\hat{\theta}_{IM} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m \tag{3.4}$$

Grâce à la multiplicité des imputations, on peut estimer correctement la variance de cet estimateur unique. Cette variance est formée par deux composantes : la composante *intra-imputations* (W_{IM}), estimée par la moyenne de la variance asymptotique sur les imputations, $\widehat{W}_{IM}$, et la composante *inter-imputations* (B_{IM}), estimée à partir de la variance observée des estimations dans les M imputations, $\widehat{B}_{IM}$:

$$\begin{aligned} \widehat{Var}(\hat{\theta}_{IM}) &= \widehat{W}_{IM} + \widehat{B}_{IM} \\ &= \frac{1}{M}\sum_{m=1}^{M}\widehat{Var}(\hat{\theta}_m) + (1+M^{-1})\frac{\sum_{m=1}^{M}(\hat{\theta}_m - \hat{\theta}_{IM})(\hat{\theta}_m - \hat{\theta}_{IM})'}{M-1} \end{aligned} \tag{3.5}$$

où $1 + M^{-1}$ est un facteur d'ajustement au fait d'utiliser un nombre fini d'imputations (Rubin et Schenker, 1991). Remarquer l'analogie avec (2.7) : dans les deux formules on ajoute à la variance attendue avec les données complètes un terme exprimant l'incertitude introduite par les données manquantes.

Rubin (1987) a montré que l'incrément relatif de la variance dû à la non-réponse peut être estime par :

$$r = \frac{(1+M^{-1})\widehat{B}_{IM}}{\widehat{W}_{IM}} \tag{3.6}$$

et que l'efficacité relative de l'imputation multiple en utilisant un nombre fini d'imputations M par rapport à un nombre infini est :

$$\text{ERA}_M \approx \left(1 + \frac{\gamma}{M}\right)^{-1/2} \tag{3.7}$$

où γ est la fraction d'information manquante :

$$\gamma \approx \frac{B_{IM}}{B_{IM} + W_{IM}} \tag{3.8}$$

Il faut noter que la fraction d'information manquante ne correspond pas au pourcentage de données manquantes. Il s'agit de la fraction d'information manquante sur le paramètre d'intérêt au sens de Fisher, qui quantifie l'information relative à un paramètre contenue dans une distribution. Le tableau 3.1 fournit sur la base des formules précédentes l'efficacité relative obtenue avec un nombre fini plutôt qu'infini d'imputations en fonction du nombre d'imputation M et du taux d'information manquante γ. On voit qu'avec un petit nombre d'imputations, de 5 à 10, on atteint généralement une efficacité relative très élevée.

Tableau 3.1 – Efficacité relative en pourcentage d'un nombre fini M plutôt qu'infini d'imputations en fonction du taux d'information manquante, γ.

	γ					
M	0,1	0,3	0,4	0,5	0,7	0,9
3	98	95	94	93	90	88
5	99	97	96	95	94	92
10	100	99	98	98	97	96
20	100	99	99	99	98	98
∞	100	100	100	100	100	100

Tableau adapté de Rubin (1987).

Nous avons vu plus haut que pour obtenir un estimateur de l'imputation multiple asymptotiquement sans biais il faut que le modèle d'imputation contienne la variable réponse et les autres covariables du modèle d'analyse. Ajouter des variables non prédictives n'introduit pas de biais mais augmente la variance de l'estimateur des paramètres associés aux variables utiles dans le modèle d'imputation, et donc la composante inter-imputations ($\widehat{B_{IM}}$). Dans ses premières années, l'imputation multiple associait des statisticiens en charge des modèles d'imputation qui com-

plétaient des bases de données économiques et des statisticiens utilisateurs qui travaillaient sur ces bases complétées. Il était donc important que les modèles d'imputation incluent un grand nombre de variables explicatives car l'exclusion d'une variable implique que l'on soit certain qu'il n'y a pas de relation entre cette variable et la variable à imputer (Rubin et Schenker, 1991). Lorsque le modèle d'analyse et le modèle d'imputation sont élaborés par le même statisticien, l'inclusion des variables inutiles dans le modèle d'imputation peut diminuer l'efficacité de l'estimateur de l'imputation multiple.

Mise en œuvre

Dans un premier temps, l'imputation multiple était réservée aux données manquantes à structure monotone. On parle de structure monotone quand il existe J sous-ensembles $E_1, E_2, ..., E_J$ de variables tels que 1) les variables de chaque sous-ensemble sont simultanément toutes observées ou non-observées ; 2) pour chaque sujet, la non-observation des variables de E_j implique la non-observation des variables de E_k, $k > j$. Un exemple typique correspond aux sorties d'étude dans les enquêtes longitudinales. Il est alors possible d'imputer les données manquantes pour les différents profils d'observation en progressant des profils les moins incomplets aux profils les plus incomplets, et en utilisant à chaque étape les données précédemment imputées pour estimer les nouveaux modèles d'imputation (figure 3.1).

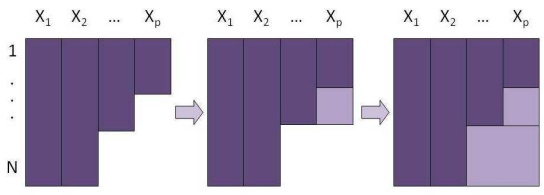


FIGURE 3.1 – Complétion séquentielle de données manquantes à structure monotone

Dans un deuxième temps, l'imputation multiple a été étendue aux structures de données manquantes non-monotones en utilisant une approche de type échantillonneur de Gibbs. La mise en œuvre peut être effectuée, entre d'autres, avec le logiciel R (R Development Core Team, 2010), en particulier, à l'aide de la bibliothèque *MICE* (*Multivariate Imputation by Chained Equations*) van Buuren et Groothuis-Oudshoorn (2011). En partant d'une première imputation plus au moins arbitraire obtenue par exemple à l'aide d'un tirage parmi les valeurs observées de chaque variable incomplète, le principe de *MICE* est d'imputer successivement les données manquantes de chaque variable incomplète conditionnellement aux données observées et aux données manquantes imputées précédemment. L'algorithme est le suivant :

Soit X_0 la matrice des variables complètes, $X_1, ..., X_p$ les p variables incomplètes, $\theta_1, ..., \theta_p$ les p vecteurs de paramètres inconnus des modèles d'imputation correspondant, X_j^{obs} la partie observée de X_j, X_j^{mis} sa partie non observée, $X_k^{(m)}$ le m-ème vecteur de données complétées de X_k, $m = 1, ..., M$, où M est le nombre d'imputations. En utilisant les distributions conditionnelles, la t-ème itération tire successivement chaque paramètre ou vecteur de donnée manquante dans sa loi conditionnelle

$$\begin{gathered}
\theta_1^{*(t)} \sim P(\theta_1 | X_0, X_1^{obs}, X_2^{(t-1)}, ..., X_p^{(t-1)}) \\
X_1^{*(t)} \sim P(X_1 | X_0, X_1^{obs}, X_2^{(t-1)}, ..., X_p^{(t-1)}, \theta_1^{*(t)}) \\
\vdots \\
\theta_p^{*(t)} \sim P(\theta_p | X_0, X_p^{obs}, X_1^{(t)}, ..., X_{p-1}^{(t)}) \\
X_p^{*(t)} \sim P(X_p | X_0, X_p^{obs}, X_1^{(t)}, ..., X_{p-1}^{(t)}, \theta_p^{*(t)})
\end{gathered}$$

où $X_j^{(t)} = (X_j^{obs}, X_j^{*(t)})$ est la j-èmè variable imputée à l'itération t.

Les trois étapes principales de *mice* sont : 1) l'imputation, donnant un objet de classe *mids (multiply imputed dataset)*, 2) l'analyse, donnant un objet de classe *mira (multiply imputed repeated analysis)* et finalement 3) la combinaison de résultats, donnant un objet de classe *mipo (multiple imputed pooled outcomes)* (Figure 3.2)

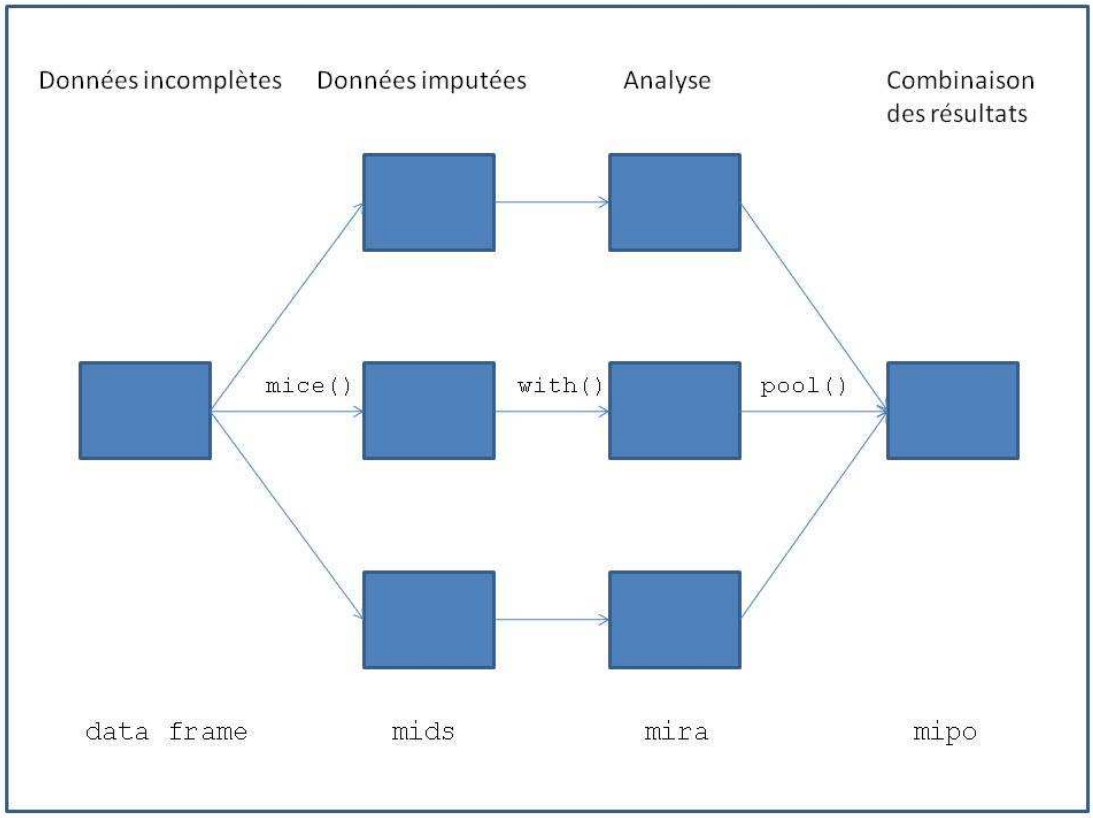


FIGURE 3.2 – Étapes principales de l'imputation multiple

Matrice de prédiction

L'ensemble des modèles d'imputations est associé à une matrice contenant des 0 et des 1, dont les lignes correspondent aux modèles d'imputation variables incomplètes et les colonnes aux variables explicatives. Chaque valeur 1 indique que la variable correspondant à cette colonne sera utilisée pour imputer la variable corres-

pondant à cette ligne. Il s'agit d'une matrice carrée de dimension égale au nombre de variables de la base de données. Les lignes correspondant aux variables complètes sont composées de 0 comme la diagonale principale de la matrice car on ne peut pas utiliser la variable à imputer dans son propre modèle d'imputation. Pour chaque variable incomplète, le modèle d'imputation impose par défaut que toutes les colonnes, sauf celle correspondant à la variable à imputer, soient utilisées en tant que prédicteurs. Mais il peut être préférable d'utiliser des modèles plus parcimonieux.

Nous donnons un exemple : supposons que les variables Z_1, Z_3 et Z_5 soient complètement observées alors que les variables Z_2 et Z_4 le sont incomplètement. La matrice de prédiction définie par défaut est :

	Z_1	Z_2	Z_3	Z_4	Z_5
Z_1	0	0	0	0	0
Z_2	1	0	1	1	1
Z_3	0	0	0	0	0
Z_4	1	1	1	0	1
Z_5	0	0	0	0	0

Ainsi, pour la variable incomplète Z_2, les modèles d'imputation définis par défaut sont

$$Z_2 = \alpha_0 + \alpha_1 Z_1 + \alpha_3 Z_3 + \alpha_4 Z_4 + \alpha_5 Z_5 + e \tag{3.9}$$

et Z_4

$$Z_4 = \alpha'_0 + \alpha'_1 Z_1 + + \alpha'_2 \tilde{Z}_2 + \alpha'_3 Z_3 + \alpha'_5 Z_5 + e \tag{3.10}$$

Noter que l'on fait intervenir la variable complétée pour Z_2, $\tilde{Z}_2$, pour imputer Z_4 et réciproquement.

Si l'on veut fait intervenir des interactions entre variables incomplètes dans un modèle d'imputation, elles doivent figurer en ligne et en colonne dans la matrice de prédiction, mais il faut veiller à ce que les interactions manquantes soient imputées de façon déterministe comme le produit des valeurs (imputées ou observées) des variables en interaction.

CHAPITRE

4 Mise en œuvre de l'imputation multiple dans les enquêtes cas-cohorte

Grâce à l'échantillonnage de la sous-cohorte, l'enquête cas-cohorte est l'un des rares cas où le caractère MA est garanti, puisque la probabilité d'observation complète ne dépend que du statut (cas ou non-cas), si l'échantillonnage est simple, et de certaines variables de phase-1, si l'échantillonnage est stratifié. L'utilisation de l'imputation multiple est donc pertinente.

Nous avons donné auparavant (2.15) l'efficacité relative asymptotique (ERA) d'une étude cas-cohorte par rapport à une étude de cohorte, qui peut être mise en relation avec la fraction d'information manquante. Ainsi :

$$\begin{aligned} \text{ERA} &\approx \left\{1 + 2\frac{1 - n_{sc}/N}{n_{sc}/N}\left[1 + \frac{1 - p_D}{p_D}\log(1 - p_D)\right]\right\}^{-1} \\ &\approx 1 - \gamma, \end{aligned} \tag{4.1}$$

Nous avons réalisé une revue de 20 études cas-cohorte publiées dans *American Journal of Epidemiology, International Journal of Epidemiology, Epidemiology* et *Atherosclerosis* entre 1993 et 2011 (Tableau 4.1). Cet échantillon non représentatif a été constitué à partir d'enquêtes cas-cohorte publiées dans l'*American Journal of Epidemiology* entre 1993 et 2007 ; nous y avons ajouté 4 article publiés dans d'autres revues. La fraction d'information manquante sur le paramètre d'intérêt, estimée par l'équation (4.1) variait de 0,05 à 0,5, avec une médiane d'environ 0,3. Avec 30% d'information manquante, $M = 5$ imputations donnent une efficacité de 0,97 et $M = 10$, 0,99 ; avec 40% d'information manquante, $M = 5$ donnent une efficacité relative de 0,96 et $M = 10$, 0,98 (Rubin, 1987). Ainsi, l'utilisation de 5 ou 10 imputation peut être considéré comme suffisant pour estimer les risques relatifs dans les enquêtes cas-cohorte.

Titre	Revue	Année	Réf	N	n_e	n_{sc}	q	ERA	γ
Impact of Institution Size, Staffing Patterns, and Infection Control Practices on Communicable Disease Outbreaks in New York State Nursing Homes	AJE	1996	143(10) :1042-9	629	61	122	0,19	0,71	0,29
Styrene exposure and ischemic heart disease : A case-cohort study	AJE	2003	158(10) :988-98	6587	498	997	0,15	0,70	0,30
Cardiac autonomic function and incident coronary heart disease : A population-based case-cohort study : The ARIC Study	AJE	1997	145(8) :696-706	15800	137	2252	0,14	0,95	0,05
Do lipids and apolipoproteins predict coronary heart disease under statin and fibrate therapy in the primary prevention setting in community-dwelling elderly subjects ? The 3C Study.	Ather.	2011	214(2) :426-31	9294	199	1081	0,12	0,86	0,14
Innate Handedness and Disease-Specific Mortality in Women	Epi	2007	18(2) :208-12	12178	252	1500	0,12	0,87	0,13
Lung cancer mortality and polynuclear aromatic hydrocarbons : A case-cohort study of aluminum production workers in Arvida, Quebec, Canada	AJE	1994	139(3) :250-62	16297	338	1138	0,07	0,78	0,22
A Prospective study of plasma ferritin level and incident diabetes : The Atherosclerosis Risk in Communities (ARIC) Study	AJE	2007	165(9) :1047-54	14406	599	690	0,05	0,54	0,46
Chlamydia pneumoniae infection and incident coronary heart disease : The Atherosclerosis Risk in Communities Study	AJE	1999	150(2) :149-56	14406	246	550	0,04	0,70	0,30
Relation of height, body mass, energy intake, and physical activity to risk of renal cell carcinoma : Results from the Netherlands cohort study	AJE	2004	160(12) :1159-67	120852	275	5000	0,04	0,95	0,05
Energy restriction in childhood and adolescence and risk of prostate cancer : Results from the Netherlands cohort study	AJE	2001	154(6) :530-7	58279	903	1630	0,03	0,65	0,35
Alcohol consumption and bladder cancer risk : Results from the Netherlands cohort study	AJE	2001	153(1) :38-41	120852	594	3170	0,03	0,85	0,15
Vegetable and fruit consumption and risks of colon and rectal cancer in a prospective cohort study the Netherlands cohort study on diet and cancer	AJE	2000	151(6) :1081-92	120852	1000	3500	0,03	0,78	0,22
Anthropometry in relation to prostate cancer risk in the Netherlands cohort study	AJE	2000	151(6) :541-9	58279	681	1688	0,03	0,72	0,28
Copper in human mammary carcinogenesis : A case-cohort study	AJE	1993	137(4) :409-14	5100	46	138	0,03	0,75	0,25
A Population-based case-cohort evaluation of the efficacy of mammographic screening for breast cancer	AJE	1994	140(10) :889-901	94656	1144	2237	0,02	0,67	0,33
Serum enterolactone concentration and the risk of coronary heart disease in a case-cohort study of finnish male smokers	AJE	2006	163(8) :687-93	29133	340	420	0,01	0,56	0,44
Occupational risk factors for esophageal and stomach cancers among female textile workers in Shanghai, China	AJE	2006	163(8) :717-25	267400	646	3188	0,01	0,83	0,17
Occupational exposures and risks of liver cancer among Shanghai female textile workers : A case-cohort study	IJE	2006	35(2) :361-9	267400	360	3186	0,01	0,90	0,10
Occupational exposures and breast cancer among women textile workers in Shanghai	Epi	2007	18(3) :383-92	267400	1709	3155	0,01	0,65	0,35
Risk in childhood of developement of severe adult obesity : retrospective, population-based case-cohort study	AJE	1998	127(1) :104-13	93800	429	938	0,01	0,69	0,31

AJE, *American Journal of Epidemiology* ; Ather., *Atherosclerosis* ; Epi, *Epidemiology* ; IJE, *International Journal of Epidemiology*

Tableau 4.1 – Revue non-exhaustive de 20 études cas-cohorte publiées. N : taille cohorte ; n_e : taille de l'ensemble de sujets qui présentent l'événement d'intérêt pendant la période de suivi ; n_{sc} : taille de la sous-cohorte ; $q = n_{sc}/N$; ERA : efficacité relative asymptotique ; γ : fraction d'information manquante.

4.1 Modèle d'imputation

Dans les études cas-cohorte, la construction du modèle d'imputation et les analyses sont, en général, menées par le même statisticien, donc seules les variables nécessaires à l'analyse doivent être incluses.

Puisque seuls les non-cas en dehors de la sous-cohorte présentent des observations manquantes, on pourrait croire que le modèle d'imputation doit être estimé à partir des non-cas de la sous-cohorte en incluant éventuellement le délai de censure parmi les variables explicatives. Si l'événement étudié est fréquent, il va y avoir un processus de sélection décalant progressivement la distribution de la variable de phase-2 chez les non-cas vers les valeurs associées à un faible risque. Cependant, si l'événement étudié est rare, comme dans la plupart des enquêtes cas-cohorte, on n'attend pas de sélection sensible et l'inclusion du délai de censure parmi les variables explicatives n'apporte pas d'information. Nous illustrerons dans la partie de validation par simulations que cette approche n'est pas pertinente.

Nous allons montrer que, sous l'hypothèse de maladie rare, le modèle d'imputation doit être estimé à partir de l'échantillon cas-cohorte en incluant l'indicatrice de statut (cas ou non-cas) parmi les variables explicatives, car les distributions de la variable de phase-2 dans ces deux sous-populations sont décalées d'une valeur qui dépend directement du risque relatif associé à la variable de phase-2.

Relation entre les distributions de Z_2 chez les cas et chez les non-cas

Soit Z_2 la variable incomplète de phase-2. On suppose que la distribution de Z_2 suit une loi de la famille exponentielle et dépend d'un vecteur de variables de

phase-1, $\tilde{Z}_2$, à travers

$$f(z_2 \mid \tilde{z}_2) = \exp\left(\frac{\theta z_2 - b(\theta) + c(z_2)}{a(\phi)}\right), \tag{4.2}$$

où ϕ est le paramètre de dispersion et le paramètre θ une fonction linéaire de $\tilde{z}_2$ et de paramètres inconnus :

$$\theta = \alpha_0 + \alpha_1' \tilde{z}_2. \tag{4.3}$$

Sous l'hypothèse de maladie rare, la distribution de Z_2 peut être considérée comme approximativement identique dans l'ensemble de la cohorte et chez les témoins.

$$f(z_2 \mid \tilde{z}_2, \Delta = 0) \simeq f(z_2 \mid \tilde{z}_2), \tag{4.4}$$

où Δ est l'indicatrice de cas. Soit $\pi(\tilde{z}_2, \mu_2, t_c)$ la probabilité d'être un cas à la fin de la période d'observation pour un sujet avec $\tilde{Z}_2 = \tilde{z}_2$, $Z_2 = \mu_2 = E[Z_2 \mid \tilde{Z}_2 = \tilde{z}_2]$ et un délai de censure $T_C = t_c$; soit $\pi(\tilde{z}_2, z_2, t_c)$ la probabilité d'être un cas à la fin de la période d'observation pour un sujet avec $\tilde{Z}_2 = \tilde{z}_2$, $Z_2 = z_2$ et $T_C = t_c$. D'après le modèle des risques proportionnels et sous l'hypothèse de maladie rare :

$$\pi(\tilde{z}_2, z_2, t_c) = \pi(\tilde{z}_2, \mu_2, t_c)\exp[\beta(z_2 - \mu_2)] \tag{4.5}$$

En supposant que la distribution du délai de censure soit indépendante de Z_2, et en intégrant sur le délai de censure, on obtient

$$\pi(\tilde{z}_2, z_2) = \pi(\tilde{z}_2, \mu_2)\exp[\beta(z_2 - \mu_2)] \tag{4.6}$$

et la distribution de Z_2 conditionnellement à $\tilde{Z}_2$ chez les cas est

$$\begin{aligned} f(z_2 \mid \Delta = 1, \tilde{z}_2) &= \frac{P[\Delta = 1 \mid \tilde{z}_2, z_2]}{P[\Delta = 1 \mid \tilde{z}_2]} f(z_2 \mid \tilde{z}_2) \\ &= \frac{\pi(\tilde{z}_2, \mu_2)\exp[\beta(z_2 - \mu_2)]}{\int \pi(\tilde{z}_2, \mu_2)\exp[\beta(z - \mu_2)] f(z \mid \tilde{z}_2) dz} f(z_2 \mid \tilde{z}_2) \\ &= \frac{\exp(\beta z_2)}{\int \exp(\beta z) f(z \mid \tilde{z}_2) dz} f(z_2 \mid \tilde{z}_2) \end{aligned} \tag{4.7}$$

Le dénominateur peut être développé comme :

$$\begin{aligned} \int \exp(\beta z) f(z \mid \tilde{z}_2) dz &= \int \exp\left(\beta z + \frac{\theta z_2 - b(\theta) + c(z_2)}{a(\phi)}\right) dz \\ &= \int \exp\left(\frac{(\theta + a(\phi)\beta)z - b(\theta + a(\phi)\beta) + b(\theta + a(\phi)\beta) - b(\theta) + c(z)}{a(\phi)}\right) dz \\ &= \exp\left[\frac{b(\theta + a(\phi)\beta) - b(\theta)}{a(\phi)}\right] \end{aligned} \tag{4.8}$$

avec θ donné par (4.3). Les résultats (4.8) et (4.4) mènent à :

$$\begin{aligned} f(z_2 \mid \Delta = 1, \tilde{z}_2) &\simeq \exp\left(\frac{a(\phi)\beta z_2 - b(\theta + a(\phi)\beta) + b(\theta)}{a(\phi)}\right) f(z_2 \mid \Delta = 0, \tilde{z}_2) \\ &= \exp\left(\frac{[\theta + a(\phi)\beta]z_2 - b[\theta + a(\phi)\beta] + c(z_2)}{a(\phi)}\right) \end{aligned} \tag{4.9}$$

Résultat

Sous l'hypothèse de maladie rare, les distributions de la variable de phase-2, Z_2, sachant le statut du sujet, cas ou non-cas, sont décalées d'une quantité de $a(\phi)\beta$ sur l'échelle du lien canonique, c-à-d, de $\sigma^2\beta$ pour un modèle linéaire, β sur l'échelle logit pour un modèle logistique ou β sur l'échelle log pour un modèle log-linéaire, où β est le coefficient associé à Z_2 dans le modèle à risques proportionnels que l'on veut estimer.

La construction du modèle d'imputation, en particulier le choix du vecteur de variables $\tilde{z}_2$, est crucial pour procéder à l'imputation multiple. En pratique, en plus de l'indicatrice de cas et des variables de stratification, si la sous-cohorte a été sélectionnée par un tirage stratifié, il est nécessaire d'ajuster sur les variables incluses dans le modèle de Cox et éventuellement d'autres variables prédictives, si elles sont disponibles.

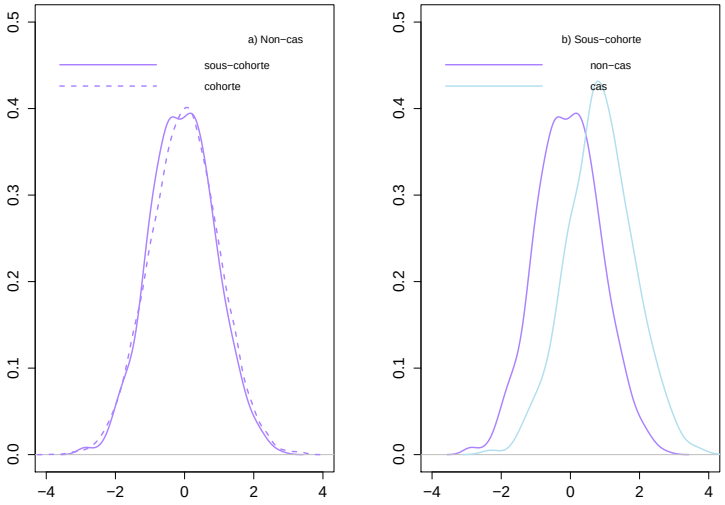


FIGURE 4.1 – Distribution observée de Z_2 : a) échantillon de non-cas, b) échantillon cas-cohorte.

Nous verrons dans la partie validation par simulations les conséquences d'une mauvaise spécification du modèle d'imputation, d'une part, au niveau du choix des variables explicatives ; d'autre part, au niveau de la loi de la famille exponentielle que la variable incomplète est supposée suivre.

4.2 Mise en œuvre avec R

Toutes les simulations et analyses ont été réalisés avec le logiciel R (R Development Core Team, 2010). En particulier, la mise en œuvre de l'imputation multiple a été effectuée à l'aide de la bibliothèque *MICE* (*Multivariate Imputation by Chained Equations*) van Buuren et Groothuis-Oudshoorn (2011).

Matrice de prédiction

Afin de faciliter la mise en œuvre de l'imputation multiple pour l'analyse des enquêtes cas-cohorte, nous avons développé la fonction *mat.pred.MI* sous R (disponible en annexe). Cette fonction nécessite les modèles d'analyse (modèle de Cox) et d'imputation (modèles linéaire généralisé), avec des interactions, le cas échéant, et la ou les variables de phase-2. A partir des seuls modèles définis par l'utilisateur, *mat.pred.MI* génère la matrice de prédiction associée aux modèles d'analyse et d'imputation, prenant en compte la relation entre les variables incomplètes et les variables qui leur sont liées ainsi que la relation avec l'événement d'intérêt. Par ailleurs, s'il y a des interactions impliquant une variable de phase-2, le terme d'interaction est calculé comme le produit des variables concernées.

Nous avons également développé la fonction *cch.MI* (disponible en annexe) pour analyser les enquêtes cas-cohorte par imputation multiple, à l'aide de la matrice de prédiction aussi définie.

CHAPITRE

5 Capacité prédictive d'un modèle ou d'une variable additionnelle

La performance des modèles de prédiction de la survenue d'un événement peut être évaluée de différents points de vue. Steyerberg *et al.* (2010) en distinguent trois : 1) la ressemblance entre les prédictions et les réponses observées comme le pourcentage de variance expliquée (coefficient de détermination R^2) ou score de Brier (1950) ; 2) la calibration, qui quantifie la proximité entre les probabilités prédites par le modèle et celles observées comme le fait la statistique du test de Hosmer-Lemeshow (2000) ; et 3) la discrimination, qui quantifie la capacité du modèle à différencier les sujets avec ou sans l'événement d'intérêt et qui utilise des mesures comme la sensibilité, la spécificité ou l'aire sous la courbe ROC (*Receiver Operating Characteristics*) (Metz, 1978; Zweig et Campbell, 1993).

Données censurées

Dans l'analyse de survie, on est très souvent confronté au phénomène de censure qui perturbe l'estimation de la capacité prédictive. Dans ce contexte, différentes mesures de la variabilité expliquée adaptées au modèle des risques proportionnels de Cox ont été proposées (Henderson, 1995; Kent et O'Quigley, 1988; Schemper, 1990). Concernant la discrimination, Harrell *et al.* (1982) ont proposé l'indice C, qui tient compte de la censure et permet d'estimer correctement la capacité prédictive d'un modèle ou d'une variable additionnelle. D'autres mesures complémentaires afin d'évaluer l'amélioration de la capacité prédictive associée à l'inclusion d'une variable additionnelle ont été proposes par Pencina *et al.* (2008). Nous nous sommes intéressés à la capacité prédictive en termes de discrimination.

5.1 C de Harrell

L'indice C proposé par Harrell *et al.* (1982) mesure la concordance entre l'ordre des délais de survie prédits et observés. Il s'agit d'une adaptation, dans le cadre de données censurées, de la statistique U de Mann-Whitney (1947), qui mesure la concordance entre deux variables quelconques V et W. La statistique U est fondée sur la considération de toutes les paires de sujets (i, j), et le décompte des paires concordantes, c'est-à-dire, celles pour lesquelles, en supposant l'absence d'ex-aequo, $V_i > V_j$ et $W_i > W_j$ ou $V_i < V_j$ et $W_i < W_j$.

$$U = \sum_{i=1}^{N} \sum_{j=1}^{N} I[(V_i > V_j \ \texttt{et} \quad W_i > W_j) \ \texttt{ou} \quad (V_i < V_j \ \texttt{et} \quad W_i < W_j)] \qquad (5.1)$$

où I désigne une variable indicatrice.

En présence de censure, la comparaison des délais de survie prédits et observés pour toutes les paires de sujets n'est pas possible. Rappelons que T_i est le délai de survie d'un sujet i, $i = 1, \ldots, N$, C_i le délai de censure et que l'on observe $X_i = min(T_i, C_i)$. Harrell *et al.* (1996) appellent paires utilisables celles dont l'ordre des délais de survie prédits peut être comparé à l'ordre des délais de survie réels, c'est-à-dire, les paires formées par deux sujets non censurés et celles formées par un sujet non censuré et un sujet censuré après la date d'événement du sujet non censuré. Les paires formées par deux sujets censurés n'apportent pas d'information sur la concordance entre les délais de survie prédits par le modèle et les délais de survie réels, car ceux-ci sont inconnus. De façon analogue, une paire formée par un sujet dont le délai de survie est observé et un sujet censuré avant cette date, n'apporte pas d'information sur la concordance car le délai de survie inconnu peut être inférieur ou supérieur au délai observé.

Pencina et D'Agostino (2004) ont montré que, sous les modèles à risques proportionnels, les probabilités de survie prédites à un temps fixé t, $Y_i = P(T_i > t)$, peuvent être substituées aux délais de survie prédits dans les comparaisons.

L'indice C est :

$$C = \frac{\pi_c}{\pi_c + \pi_d} \tag{5.2}$$

où π_c est la probabilité de concordance de la paire (i, j) :

$$\pi_c = P(X_i < X_j \ \texttt{et} \quad Y_i < Y_j) + P(X_i > X_j \ \texttt{et} \quad Y_i > Y_j) \tag{5.3}$$

et π_d est la probabilité de discordance de la paire (i, j) :

$$\pi_d \;=\; P(X_i < X_j \;\texttt{et}\quad Y_i > Y_j) + P(X_i > X_j \;\texttt{et}\quad Y_i < Y_j) \tag{5.4}$$

Nous supposons que les délais de survie comme les probabilités de survie prédites sont continus, $P(X_i = X_j) = P(Y_i = Y_j) = 0$, donc, $\pi_c + \pi_d = 1$.

La probabilité de concordance, π_c, peut être estimée par la proportion de paires concordantes, p_c,

$$p_c = \frac{1}{Q} \sum_i \sum_{j<i} c_{ij}, \tag{5.5}$$

où Q est le nombre de paires utilisables et où

$$c_{ij} = \begin{cases} 1 \quad \texttt{si} \quad (X_i < X_j \;\texttt{et}\quad Y_i < Y_j) \;\texttt{ou}\quad (X_i > X_j \;\texttt{et}\quad Y_i > Y_j) \\ 0 \quad \text{sinon} \end{cases}$$

De façon analogue, la probabilité de discordance, π_d, peut être estimé par le nombre de paires discordantes, p_d,

$$p_d = \frac{1}{Q} \sum_i \sum_{j<i} d_{ij}, \tag{5.6}$$

où

$$d_{ij} = \begin{cases} 1 \quad \texttt{si} \quad (X_i < X_j \;\texttt{et}\quad Y_i > Y_j) \;\texttt{ou}\quad (X_i > X_j \;\texttt{et}\quad Y_i < Y_j) \\ 0 \quad \text{sinon} \end{cases}$$

L'indice C peut alors être estimé par la proportion de paires concordantes parmi les paires utilisables

$$\widehat{C} = \frac{p_c}{p_c + p_d} \tag{5.7}$$

et sa variance peut être estimée selon Kremers (2007) par :

$$\widehat{var}(\widehat{C}) = 4\frac{(\sum_i c_i)^2}{Q^2}\sum_i \left[\frac{(\sum_i c_i)^2}{c_i^2} + \frac{(\sum_i (c_i + d_i))^2}{Q^2} - 2\frac{\sum_i c_i \sum_i d_i}{Q c_i}\right] \tag{5.8}$$

Supposons que nous voulions estimer la capacité prédictive d'une variable additionnelle. Soient C_1 la capacité prédictive d'un modèle M_1, et C_2 la capacité prédictive du modèle M_2 défini comme le modèle M_1 plus la variable additionnelle. La valeur prédictive de la variable additionnelle est définie par $\Delta = C_2 - C_1$ et estimée par $\widehat{\Delta} = \widehat{C_2} - \widehat{C_1}$. Les variances de $\widehat{C_1}$ et $\widehat{C_2}$, respectivement Var_{C1} et Var_{C2}, sont estimées par 5.8, mais la variance de $\widehat{\Delta}$ doit être estimée par bootstrap car elle n'a pas de forme analytique.

Données cas-cohorte

Dans les données issues d'une enquête cas-cohorte les paires utilisables ne sont définies que dans l'échantillon cas-cohorte puisque les valeurs prédites ne sont pas calculables pour les sujets incomplètement observés. Cela peut entrainer un biais sur l'estimation de la capacité prédictive : l'échantillon des cas est sur-représenté, par conséquent, une proportion artificiellement élevée de paires formées par deux cas est observée. Généralement, les cas ont, en moyenne, un risque d'événement plus élevé que les non-cas. En moyenne, les paires formées par deux cas seront

plus semblables en termes de risque, et donc plus facilement discordantes que les paires formées par un cas et un non-cas. On peut donc craindre que l'application naïve de formules précédentes aux sujets de l'échantillon cas-cohorte ne fournisse des mesures biaisées des valeurs prédictives auxquelles on s'intéresse. En revanche, l'imputation multiple reconstitue des cohortes complètes, et permet d'estimer facilement n'importe quelle statistique que l'on sait estimer sur une cohorte, comme la capacité prédictive d'un modèle ou d'une variable additionnelle.

La mise en œuvre de l'imputation multiple dans ce cadre ne pose aucun problème spécifique. Nous reconstruisons des cohortes complètes à l'aide de l'imputation multiple. Pour chaque cohorte complétée, nous pouvons estimer directement les quantités C_1, Var_{C1}, C_2, Var_{C2} et Δ. Puis, à l'aide des équations 3.4 et 3.5, nous pouvons obtenir les estimations de l'imputation multiple de ces quantités. Concernant la variance de $\widehat{\Delta}$, la composante inter-imputations peut être estimée par la variance empirique des M estimations de Δ, fournies par les M jeux de données complétés. Pour obtenir la composante intra-imputations de la variance de $\widehat{\Delta}$ il est nécessaire de bootstraper chacun des jeux des données complétées.

5.2 NRI et IDI

Récemment, les indices NRI (*Net Reclassification Improvement*) et IDI (*Integrated Discrimination Index*) ont été proposées par Pencina *et al.* (2008) comme mesures complémentaires pour évaluer l'amélioration de la capacité prédictive associée à l'inclusion d'une ou plusieurs variables additionnelles dans le modèle dans le cadre d'une cohorte.

NRI

L'indice NRI suppose que l'on ait défini un petit nombre de niveaux de risque pertinents cliniquement. Il quantifie l'amélioration de la classification des sujets lorsqu'une variable additionnelle est introduite dans le modèle. On aimerait que le risque pour les sujets non censurés augmente après l'ajout d'une variable et qu'il diminue pour les sujets censurés. On attribue un score 1 aux sujets qui passent dans une catégorie de risque plus élevé (passage que nous appelons « $+$ »), un score 0 à ceux qui restent dans la même catégorie et un score -1 à ceux qui passent dans une catégorie de risque plus faible (passage que nous appelons « $-$ »). L'indice NRI est la somme de deux composantes concernant l'une les sujets non censurés ($\Delta = 1$), l'autre les censurés ($\Delta = 0$) :

$$\text{NRI} = [P(+|\Delta = 1) - P(-|\Delta = 1)] + [P(-|\Delta = 0) - P(+|\Delta = 0)] \tag{5.9}$$

et peut être estimé par

$$\widehat{\text{NRI}} = [\hat{P}(+|\Delta = 1) - \hat{P}(-|\Delta = 1)] + [\hat{P}(-|\Delta = 0) - \hat{P}(+|\Delta = 0)] \tag{5.10}$$

où

$$\hat{P}(+|\Delta = 1) = \frac{\text{Nombre de cas passant dans une catégorie à plus haut risque}}{\text{Nombre de cas}} \tag{5.11}$$

$$\hat{P}(-|\Delta = 1) = \frac{\text{Nombre de cas passant dans une catégorie à plus faible risque}}{\text{Nombre de cas}} \tag{5.12}$$

$$\hat{P}(+|\Delta = 0) = \frac{\text{Nombre de non-cas passant dans une catégorie à plus haut risque}}{\text{Nombre de non-cas}} \tag{5.13}$$

$$\hat{P}(-|\Delta = 0) = \frac{\text{Nombre de non-cas passant dans une catégorie à plus faible risque}}{\text{Nombre de non-cas}} \tag{5.14}$$

Sa variance est estimée par (Pencina, 2008) :

$$\widehat{Var}(\widehat{\text{NRI}}) = \frac{\hat{P}(+|\Delta = 1) + \hat{P}(-|\Delta = 1)}{\text{Nombre de cas}} + \frac{\hat{P}(+|\Delta = 0) + \hat{P}(-|\Delta = 0)}{\text{Nombre de non-cas}} \tag{5.15}$$

IDI

L'indice IDI peut être envisagé comme une version continue de l'indice NRI, en utilisant les probabilités au lieu des catégories de risque. Pencina *et al.* (2008) appelle X la probabilité prédite d'événement dans l'intervalle [0,T] et f la densité de X. Au seuil u, $0 < u < 1$, la sensibilité, $S(u)$, et 1 - spécificité, $P(u)$, sont respectivement définies comme :

$$S(u) = P(X > u|\Delta = 1) = \int_u^1 f(x|\Delta = 1)dx \tag{5.16}$$

$$P(u) = P(X > u|\Delta = 0) = \int_u^1 f(x|\Delta = 0)dx \tag{5.17}$$

Les intégrales de ces deux quantités sont respectivement

$$\text{IS} = \int_0^1 S(u)du = \int_0^1 \int_u^1 f(x|\Delta = 1)dxdu \tag{5.18}$$

$$\text{IP} = \int_0^1 P(u)du = \int_0^1 \int_u^1 f(x|\Delta = 0)dxdu \tag{5.19}$$

En échangeant l'ordre d'intégration, on obtient

$$\begin{aligned} \mathsf{IS} &= \int_0^1 \int_0^x f(x|\Delta=1)dudx = \int_0^1 xf(x|\Delta=1)dx \\ &= E(X|\Delta=1) \end{aligned} \tag{5.20}$$

$$\begin{aligned} \mathsf{IP} &= \int_0^1 \int_0^x f(x|\Delta=0)dudx = \int_0^1 xf(x|\Delta=0) \\ &= E(X|\Delta=0) \end{aligned} \tag{5.21}$$

Les espérances conditionnelles sont estimés par les probabilités moyennes d'événement chez les cas et chez les non-cas. Ainsi, l'IDI entre deux modèle, l'un sans la variable additionnelle (« sans »), l'autre avec (« avec »), est défini comme :

$$\mathsf{IDI} = (\mathsf{IS}_{avec} - \mathsf{IS}_{sans}) - (\mathsf{IP}_{avec} - \mathsf{IP}_{sans}) \tag{5.22}$$

Il s'agit d'une mesure de l'écart entre les probabilités prédites des sujets non censurés et censurés selon les modèles avec et sans la variable additionnelle.

$$\widehat{\mathsf{IDI}} = (\bar{\hat{P}}_{avec|\Delta=1} - \bar{\hat{P}}_{sans|\Delta=1}) - (\bar{\hat{P}}_{avec|\Delta=0} - \bar{\hat{P}}_{sans|\Delta=0}) \tag{5.23}$$

où $\bar{\hat{P}}_{avec|\Delta=1}$ (respectivement $\bar{\hat{P}}_{sans|\Delta=1}$) est la probabilité moyenne d'événement pour les sujets non censures prédite par le modèle avec (respectivement sans) la variable additionnelle et $\bar{\hat{P}}_{avec|\Delta=0}$ (respectivement $\bar{\hat{p}}_{sans|\Delta=0}$) est la probabilité moyenne d'événement pour les sujets censurés prédite par le modèle avec (respectivement sans) la variable additionnelle.

Ou de façon équivalente

$$\widehat{\mathsf{IDI}} = (\bar{\hat{P}}_{avec|\Delta=1} - \bar{\hat{P}}_{avec|\Delta=0}) - (\bar{\hat{P}}_{sans|\Delta=1} - \bar{\hat{P}}_{sans|\Delta=0}) \tag{5.24}$$

La variance de $\widehat{\mathsf{IDI}}$ peut être estimée par :

$$\widehat{Var}(\widehat{\mathsf{IDI}}) = \hat{Var}_{\Delta=1} + \hat{Var}_{\Delta=0} \tag{5.25}$$

où $\hat{Var}_{\Delta=1}$ (respectivement $\hat{Var}_{\Delta=0}$) est la variance observée des différences de prédiction selon les modèles avec ou sans la variable additionnelle parmi les sujets non censurés (respectivement censurés).

Données cas-cohorte

Avec les cohortes complètes reconstituées par imputation multiple et les équations 3.4 et 3.5, nous pouvons obtenir les estimations de l'imputation multiple des indices NRI, IDI et leurs variances respectives.

CHAPITRE

6 Validation par simulations

Nous avons réalisé une étude de simulation pour valider l'utilisation de l'imputation multiple dans l'analyse des enquêtes cas-cohorte. En ce qui concerne les risques relatifs, nous avons évalué le biais des différents estimateurs, comparé leur précision, et évalué la robustesse de l'estimateur de l'imputation multiple quand le modèle d'imputation est mal spécifié. Nous avons également illustré les problèmes rencontrés quand on estime le modèle d'imputation à partir des seuls non-cas. En ce qui concerne les mesures de valeurs prédictives nous avons comparé les estimations fournies par l'imputation multiple dans les enquêtes cas-cohorte à celles obtenues à partir de la cohorte entière et nous avons illustré les problèmes rencontrés avec l'application naïve des formules usuelles aux seuls sujets de l'échantillon cas-cohorte. Pour cela, nous avons utilisé d'une part des données entièrement simulées, d'autre part, des enquêtes cas-cohorte simulées à partir de la cohorte PRIME et de la cohorte NWTS.

6.1 Données complètement simulées

6.1.1 Conditions générales des simulations

Nous avons généré trois variables de phase-1, Z_1, $\widetilde{Z}_2$, Z_3, observées sur la cohorte entière, et une variable de phase-2, Z_2, disponible seulement sur l'échantillon cas-cohorte : Z_1 était une variable binaire, $\widetilde{Z}_2$ était une variable prédictive de la variable de phase-2, $\widetilde{Z}_2 \equiv Z_2 + \varepsilon$ où $\varepsilon \sim N(0, \sigma^2)$ était indépendante de Z_2, et Z_3 était une variable gaussienne. La variable de phase-2, Z_2, était aussi une variable gaussienne. Z_1 et Z_2 étaient indépendantes, tandis que la corrélation entre Z_2 et Z_3 était de 0,2. Le délai jusqu'à l'événement suivait une distribution exponentielle de paramètre $\lambda = exp(\beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3)$, avec β_1, β_2 et β_3 fixés à 0 ou $log(2) = 0,693$. Le délai de censure suivait une distribution uniforme sur l'intervalle $[0, \tau]$, où τ était tel que la probabilité d'événement fût 0,03 (τ=0,025). Nous avons simulé 1000 cohortes de taille N =10000 et un échantillon cas-cohorte pour chacune des 1000 cohortes pour lequel 5 imputations ont été réalisées.

Nous désirions estimer l'effet de Z_2 sur la survenue de l'événement après ajustement sur Z_1 et Z_3, ainsi que sa capacité prédictive.

Pour ce qui est de l'estimation de l'effet de Z_2 sur la survenue d'événement et la comparaison des estimations fournies par imputation multiple à celles de l'analyse pondérée, dans un premier temps nous avons sélectionné des sous-cohortes de taille $n_{sc} = 300$ et $n_{sc} = 1000$ par tirage simple, i.e. des fréquences d'échantillonnage de $q = 0,03$ et $q = 0,1$ respectivement. Pour sélectionner des sous-cohortes par tirage stratifié, nous avons divisé la cohorte en neuf strates basées sur les tertiles de $\widetilde{Z}_2$ et Z_3, considérant deux niveaux de corrélation ρ entre Z_2 et sa variable prédictive $\widetilde{Z}_2$,

$\rho \approx 0{,}7$ (σ^2=1) et $\rho \approx 0{,}3$ (σ^2=9). Puis nous avons sélectionné des sous-cohortes de taille $n_{sc} = 300$ et $n_{sc} = 1000$ par tirage stratifié. La distribution des sujets par strate est donnée au tableau 6.1.

Tableau 6.1 – Distribution des sujets de la sous-cohorte par strate.

		Corrélation $(Z_2, \widetilde{Z}_2)$=0.7		Corrélation $(Z_2, \widetilde{Z}_2)$=0.3	
Strate	SC	Cohorte[1]	Cas[1]	Cohorte[1]	Cas[1]
Tertile 1 $\widetilde{Z}_2$, tertile 1 Z_3	80	1296	7	1196	10
Tertile 1 $\widetilde{Z}_2$, tertile 2 Z_3	80	1107	14	1109	21
Tertile 2 $\widetilde{Z}_2$, tertile 1 Z_3	80	1106	10	1109	11
Tertile 3 $\widetilde{Z}_2$, tertile 1 Z_3	80	931	15	1128	12
Tertile 1 $\widetilde{Z}_2$, tertile 3 Z_3	100	929	27	1030	47
Tertile 2 $\widetilde{Z}_2$, tertile 2 Z_3	100	1121	23	1116	25
Tertile 2 $\widetilde{Z}_2$, tertile 3 Z_3	150	1106	55	1108	63
Tertile 3 $\widetilde{Z}_2$, tertile 2 Z_3	150	1105	38	1106	30
Tertile 3 $\widetilde{Z}_2$, tertile 3 Z_3	180	1298	117	1198	89
Total	1000	10000	308	10000	308

SC : Sous-cohorte
[1]Moyenne des 1000 réplications.

Pour les deux modes d'échantillonnage, nous avons modélisé la relation entre la variable incomplète, Z_2, et les variables complètes qui lui sont liées, $\widetilde{Z}_2$ et Z_3 à l'aide d'un modèle linéaire ajusté à l'échantillon cas-cohorte et incluant l'indicatrice des cas, les variables de stratification et les variables du modèle de Cox comme variables explicatives :

$$Z_2 = \alpha_0 + \alpha_1 Ind_{cas} + \alpha_2 \widetilde{Z}_2 + \alpha_3 Z_3 + \alpha_4 Z_1 + e \tag{6.1}$$

Pour ce qui est des analyses pondérées, l'estimateur proposé par Lin et Ying (1993) a été mis en œuvre pour les sous-cohortes sélectionnées par un tirage simple, l'estimateur II proposé par Borgan *et al.* (2000) quand le tirage a été stratifié.

6.1.2 Résultats

6.1.2.1 Estimation des log de risques relatifs

Les résultats des simulations avec un échantillonnage simple figurent dans le tableau 6.2, celles des simulations avec un échantillonnage stratifié pour une sous-cohorte de taille 300 dans les tableau 6.3 et pour une sous-cohorte de taille 1000 dans le tableau 6.4. Ils conduisent à des conclusions similaires.

Quand l'effet des covariables était nul, $\beta_1 = \beta_2 = \beta_3 = 0$, aucun biais n'a été observé sur les effets des variables de phase-1, β_1 et β_3, ni sur celui de la variable de phase-2, β_2, quel que soit le type d'échantillonnage ou la taille de la sous-cohorte. Quand les variables étaient associées à une augmentation du risque d'événement, $\beta_1 = \beta_2 = \beta_3 = log(2) = 0,693$, les trois analyses fournissaient des estimations moyennes analogues : entre 0,683 et 0,727 avec la cohorte entière, entre 0,684 et 0,728 avec l'analyse pondérée, entre 0,659 et 0,724 avec l'imputation multiple. Cependant, les simulations réalisées avec un échantillonnage stratifié de la sous-cohorte conduisait à une légère sous-estimation (de -2% à -5%) des log de risques relatifs.

Globalement (tableaux 6.2, 6.3, 6.4), les écart-types estimés coïncidaient avec les écart-types observés des estimations sauf en ce qui concerne les résultats de l'analyse pondérée en présence d'une petite sous-cohorte et d'un effet des variables sur la survenue d'événement (tableau 6.3). Pour les variables de phase-1, les estimations fournies par l'imputation multiple avaient une précision similaire à celles de la cohorte entière, et supérieure à celle de l'analyse pondérée (rapport des écart-types de l'analyse pondérée et l'imputation multiple entre 1,07 et 1,70 quand l'effet de la

variable de phase-2 était nul, et entre 1,07 et 2,01 quand l'effet de la variable de phase-2 était non nul). Pour la variable de phase-2, comme attendu, les estimations obtenues par l'imputation multiple étaient moins précises que celles de la cohorte entière mais plus précises que celles de l'analyse pondérée (rapport des écart-types de l'analyse pondérée et l'imputation multiple entre 1,05 et 1,20 quand l'effet de la variable de phase-2 était nul, et entre 1,10 et 1,50 quand l'effet de la variable de phase-2 était non nul).

Tableau 6.2 – Paramètres estimés : échantillonnage simple

	SC = 300					SC = 1000				
	Est	$\widehat{ET}$	ET	PR	Ratio	Est	$\widehat{ET}$	ET	PR	Ratio
$\beta_1 = 0$										
Cohorte	-0,001	0,106	0,100	98,6		-0,001	0,106	0,100	98,6	
IM	-0,001	0,106	0,100	98,6		-0,001	0,106	0,100	98,6	
LY	-0,002	0,172	0,171	97,3	1,62	-0,002	0,127	0,122	98,3	1,20
$\beta_2 = 0$										
Cohorte	-0,003	0,054	0,055	97,5		-0,003	0,054	0,055	97,5	
IM	-0,004	0,070	0,073	96,3		-0,002	0,059	0,060	97,3	
LY	-0,002	0,080	0,082	97,4	1,14	-0,002	0,062	0,063	97,2	1,05
$\beta_3 = 0$										
Cohorte	-0,001	0,052	0,050	97,7		-0,001	0,052	0,050	97,7	
IM	-0,001	0,053	0,051	97,9		-0,001	0,053	0,050	97,6	
LY	-0,002	0,069	0,068	97,5	1,30	-0,002	0,057	0,056	97,6	1,08
$\beta_1 = 0.693$										
Cohorte	0,727	0,279	0,283	97,3		0,727	0,279	0,283	97,3	
IM	0,723	0,284	0,287	97,2		0,724	0,283	0,286	97,1	
LY	0,728	0,323	0,323	96,8	1,14	0,723	0,303	0,303	96,9	1,07
$\beta_2 = 0.693$										
Cohorte	0,700	0,093	0,094	94,6		0,700	0,093	0,094	94,6	
IM	0,687	0,132	0,130	89,9		0,690	0,109	0,105	91,0	
LY	0,687	0,160	0,163	96,8	1,21	0,713	0,120	0,120	95,9	1,10
$\beta_3 = 0.693$										
Cohorte	0,688	0,092	0,092	92,8		0,688	0,092	0,092	92,8	
IM	0,684	0,098	0,095	93,3		0,682	0,094	0,093	91,8	
LY	0,701	0,147	0,147	95,5	1,50	0,684	0,108	0,109	95,3	1,15

Est : Paramètre estimé moyen des 1000 simulations, $\widehat{ET}$: écart-type estimé moyen,
ET :écart-type observé, PR : % de recouvrement, Ratio : rapport des $\widehat{ET}$ de l'analyse pondérée et l'imputation multiple, LY : estimateur de Lin et Ying,
IM Modèle d'imputation : $Z_2 = \alpha_0 + \alpha_1 Ind_{cas} + \alpha_2 \widetilde{Z}_2 + \alpha_3 Z_3 + \alpha_4 Z_1 + e$.

Tableau 6.3 – Paramètres estimés : échantillonnage stratifié, $n_{sc} = 300$

	Corrélation $(Z_2, \widetilde{Z}_2)$=0.3					Corrélation $(Z_2, \widetilde{Z}_2)$=0.7				
	Est	$\widehat{ET}$	ET	PR	Ratio	Est	$\widehat{ET}$	ET	PR	Ratio
$\beta_1 = 0$										
Cohorte	-0,003	0,107	0,100	99,0		-0,003	0,107	0,100	99,0	
IM	-0,003	0,107	0,100	99,0		-0,003	0,107	0,100	99,0	
BII	-0,000	0,181	0,180	97,2	1.69	-0,002	0,182	0,173	98,1	1.70
$\beta_2 = 0$										
Cohorte	-0,000	0,054	0,058	97,0		-0,000	0,054	0,058	97,0	
IM	-0,001	0,079	0,084	96,5		0,000	0,070	0,074	97,1	
BII	0,003	0,090	0,097	96,9	1.14	0,000	0,084	0,088	97,1	1.20
$\beta_3 = 0$										
Cohorte	-0,004	0,053	0,056	96,7		-0,004	0,053	0,056	96,7	
IM	-0,004	0,055	0,058	96,9		-0,004	0,054	0,0570	96,3	
BII	-0,001	0,068	0,067	96,7	1.24	-0,003	0,068	0,068	97,2	1.26
$\beta_1 = 0.693$										
Cohorte	0,689	0,118	0,113	97,8		0,689	0,118	0,113	97,8	
IM	0,667	0,120	0,112	96,9		0,675	0,120	0,112	96,9	
BII	0,685	0,228	0,237	96,4	1.90	0,681	0,241	0,253	96,4	2.01
$\beta_2 = 0.693$										
Cohorte	0,687	0,058	0,057	97,6		0,687	0,058	0,057	97,6	
IM	0,659	0,090	0,096	91,2		0,672	0,082	0,085	93,6	
BII	0,707	0,121	0,143	95,8	1.34	0,702	0,123	0,145	96,2	1.50
$\beta_3 = 0.693$										
Cohorte	0,683	0,057	0,057	96,0		0,683	0,057	0,057	96,0	
IM	0,677	0,062	0,062	95,2		0,679	0,060	0,059	95,2	
BII	0,687	0,104	0,118	94,7	1.68	0,685	0,107	0,120	96,1	1.78

Est : Paramètre estimé moyen des 1000 simulations, $\widehat{ET}$: écart-type estimé moyen, ET :écart-type observé, PR : % de recouvrement, Ratio : rapport des $\widehat{ET}$ de l'analyse pondérée et l'imputation multiple, IM Modèle d'imputation : $Z_2 = \alpha_0 + \alpha_1 Ind_{cas} + \alpha_2 Strata + \alpha_3 Z_1 + e$, BII : estimateur II de Borgan.

Tableau 6.4 – Paramètres estimés : échantillonnage stratifié $n_{sc} = 1000$

	Corrélation $(Z_2, \widetilde{Z}_2)$=0.3					Corrélation $(Z_2, \widetilde{Z}_2)$=0.7				
	Est	$\widehat{ET}$	ET	PR	Ratio	Est	$\widehat{ET}$	ET	PR	Ratio
$\beta_1 = 0$										
Cohorte	-0,003	0,107	0,100	99,0		-0,003	0,107	0,100	99,0	
IM	-0,003	0,107	0,100	99,0		-0,003	0,107	0,100	99,0	
BII	-0,005	0,132	0,127	97,7	1.23	-0,001	0,133	0,128	98,0	1.24
$\beta_2 = 0$										
Cohorte	-0,000	0,054	0,058	97,0		-0,000	0,054	0,058	97,0	
IM	0,001	0,063	0,066	97,5		-0,001	0,060	0,062	97,3	
BII	0,000	0,067	0,072	96,4	1.06	0,000	0,065	0,068	97,6	1.08
$\beta_3 = 0$										
Cohorte	-0,004	0,053	0,056	96,7		-0,004	0,053	0,056	96,7	
IM	-0,004	0,054	0,057	96,7		-0,004	0,054	0,057	96,4	
BII	-0,004	0,058	0,061	97,9	1.07	-0,003	0,057	0,060	97,3	1.09
$\beta_1 = 0.693$										
Cohorte	0,689	0,118	0,113	97,8		0,689	0,118	0,113	97,8	
IM	0,669	0,120	0,111	96,7		0,676	0,119	0,112	97,5	
BII	0,692	0,162	0,157	97,6	1.35	0,696	0,168	0,165	97,2	1.41
$\beta_2 = 0.693$										
Cohorte	0,687	0,058	0,057	97,6		0,687	0,058	0,057	97,6	
IM	0,664	0,073	0,072	94,1		0,670	0,070	0,068	95,8	
BII	0,698	0,086	0,090	97,3	1.18	0,701	0,088	0,097	97,9	1.26
$\beta_3 = 0.693$										
Cohorte	0,683	0,057	0,057	96,0		0,683	0,057	0,057	96,0	
IM	0,676	0,059	0,059	94,5		0,679	0,058	0,058	95,2	
BII	0,686	0,078	0,081	95,8	1.32	0,689	0,080	0,086	96,1	1.38

Est : Paramètre estimé moyen des 1000 simulations, $\widehat{ET}$: écart-type estimé moyen,
ET :écart-type observé, PR : % de recouvrement, Ratio : rapport des $\widehat{ET}$ de l'analyse pondérée et l'imputation multiple, IM Modèle d'imputation : $Z_2 = \alpha_0 + \alpha_1 Ind_{cas} + \alpha_2 Strata + \alpha_3 Z_1 + e$, BII : estimateur II de Borgan.

6.1.2.2 Estimation de la capacité prédictive

Les résultats concernant la capacité prédictive figurent dans le tableau 6.5. Nous avons estimé la capacité prédictive C_1 (respectivement C_2) du modèle sans (respectivement avec) la variable de phase-2, la capacité prédictive de celle-ci, $\Delta = C_2 - C_1$, et ses indices NRI et IDI. La variance de la capacité prédictive de la variable additionnelle de phase-2 a été obtenue par bootstrap, celles des autres estimateurs à partir des formules (5.8), (5.15) et (5.25). Nous avons estimé ces quantités dans les cohortes entières, dans les échantillons cas-cohorte et dans les cohortes reconstituées par imputation multiple. Les sous-cohortes étaient de taille n_{sc} =1000 ou n_{sc}=300.

Pour un effet nul des covariables (H_0), les trois méthodes fournissaient des résultats similaires, centrés sur la même valeur. Les écart-types estimés de C étaient également analogues mais très supérieurs aux écart-types observés (0,033 au lieu de 0,012 avec la cohorte entière). Au contraire, un biais négatif était observé pour l'écart-type de l'indice IDI dans les analyses limitées à l'échantillon cas-cohorte, ce qui entrainait un mauvais contrôle du risque de première espèce. Pour la valeur prédictive de la variable de phase-2, Δ, et pour l'indice NRI, les écart-types étaient correctement estimés.

Quand l'effet des covariables était non nul ($\beta_1 = \beta_3 = log(2)$ et $\beta_2 = log(1,5)$ ou $\beta_2 = log(2)$), l'imputation multiple fournissait des estimations équivalentes à celles de l'analyse de la cohorte entière. En revanche, dans l'analyse limitée à l'échantillon cas-cohorte, un biais négatif sensible était observé lors de l'utilisation de l'estimateur naïf de C alors qu'un biais également sensible mais positif était observé sur Δ, NRI et IDI. Le biais était évidemment plus élevé quand la taille de la sous-cohorte était plus petite. Comme sous H_0, les écart-types de C étaient surestimés, ceux de

IDI sous-estimés et ceux de Δ et NRI sans biais. Sous $\beta_2 = log(2)$, la puissance des test des hypothèses $\Delta = 0$, NRI=0, IDI=0 effectués avec la cohorte entière ou par imputation multiple atteignait 100%. Sous $\beta_2 = log(1,5)$, les puissances observées étaient inférieures, en particulier pour NRI, mais du même ordre de grandeur avec la cohorte entière et avec l'imputation multiple.

Tableau 6.5 – Estimation moyenne de la capacité prédictive (Est), écart-type moyen estimé ($\widehat{ET}$) et écart-type observé (ET). Résultats des 1000 simulations.

	$\beta_1 = \beta_2 = \beta_3 = 0$				$\beta_1 = \beta_3 = log(2), \beta_2 = log(1.5)$				$\beta_1 = \beta_2 = \beta_3 = log(2)$			
	Est	$\widehat{ET}$	ET	% H_0 rejeté	Est	$\widehat{ET}$	ET	% H_0 rejeté	Est	$\widehat{ET}$	ET	% H_0 rejeté
Cohorte												
C_1	0.518	0.033	0.012		0.727	0.032	0.015		0.733	0.029	0.014	
C_2	0.524	0.033	0.013		0.747	0.031	0.015		0.782	0.029	0.014	
Δ	0.006	0.010	0.009	3.7	0.020	0.007	0.007	91.6	0.049	0.010	0.010	100
NRI	0.007	0.017	0.019	4.8	0.071	0.030	0.033	52.5	0.167	0.034	0.035	99.9
IDI	$2e^{-4}$	$2e^{-4}$	$3e^{-4}$	6.0	0.014	0.003	0.005	99.9	0.048	0.006	0.009	99.9
IM1000												
C_1	0.518	0.033	0.012		0.724	0.032	0.016		0.733	0.029	0.014	
C_2	0.526	0.033	0.013		0.745	0.031	0.016		0.783	0.027	0.014	
Δ	0.008	0.012	0.010	3.4	0.021	0.008	0.008	90.6	0.049	0.010	0.011	100
NRI	0.009	0.019	0.017	1.5	0.076	0.033	0.033	64.8	0.172	0.037	0.036	100
IDI	$3e^{-4}$	$3e^{-4}$	$4e^{-4}$	3.5	0.014	0.004	0.005	99.0	0.045	0.008	0.010	100
IM300												
C_1	0.518	0.033	0.012		0.724	0.032	0.016		0.733	0.029	0.014	
C_2	0.528	0.033	0.012		0.745	0.031	0.017		0.783	0.027	0.015	
Δ	0.010	0.014	0.011	3.0	0.021	0.008	0.009	84.6	0.050	0.011	0.012	100
NRI	0.013	0.023	0.018	1.3	0.076	0.035	0.035	57.0	0.172	0.039	0.039	99.7
IDI	$4e^{-4}$	$4e^{-4}$	$5e^{-4}$	1.8	0.014	0.005	0.006	87.5	0.046	0.010	0.012	100
CC1000												
C_1	0.528	0.032	0.013		0.667	0.033	0.015		0.670	0.031	0.014	
C_2	0.534	0.033	0.015		0.709	0.032	0.022		0.737	0.029	0.014	
Δ	0.006	0.010	0.010	4.7	0.043	0.011	0.017	100	0.067	0.012	0.012	100
NRI	0.017	0.031	0.033	6.7	0.147	0.039	0.043	96.7	0.261	0.041	0.043	100
IDI	0.002	0.001	0.003	15.2	0.058	0.009	0.014	100	0.114	0.011	0.017	100
CC300												
C_1	0.523	0.034	0.013		0.620	0.037	0.016		0.620	0.034	0.015	
C_2	0.529	0.034	0.015		0.647	0.036	0.016		0.668	0.032	0.015	
Δ	0.006	0.010	0.009	3.6	0.027	0.011	0.011	83.3	0.048	0.013	0.013	99.8
NRI	0.019	0.039	0.043	6.2	0.154	0.043	0.050	94.4	0.257	0.046	0.051	99.9
IDI	0.002	0.001	0.003	13.9	0.040	0.008	0.014	99.8	0.078	0.010	0.017	100

IM300, IM1000 : estimations de l'imputation multiple avec des sous-cohortes de taille, respectivement, 300 et 1000 ;
CC300, CC1000 : estimations de l'échantillon cas-cohorte avec des sous-cohortes de taille, respectivement, 300 et 1000 ;
$\widehat{ET}$, écart-type estimé moyen, ET, écart-type observé ; C_1, indice C du modèle de Cox sans la variable de phase-2 ;
C_2, indice C du modèle de Cox avec la variable de phase-2 ; Δ, Capacité prédictive de la variable de phase-2 ; $H_0 : \Delta = 0$;
NRI, *Net reclassification index* en incluant la variable de phase-2, H_0 : NRI = 0.
IDI, *Integrated discrimination index* en incluant la variable de phase-2, H_0 : IDI = 0.

6.1.2.3 Robustesse de l'imputation multiple quand le modèle d'imputation est mal spécifié

Nous avons envisagé deux types de mauvaise spécification. D'une part, contrairement au résultat démontré dans le chapitre 4, l'estimation du modèle d'imputation à partir des seuls non-cas de la sous-cohorte, d'autre part, l'imputation de la variable de phase-2 selon une distribution gaussienne alors qu'elle était simulée sous une autre loi.

- Imputation à partir des seuls non-cas de la sous-cohorte

Un scénario illustratif où les modèles d'imputation ont été construits à partir des non-cas de la sous-cohorte est présenté. Nous avons repris les conditions de simulation précédentes en ce qui concerne la génération des autres variables (variables de phase-1, délai jusqu'à l'événement et délai de censure), et en particulier le choix des paramètres $\beta_1 = \beta_2 = \beta_3$ fixés à 0 ou $log(2)$. Deux modèles d'imputation distincts ont été envisagés :

$$\texttt{MI1}: Z_2 = \alpha_0 + \alpha_1 Z_1 + \alpha_2 \widetilde{Z}_2 + \alpha_3 Z_3 + e \tag{6.2}$$

$$\texttt{MI2}: Z_2 = \alpha_0 + \alpha_1 Z_1 + \alpha_2 \widetilde{Z}_2 + \alpha_3 Z_3 + \alpha_4 \texttt{délai de censure} + e \tag{6.3}$$

Les résultats sont donnés au tableau 6.6. Sous H_0 aucun biais n'était observé. Sous $\beta_1 = \beta_2 = \beta_3 = log(2)$ aucun biais n'était observé sur l'effet β_1 de la variable de phase-1 indépendante de celle de phase-2. En revanche, un biais notable (environ 15%) était observé sur l'effet de la variable de phase-2 comme sur celui de la variable de phase-1 qui lui était corrélée, Z_3, (ρ=0,2). Les résultats étaient pratiquement identiques que le modèle d'imputation inclue (MI2) ou non (MI1) le délai de censure et que la variable $\widetilde{Z}_2$ soit moyennement (ρ=0,3) ou fortement (ρ=0,7) prédictive de Z_2.

Tableau 6.6 – Paramètres estimés à partir de l'échantillon des non-cas. Échantillonnage stratifié. $n_{sc} = 1000$

	Corrélation $(Z_2, \widetilde{Z}_2)$*=0,3*				*Corrélation* $(Z_2, \widetilde{Z}_2)$*=0,7*			
	Est	$\widehat{ET}$	*ET*	*PR*	*Est*	$\widehat{ET}$	*ET*	*PR*
$\beta_1 = 0$								
Cohorte	0,003	0,121	0,121	95,2	0,002	0,124	0,121	95,2
MI1	0,003	0,121	0,121	95,0	0,002	0,124	0,121	95,4
MI2	0,003	0,121	0,121	95,5	0,002	0,124	0,121	95,4
$\beta_2 = 0$								
Cohorte	0,002	0,060	0,062	95,0	-0,001	0,063	0,062	95,0
MI1	0,002	0,066	0,068	71,6	0,002	0,067	0,065	70,7
MI2	0,002	0,066	0,068	70,9	0,002	0,066	0,065	70,7
$\beta_3 = 0$								
Cohorte	0,000	0,062	0,061	92,7	-0,001	0,058	0,061	92,7
MI1	0,000	0,060	0,059	55,3	-0,001	0,056	0,059	59,1
MI2	0,000	0,060	0,059	56,2	-0,001	0,056	0,059	58,4
$\beta_1 = 0.6931$								
Cohorte	0,690	0,136	0,131	95,0	0,695	0,133	0,131	95,0
MI1	0,677	0,135	0,132	95,0	0,681	0,131	0,132	95,0
MI2	0,677	0,135	0,132	95,0	0,681	0,131	0,132	95,0
$\beta_2 = 0.6931$								
Cohorte	0,691	0,062	0,063	95,4	0,690	0,063	0,063	95,4
MI1	0,794	0,074	0,074	95,7	0,795	0,071	0,073	94,9
MI2	0,792	0,074	0,074	95,7	0,794	0,072	0,073	95,5
$\beta_3 = 0.6931$								
Cohorte	0,696	0,064	0,062	93,9	0,694	0,060	0,062	93,9
MI1	0,800	0,060	0,062	94,5	0,801	0,057	0,061	94,5
MI2	0,800	0,060	0,062	94,5	0,801	0,057	0,061	94,5

Est : Paramètre estimé moyen des 1000 simulations, $\widehat{ET}$: écart-type estimé moyen,
ET : écart-type observé, PR : % de recouvrement,
MI1 : Modèle d'imputation 1 : $Z_2 = \alpha_0 + \alpha_1 Z_1 + \alpha_2 \widetilde{Z}_2 + \alpha_3 Z_3 + e$,
MI2 : MI1 + délai de censure

- Mauvaise spécification de la distribution de la variable de phase-2

Nous avons simulé la variable de phase-2 sous trois distributions différentes : log-normale, uniforme ou Student à 5 degrés de liberté, toutes avec une variance égale à 1. Les autres conditions de simulations étaient identiques à celles utilisées précédemment. Nous avons réalisé les imputations avec le modèle 6.1, dans lequel la distribution de la variable de phase-2 est supposée gaussienne.

Les résultats sont donnés dans le tableau 6.7. Comme attendu, l'analyse de la cohorte entière et l'analyse pondérée fournissaient des estimations des log de risques relatifs non-biaisées. Pour un modèle d'imputation mal spécifié, l'estimation de l'effet $\beta_2 = log(2)$ de la variable de phase-2, Z_2, fournie par l'imputation multiple était biaisé (-14%) quand la distribution de Z_2 était log-normale. Quand Z_2 était distribuée selon une loi uniforme ou Student à 5 degrés de liberté, l'estimation de l'effet de la variable de phase-2 était moins biaisée (respectivement -5% et -7%). Pour un modèle d'imputation mal spécifié et un effet de la variable de phase-2 nul, aucun biais n'était observé. L'écart-type de l'imputation multiple et celui de l'analyse pondérée coïncidaient avec les dispersions des estimations observées. La dispersion observée était toujours plus petite avec l'imputation multiple qu'avec l'estimateur pondéré. Pour les variables de phase-1, la dispersion était similaire avec la cohorte entière et avec l'imputation multiple, quelle que soit la distribution de la variable de phase-2. Pour l'estimation de l'effet de la variable de phase-2, les dispersions observées étaient plus petites avec l'imputation multiple qu'avec l'estimateur pondéré et comme attendu, légèrement plus grands que l'analyse de la cohorte complète.

Tableau 6.7 – Mauvaise spécification de la distribution de la variable de phase-2 dans le modèle d'imputation

	Cohorte entière			Imputation multiple[a]			Estimateur pondéré (Lin et Ying)		
	Est	$\widehat{ET}$	ET	Est	$\widehat{ET}$	ET	Est	$\widehat{ET}$	ET
Distribution log-normale de Z_2									
$\beta_1 = \beta_2 = \beta_3 = 0$									
β_1	-0,017	0,107	0,100	-0,017	0,107	0,100	-0,003	0,132	0,124
β_2	-0,006	0,027	0,034	0,005	0,031	0,032	-0,001	0,034	0,037
β_3	-0,008	0,053	0,056	-0,015	0,054	0,058	0,001	0,059	0,062
$\beta_1 = \beta_2 = \beta_3 = 0,693$									
β_1	0,685	0,058	0,056	0,638	0,061	0,060	0,686	0,112	0,117
β_2	0,685	0,013	0,015	0,598	0,015	0,011	0,695	0,020	0,023
β_3	0,678	0,029	0,031	0,680	0,032	0,027	0,687	0,049	0,053
Distribution uniforme de Z_2									
$\beta_1 = \beta_2 = \beta_3 = 0$									
β_1	-0,010	0,179	0,170	-0,010	0,179	0,170	0,007	0,197	0,188
β_2	-0,002	0,091	0,092	0,002	0,093	0,092	-0,002	0,098	0,095
β_3	0,003	0,089	0,096	0,002	0,089	0,096	0,004	0,093	0,093
$\beta_1 = \beta_2 = \beta_3 = 0,693$									
β_1	0,688	0,119	0,118	0.686	0.123	0.122	0,699	0,166	0,167
β_2	0,692	0,069	0,065	0.657	0.076	0.069	0,703	0,087	0,083
β_3	0,692	0,057	0,051	0.691	0.059	0.055	0,699	0,080	0,080
Distribution de Student à 5 ddl de Z_2									
$\beta_1 = \beta_2 = \beta_3 = 0$									
β_1	0,010	0,107	0,109	0,010	0,107	0,109	0,010	0,132	0,133
β_2	-0,007	0,053	0,051	-0,004	0,062	0,059	-0,006	0,065	0,063
β_3	0,001	0,053	0,049	0,000	0,054	0,049	-0,001	0,058	0,053
$\beta_1 = \beta_2 = \beta_3 = 0,693$									
β_1	0,696	0,118	0,117	0,683	0,121	0,120	0,705	0,170	0,180
β_2	0,678	0,043	0,041	0,641	0,053	0,046	0,687	0,070	0,079
β_3	0,689	0,057	0,048	0,682	0,059	0,050	0,700	0,081	0,077

[a] Estimations fournies par l'imputation multiple avec le modèle d'imputation :
$Z_2 = \alpha_0 + \alpha_1 Ind_{cas} + \alpha_2 Z_1 + \alpha_3 Strata + e$, ddl : degrés de liberté.

Nous avons également estimé la capacité prédictive dans des situations où la distribution utilisée pour effectuer les imputations ne coïncidait pas avec la vraie distribution de la variable de phase-2 (Tableau 6.8). En présence d'une mauvaise spécification de la distribution de la variable de phase-2, l'imputation multiple continuait à donner des résultats voisins de ceux de la cohorte entière, et le risque de première espèce du test $\Delta = 0$ était bien contrôlé.

Tableau 6.8 – Capacité prédictive des modèles sans et avec la variable de phase-2. Résultats des 1000 simulations.

	Cohorte entière				Imputation multiple			
	Est	$\widehat{ET}$	ET	% H_0 rejeté	Est	$\widehat{ET}$	ET	% H_0 rejeté
Distribution log-normale de Z_2								
$\beta_1 = \beta_2 = \beta_3 = 0$								
C_1	0,518	0,033	0,012		0,518	0,033	0,012	
C_2	0,524	0,033	0,013		0,520	0,031	0,016	
Δ	0,006	0,010	0,009	5,5	0,002	0,013	0,012	4,2
$\beta_1 = \beta_2 = \beta_3 = log(2)$								
C_1	0,784	0,013	0,006		0,784	0,013	0,006	
C_2	0,881	0,011	0,006		0,866	0,011	0,006	
Δ	0,097	0,005	0,005	100	0,082	0,005	0,004	100
Distribution uniforme de Z_2								
$\beta_1 = \beta_2 = \beta_3 = 0$								
C_1	0,532	0,055	0,019		0,532	0,055	0,019	
C_2	0,540	0,055	0,019		0,541	0,055	0,020	
Δ	0,008	0,015	0,013	2,2	0,009	0,017	0,013	4,0
$\beta_1 = \beta_2 = \beta_3 = log(2)$								
C_1	0,733	0,029	0,014		0,733	0,029	0,014	
C_2	0,781	0,027	0,012		0,785	0,027	0,012	
Δ	0,048	0,009	0,009	100	0,052	0,010	0,010	100
Distribution de Student à 5 ddl de Z_2								
$\beta_1 = \beta_2 = \beta_3 = 0$								
C_1	0,518	0,033	0,011		0,518	0,033	0,011	
C_2	0,523	0,033	0,011		0,525	0,033	0,011	
Δ	0,005	0,009	0,007	2,0	0,007	0,012	0,008	5,0
$\beta_1 = \beta_2 = \beta_3 = log(2)$								
C_1	0,741	0,029	0,013		0,741	0,029	0,013	
C_2	0,793	0,027	0,013		0,786	0,027	0,013	
Δ	0,052	0,009	0,009	100	0,045	0,010	0,009	100

$\widehat{ET}$, écart-type estimé moyen, ET, écart-type observé,
C_1, indice C du modèle de Cox sans la variable de phase-2,
C_2, indice C du modèle de Cox avec la variable de phase-2.
Δ, Capacité prédictive de la variable de phase-2, $H_0 : \Delta = 0$,
ddl, degrés de liberté.

6.2 Données cas-cohorte simulées depuis la cohorte PRIME

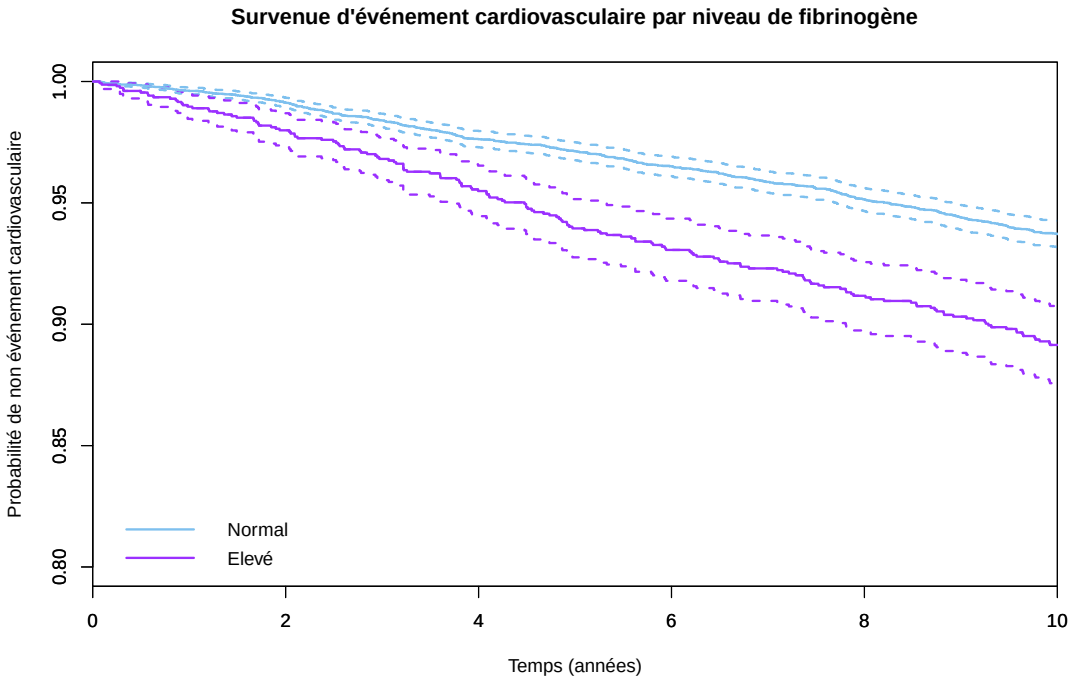
6.2.1 Description des données

L'étude PRIME (PRospective sur l'Infarctus du MyocardE) est une étude de cohorte multicentrique à laquelle ont participé trois centres en France (Lille, Strasbourg, Toulouse) et un centre en Irlande du Nord (Belfast). Le recrutement des sujets, tirés au sort sur les listes électorales et basé sur le volontariat, avait eu lieu entre 1991 et 1993. Chaque centre avait inclus environ 2500 sujets, ce qui a permis de constituer une cohorte totale de 9510 hommes, dont 642 (6,7%) ont présenté un événement cardio-vasculaire pendant les 10 ans de suivi.

Nous avons simulé des enquêtes cas-cohorte à partir de la cohorte entière afin de comparer les estimations des risques relatifs obtenues à partir de la cohorte complètement observée et à partir des échantillons cas-cohorte simulés, analysés par imputation multiple ou par estimation pondérée. Dans un second temps, nous nous sommes intéressés à la mesure de la valeur prédictive de la variable de phase-2. La démarche de validation a comporté la simulation de 1000 sous-cohortes de taille 700 (environ 7% de la cohorte), et 1000 sous-cohortes de taille 2100 (environ 20% de la cohorte) dans la cohorte PRIME.

6.2.2 Mise en œuvre

Le fibrinogène est une protéine dont la concentration plasmatique augmente dans les états inflammatoires. Elle joue également un rôle important dans la formation de caillots et les niveaux de fibrinogène élevés sont associés à une augmentation du risque d'événement cardiovasculaire (Scarabin *et al.*, 2003).



Dans les enquêtes cas-cohorte simulées, nous avons supposé que le fibrinogène était une variable de phase-2, donc non disponible pour les non-cas en dehors de la sous-cohorte. Conformément aux analyses réalisées par Scarabin *et al.* (2003), nous avons choisi d'estimer l'effet du fibrinogène sur la survenue d'un événement cardiovasculaire, après ajustement sur le pays de recrutement, qui valait 0 pour les trois centres en France et 1 pour le centre en Irlande du Nord, et les caractéristiques du sujet à l'entrée dans l'étude : âge (années), cholestérol total (g/l), cholestérol HDL (g/l), pression artérielle systolique (mmHg) et consommation de tabac (g/j).

Un des principaux facteurs déterminant le niveau de fibrinogène est le tabac (Scarabin *et al.*, 2003). La consommation totale de tabac en grammes/jour était calculée en fonction du nombre de cigarettes, cigares, cigarillos et pipes consommés, autorapporté par les sujets. Ainsi, la consommation journalière correspondait au nombre de cigarettes + 4*(nombre de cigares) + nombre de cigarillos + 1,5*(nombre de pipes). Dans ces données, la corrélation entre la consommation de tabac et les niveaux de fibrinogène était de 0,15 (Figure 6.1), nous ne disposions donc pas d'une variable fortement prédictive du niveau de fibrinogène.

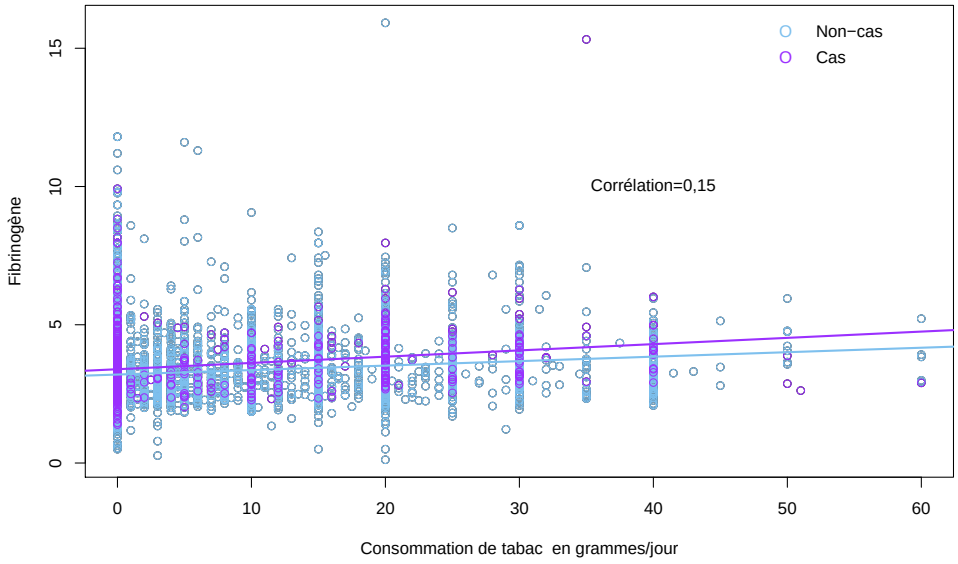


FIGURE 6.1 – Relation entre les niveaux de fibrinogène et la consommation de tabac

Cependant, le tabac est un facteur de risque cardiovasculaire important et nous l'avons utilisé pour stratifier la cohorte de la façon suivante : strate 1 : Non-fumeurs, strate 2 : Ex-fumeurs, strate 3 : Fumeurs (0 - 10) gr/jour, strate 4 : Fumeurs [10 - 20) gr/jour, strate 5 : Fumeurs [20 -) gr/jour.

Les sous-cohortes, de taille 2100 (environ 20% de la cohorte) ou de taille 700 (environ 7% de la cohorte), ont été choisies par échantillonnage stratifié, avec des fréquences d'échantillonnage de 0,16, 0,20, 0,21, 0,25 et 0,46 respectivement, pour les strates 1-5, pour la sous-cohorte de taille 2100 et un tiers de celles-ci pour la sous-cohortes de taille 700 (tableau 6.9). Donc, 80% ou 93% des sujets de la cohorte avaient des données manquantes pour la variable de phase-2.

Tableau 6.9 – Distribution des cas par strate.

Strate	Cas (%)	Taille strate	$n_{sc} = 1000$	$n_{sc} = 300$
1 Non-fumeurs	153 (5,3)	2888	475	158
2 Ex-fumeurs	261 (6,4)	4076	800	268
3 Fumeurs 1-9 gr/jour	51 (6,3)	814	175	58
4 Fumeurs 10-19 gr/jour	55 (7,8)	705	175	58
5 Fumeurs $\geq$ 20 gr/jour	122 (11,9)	1027	475	158
Total	642 (6,7)	9510	2100	700

Les distributions des niveaux de fibrinogène selon la strate pour les cas et les non-cas sont données par la figure 6.2.

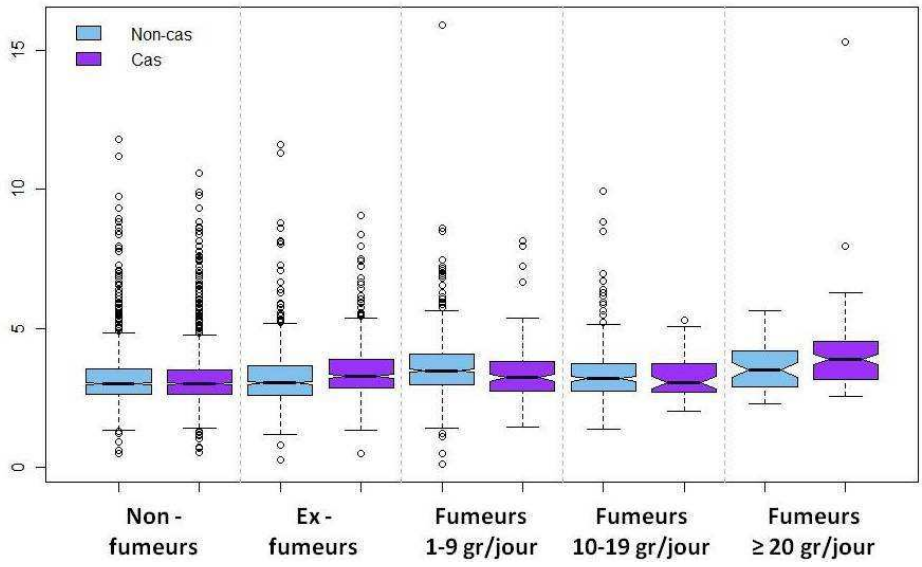


FIGURE 6.2 – Distribution du fibrinogène par strate et statut

Nous avons estimé les risques relatifs sur l'ensemble de la cohorte PRIME. Puis, pour analyser les enquêtes cas-cohorte simulées, nous avons utilisé l'estimateur II de Borgan *et al.* (2000) proposé pour un échantillonnage stratifié, la calibration proposée par Breslow *et al.* (2009b) en calibrant, d'une part, sur les seuls résidus *dfbeta* du fibrinogène, d'autre part, sur les résidus de toutes les variables incluses dans le modèle d'analyse et l'imputation multiple. Puis, nous avons calculé l'efficacité relative des différentes méthodes par rapport à l'analyse de la cohorte complète comme le rapport de la variance estimée de l'estimation unique fournie par la cohorte entière et la moyenne des variances estimées par chacune des méthodes utilisées.

La mise en œuvre de l'imputation multiple a comporté la construction d'un modèle linéaire d'imputation du fibrinogène (Fb), ajusté sur l'échantillon cas-cohorte incluant l'indicatrice de cas (Statut), la variable de stratification (Strate), l'âge (Age), l'indicatrice de centre=Belfast (Centre), le cholestérol total (Chol), le cholestérol HDL (HDL) et la tension artérielle systolique (TAS).

$$\mathsf{Fb} = \alpha_0 + \alpha_1 \mathsf{Statut} + \alpha_2 \mathsf{Strate} + \alpha_3 \hat{\mathsf{A}}\mathsf{ge} + \alpha_4 \mathsf{Centre} + \alpha_5 \mathsf{Chol} + \alpha_6 \mathsf{HDL} + \alpha_7 \mathsf{TAS} + e \qquad (6.4)$$

Nous avons également estimé la capacité prédictive du fibrinogène, à l'aide de l'indice C de Harrell *et al.* (1996) et des indices NRI et IDI proposées par Pencina *et al.* (2008).

6.2.3 Résultats

Globalement, les estimations des risques relatifs fournies par les analyses pondérées classique ou calibrée et par imputation multiple étaient proches de celles fournies par l'analyse de la cohorte entière, pour les sous-cohortes de taille 2100 (tableau 6.10), comme pour les sous-cohortes de taille 700 (tableau 6.11).

Pour les variables de phase-1, les analyses pondérées classique et calibrées fournissaient des estimations moins précises que l'analyse de la cohorte entière. L'analyse calibrée sur les résidus *dfbeta* de toutes les variables d'ajustement, fournissait des estimations légèrement plus précises que les deux autres analyses pondérées (classique et calibré sur les résidus du fibrinogène), sans atteindre la précision des estimations fournies par l'analyse de la cohorte entière. En revanche, la précision des estimations fournies par l'imputation multiple était très proche de celle fournie par la cohorte entière. L'efficacité relative des analyses pondérées par rapport à la cohorte entière variait entre 0,84 et 0,94 pour une grande sous-cohorte (tableau 6.10) et entre 0,59 et 0,79 pour une petite sous-cohorte (tableau 6.11). L'efficacité relative de l'analyse par imputation multiple par rapport à la cohorte entière était au moins de 0,98 pour les deux tailles de sous-cohorte.

Pour le fibrinogène, variable de phase-2, et une grande sous-cohorte, l'efficacité relative par rapport à la cohorte entière était de 0,84 pour l'estimation fournie par les différentes analyses pondérées. Elle était légèrement plus petite (0,81) pour l'imputation multiple (tableau 6.10). Comme attendu, l'efficacité relative était sensiblement plus faible avec une petite sous-cohorte, (0,62), pour les estimations de l'analyse pondérée classique et calibrée uniquement sur les résidus *dfbeta* du fibrinogène et 0,69 pour celle de l'analyse calibrée sur les résidus *dfbeta* de toutes les variables d'ajustement comme pour l'imputation multiple (tableau 6.11).

Tableau 6.10 – Estimation des risques relatifs (RR), intervalles de confiance à 95% (IC 95%) et efficacité relative (ER) de l'analyse cas-cohorte par rapport à la cohorte entière. Échantillonnage stratifié. Sous-cohortes de taille 2100. Résultats des 1000 simulations.

	Cohorte		Cas-cohorte									Imputation		
	PRIME		Borgan II			Cal1[a]			Cal2[b]			Multiple[c]		
	RR	IC 95%	RR	IC 95%	ER	RR	IC 95%	ER	RR	IC 95%	ER	RR	IC 95%	ER
Fibrinogène	1,14	(1,06 ; 1,22)	1,14	(1,05 ; 1,24)	0,84	1,14	(1,05 ; 1,24)	0,84	1,14	(1,05 ; 1,24)	0,84	1,15	(1,06 ; 1,25)	0,81
Tabac[d]	1,24	(1,16 ; 1,34)	1,24	(1,15 ; 1,34)	0,94	1,24	(1,15 ; 1,34)	0,93	1,24	(1,15 ; 1,34)	0,94	1,24	(1,16 ; 1,34)	0,99
Âge	1,06	(1,03 ; 1,09)	1,06	(1,03 ; 1,09)	0,85	1,06	(1,03 ; 1,09)	0,85	1,06	(1,03 ; 1,09)	0,87	1,06	(1,03 ; 1,09)	1,00
Centre	1,32	(1,12 ; 1,56)	1,33	(1,09 ; 1,61)	0,85	1,33	(1,09 ; 1,61)	0,85	1,33	(1,09 ; 1,61)	0,87	1,32	(1,11 ; 1,56)	0,99
Cholestérol total[d]	1,08	(1,06 ; 1,10)	1,08	(1,06 ; 1,11)	0,85	1,08	(1,06 ; 1,11)	0,85	1,08	(1,06 ; 1,11)	0,88	1,08	(1,06 ; 1,10)	1,00
Cholestérol HDL[d]	0,59	(0,52 ; 0,68)	0,59	(0,50 ; 0,69)	0,84	0,59	(0,50 ; 0,70)	0,84	0,59	(0,50 ; 0,69)	0,86	0,59	(0,51 ; 0,68)	1,00
PAS[d,e]	1,11	(1,08 ; 1,14)	1,11	(1,07 ; 1,15)	0,85	1,11	(1,07 ; 1,15)	0,85	1,11	(1,07 ; 1,15)	0,87	1,11	(1,08 ; 1,16)	1,00

[a] Cal1 : Poids calibrés sur les résidus *dfbeta* de la variable fibrinogène

[b] Cal2 : Poids calibrés sur les résidus *dfbeta* de toutes les variables du modèle l'analyse

[c] Modèle d'imputation incluant l'indicatrice de cas et les variables d'ajustement du modèle d'analyse

[d] coefficients multipliés par 10

[e] PAS, pression artérielle systolique

Tableau 6.11 – Estimation des risques relatifs (RR), intervalles de confiance à 95% (IC 95%) et efficacité relative (ER) de l'analyse cas-cohorte par rapport à la cohorte entière. Échantillonnage stratifié. Sous-cohortes de taille 700. Résultats des 1000 simulations.

	Cohorte		Cas-cohorte									Imputation		
	PRIME		Borgan II			Cal1[a]			Cal2[b]			Multiple[c]		
	RR	IC 95%	RR	IC 95%	ER	RR	IC 95%	ER	RR	IC 95%	ER	RR	IC 95%	ER
Fibrinogène	1,14	(1,06 ; 1,22)	1,15	(1,03 ; 1,28)	0,62	1,15	(1,03 ; 1,28)	0,62	1,15	(1,03 ; 1,28)	0,69	1,14	(1,03 ; 1,27)	0,69
Tabac[d]	1,24	(1,16 ; 1,34)	1,24	(1,14 ; 1,36)	0,79	1,24	(1,13 ; 1,36)	0,78	1,24	(1,14 ; 1,36)	0,79	1,24	(1,16 ; 1,34)	0,99
Âge	1,06	(1,03 ; 1,09)	1,06	(1,02 ; 1,10)	0,66	1,06	(1,02 ; 1,10)	0,66	1,06	(1,02 ; 1,10)	0,71	1,06	(1,03 ; 1,09)	0,98
Centre	1,32	(1,12 ; 1,56)	1,32	(1,02 ; 1,71)	0,65	1,32	(1,02 ; 1,71)	0,65	1,33	(1,04 ; 1,69)	0,70	1,32	(1,11 ; 1,57)	1,00
Cholestérol total[d]	1,08	(1,06 ; 1,10)	1,08	(1,05 ; 1,12)	0,65	1,08	(1,05 ; 1,12)	0,65	1,08	(1,05 ; 1,11)	0,70	1,08	(1,06 ; 1,10)	0,99
Cholestérol HDL[d]	0,59	(0,52 ; 0,68)	0,59	(0,48 ; 0,72)	0,67	0,58	(0,48 ; 0,72)	0,67	0,59	(0,48 ; 0,71)	0,71	0,59	(0,51 ; 0,68)	0,99
PAS[d,e]	1,11	(1,08 ; 1,14)	1,11	(1,06 ; 1,17)	0,59	1,11	(1,06 ; 1,17)	0,59	1,11	(1,06 ; 1,17)	0,66	1,11	(1,08 ; 1,15)	0,98

[a] Cal1 : Poids calibrés sur les résidus *dfbeta* de la variable fibrinogène

[b] Cal2 : Poids calibrés sur les résidus *dfbeta* de toutes les variables du modèle d'analyse

[c] Modèle d'imputation incluant l'indicatrice de cas et les variables d'ajustement du modèle d'analyse

[d] coefficients multipliés par 10

[e] PAS, pression artérielle systolique

Les valeurs prédictives fournies par la cohorte entière et par imputation multiple étaient similaires. Alors que le fibrinogène augmentait significativement le risque d'événement cardio-vasculaire, il était associé à des indices NRI et IDI significativement positifs, mais sa capacité prédictive mesurée par Δ n'était pas significative (tableau 6.12).

Tableau 6.12 – Distribution des cas par strate.

	Événement cardiovasculaire			
	Cohorte PRIME		Imputation multiple	
	Estimation	IC 95%	Estimation	IC 95%
C_1	0,6795	(0,6399 - 0,7191)	0,6795	(0,6399 - 0,7191)
C_2	0,6834	(0,6438 - 0,7230)	0,6831	(0,6434 - 0,7227)
Δ	0,0039	(-0,0073 - 0,0151)	0,0035	(-0,0016 - 0,0086)
NRI	0,0152	(0,0040 - 0,0264)	0,0153	(0,0039 - 0,0267)
IDI	0,0005	(0,0001 - 0,0009)	0,0003	(0,0000 - 0,0005)

C_1, C indice du modèle d'analyse sans le fibrinogène.
C_2, C indice du modèle d'analyse avec le fibrinogène.
Δ, Capacité prédictive du fibrinogène.
NRI, *Net reclassification index* en incluant le fibrinogène
IDI, *Integrated discrimination index* en incluant le fibrinogène

6.3 Données NWTS

La cohorte NWTS (D'Angio *et al.*, 1989) issue de *The National Wilms Tumor Study Group* est constituée de 3915 enfants atteints d'une tumeur de Wilms diagnostiquée entre 1980 et 1994. Ils ont été suivis jusqu'à la détection d'un signe de progression de la maladie ou au décès, ce que nous avons considéré comme l'événement d'intérêt. La figure 6.3 montre le taux de survie sans rechute et son intervalle de confiance à 95% en fonction de l'évaluation histologique de la tumeur, favorable ou défavorable.

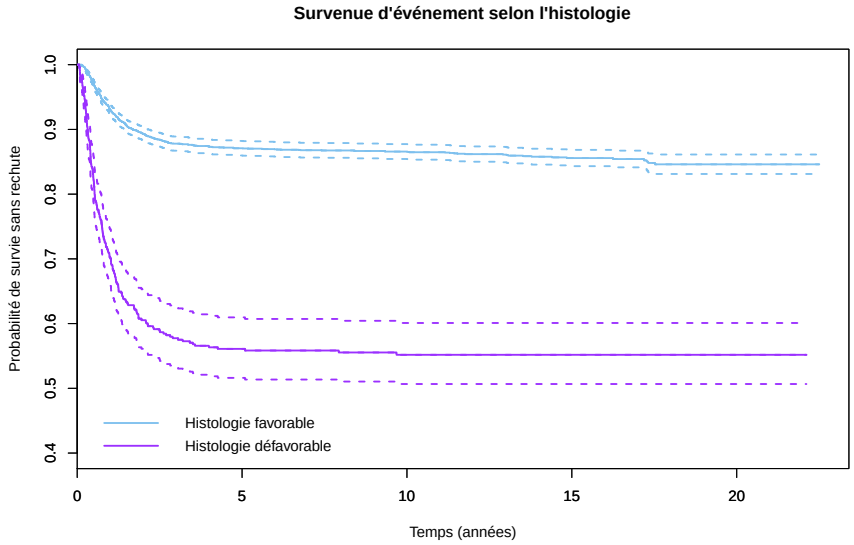


FIGURE 6.3 – Probabilité de survie sans rechute et son intervalle de confiance à 95% en fonction de l'évaluation histologique centrale de la tumeur

L'information de base incluait les variables :

- Stade : Stade de la maladie, de I à IV

- Âge : Âge au diagnostic, en années

- Diamètre : Diamètre de la tumeur, en cm

Et deux variables binaires concernant l'évaluation histologique de la tumeur, l'une réalisée localement dans les hôpitaux, l'autre réalisée dans le laboratoire de référence :

- HL : Évaluation histologique locale (favorable versus défavorable)

- HC : Évaluation histologique centrale (favorable versus défavorable)

La première était fortement prédictive de la deuxième (spécificité 98%, sensibilité 74%).

L'objectif de cette application était d'analyser des enquêtes cas-cohorte simulées à partir de la cohorte NWTS, afin de comparer les estimations obtenues par imputation multiple à celles fournies par les analyses pondérées réalisées par Breslow *et al.* (2009b).

Suivant la démarche de Breslow *et al.* (2009b), nous avons considéré l'évaluation histologique centrale comme la variable de phase-2 et nous avons défini 16 strates selon le statut du sujet (événement versus non-événement), le stade (I/II ou III/IV), l'histologie locale favorable (oui ou non) et l'indicatrice d'âge < 1 an. Pour les données cas-cohorte simulées, les non-cas n'ont été échantillonnés que dans les trois strates définies par les combinaisons 1) histologie locale favorable, indicatrice d'âge < 1 an et stade I/II (n=120), 2) histologie locale favorable, age ≥ 1 an et stade I/II (n=160), 3) histologie locale favorable, age ≥ 1 an et stade III/IV (n=120). Tous les sujets appartenant aux autres strates ont été inclus (Tableau 6.13).

Tableau 6.13 – Stratification et échantillonnage stratifié des données NWTS.

		HL favorable				HL défavorable			
	Totaux	Stade I-II		Stade III-IV		Stade I-II		Stade III-IV	
		âge<1	âge≥1	âge<1	âge≥1	âge<1	âge≥1	âge<1	âge≥1
Cas	669	57	232	10	208	15	41	29	77
Non-cas	3246	452[1]	1620[1]	40	914[1]	12	107	2	99
% événement	17,1	11,2	12,5	20,0	18,5	55,5	27,7	93,5	43,8
Cas	669	57	232	10	208	15	41	29	77
Non-cas	660	120	160	40	120	12	107	2	99

Tableau adapté de Breslow *et al.* (2009b).
[1] Strates dont les enfants ont été échantillonnés.

Breslow *et al.* (2009b) ont modélisé la survenue d'événement à l'aide du modèle de Cox utilisé auparavant par Kulich et Lin (2004) qui avaient déjà analysé ces données :

$$\begin{aligned}\lambda(t;Z) &= \lambda_0(t)exp(\beta_0 + \beta_1 \text{HC} + \beta_2 \hat{\text{A}}\text{ge0} + \beta_3 \hat{\text{A}}\text{ge1} + \beta_4 \text{Stade34} + \beta_5 \text{Diamètre} \\ &+ \beta_6 \text{Stade34*Diamètre} + \beta_7 \text{HC*}\hat{\text{A}}\text{ge0} + \beta_8 \text{HC*}\hat{\text{A}}\text{ge1}) \end{aligned} \quad (6.5)$$

où l'effet de l'âge était estimé moyennant une fonction linéaire par morceaux avec un point d'inflexion à 1 an, c-à-d, Âge0 correspondait à l'âge si l'âge était < 1 an et valait 1 sinon, et Âge1 correspondait à l'âge-1 si l'âge était $\geq$ 1 an et valait 0 sinon. Stade34 correspondait à l'indicatrice de stade III/IV versus I/II.

Breslow *et al.* (2009b) ont utilisé l'estimateur pondéré classique, la calibration et la re-estimation des poids pour étudier la survenue d'événement. La calibration nécessite un modèle de prédiction pour la variable de phase-2, qu'ils ont construit suivant Kulich et Lin (2004) :

$$logit(P[\text{HC=1}]) = \alpha_0 + \alpha_1 \text{HL} + \alpha_2 \text{Stade4} + \alpha_3 \hat{\text{A}}\text{ge10} + \alpha_4 \text{Diamètre} + \alpha_5 \text{HL*Stade4} + e \quad (6.6)$$

où Stade4 correspondait à l'indicatrice de stade IV versus I-III et Âge10 à l'indicatrice d'âge > 10 ans.

En revanche, nous avons construit le premier modèle d'imputation (6.7) selon les règles que nous avons définies pour obtenir un modèle d'imputation parcimonieux, c-à-d, en prenant en compte l'indicatrice de statut et les variables du modèle d'analyse, car les variables de stratification étaient comprises dans ce dernier :

$$\begin{aligned} logit(P[\text{HC=1}]) &= \alpha_0 + \alpha_1 \text{HL} + \alpha_2 \text{Âge0} + \alpha_3 \text{Âge1} + \alpha_4 \text{Stade34} + \alpha_5 \text{Diamètre} \\ &+ \alpha_6 \text{Stade34*Diamètre} + \alpha_7 \text{HL*Âge0} + \alpha_8 \text{HL*Âge1} + \alpha_9 \text{Statut} + e \end{aligned} \quad (6.7)$$

Ces données présentaient deux particularités : d'une part, l'observation d'une variable de phase-1 (évaluation histologique locale) fortement prédictive de la variable de phase-2 ; d'autre part, l'existence d'une interaction entre la variable de phase-2 et une variable de phase-1 (âge) dans le modèle d'intérêt.

Parmi les données réelles, dans les strates dont les non-cas ont été échantillonnés (tous avec histologie locale favorable) très peu de non-cas présentaient une histologie centrale défavorable. Donc, pour certaines sous-cohortes, le modèle d'imputation ne pouvait pas être estimé dans ces strates à cause des odds ratio observés infinis. Par ailleurs, quand les sous-cohortes incluaient quelques non-cas présentant une histologie centrale défavorable dans ces strates, ce nombre était très petit et l'estimateur du modèle d'imputation risquait de ne pas être distribué selon les conditions asymptotiques.

À cause de ces spécificités, nous avons considéré deux autres modèles d'imputation : L'un (modèle 6.8), correspondait au modèle de prédiction utilisé par Breslow *et al.* (2009b) dans la calibration plus l'indicatrice de cas

$$logit[P(\text{HC=1})] = \alpha_0 + \alpha_1 \text{Statut} + \alpha_2 \text{HL} + \alpha_3 \text{Stade4} + \alpha_4 \text{Age10} + \alpha_5 \text{Diamètre} + \alpha_6 \text{HL*Stade4} + e \quad (6.8)$$

L'autre (modèle 6.9), incluait seulement l'indicatrice de cas et l'histologie locale, puisque cette dernière variable était fortement prédictive de l'histologie centrale :

$$logit[P(\text{HC=1})] = \alpha_0 + \alpha_1 \text{Statut} + \alpha_2 \text{HL} + e \quad (6.9)$$

Résultats

Les résultats sont donnés au tableau 6.14. Les analyses pondérées fournissaient toujours des estimations en accord avec la cohorte complète. L'estimateur pondéré classique fournissait, comme attendu, des écart-types plus larges que l'analyse de la cohorte complète. Les approches pondérées par calibration ou ré-estimation améliioraient la précision par rapport à l'approche classique, sans atteindre toutefois la précision des estimations fournies par la cohorte entière.

Pour le modèle d'imputation 6.7 (modèle 1), l'imputation multiple fournissait des estimations différentes de la cohorte complète pour les effets impliquant l'histologie : effet principal de l'histologie centrale, qui représentait l'écart attendu à 1 an entre deux enfants avec une histologie centrale favorable ou défavorable différait de l'effet estimé dans la cohorte complète (respectivement, 3,72 versus 4,04) ; différence entre

les effets de l'âge < 1 an selon l'histologie (respectivement, -2,63 versus -2,33), effet de l'âge > 1 an selon l'histologie (respectivement, -0,06 versus -0,03). Les écart-types des variables de phase-1 étaient proches de ceux obtenus avec la cohorte complète.

Avec les modèles 6.8 (modèle 2) et 6.9 (modèle 3), l'imputation multiple fournissait également des résultats qui différaient de ceux obtenus avec la cohorte entière pour l'estimation des effets impliquant la variable de phase-2.

Tableau 6.14 – NWTS : Resultats moyens des 1000 échantillons cas-cohorte simulés.

	Cohorte[1]	Imputation Multiple			Estimateurs pondérés		
	entière	Modèle 1[2]	Modèle 2[3]	Modèle 3[4]	Standard	Calibré	Re-estimé
	β	β	β	β	β	β	β
	(ET)	(ET[5])	(ET[5])	(ET[5])	(ET[5])	(ET[5])	(ET[5])
HC	4,042	3,717	3,596	4,106	4,046	4,046	4,050
	(0,413)	(0,493)	(0,555)	(0,469)	(0,537)	(0,520)	(0,518)
Âge0	-0,661	-0,669	-0,702	-0,537	-0,687	-0,663	-0,676
	(0,326)	(0,327)	(0,329)	(0,329)	(0,359)	(0,324)	(0,324)
Âge1	0,104	0,106	0,102	0,106	0,102	0,104	0,107
	(0,017)	(0,017)	(0,017)	(0,016)	(0,026)	(0,017)	(0,017)
Stade	1,346	1,344	1,441	1,353	1,431	1,345	1,344
	(0,244)	(0,252)	(0,257)	(0,251)	(0,346)	(0,273)	(0,272)
Diamètre	0,069	0,072	0,073	0,072	0,070	0,070	0,070
	(0,014)	(0,014)	(0,014)	(0,014)	(0,021)	(0,015)	(0,015)
Std*Diam	-0,076	-0,083	-0,082	-0,083	-0,076	-0,076	-0,076
	(0,019)	(0,020)	(0,020)	(0,020)	(0,029)	(0,021)	(0,021)
HC*age0	-2,635	-2,334	-2,239	-3,097	-2,648	-2,655	-2,651
	(0,464)	(0,543)	(0,611)	(0,525)	(0,612)	(0,592)	(0,590)
HC*age1	-0,058	-0,033	-0,041	-0,065	-0,051	-0,050	-0,052
	(0,034)	(0,038)	(0,041)	(0,040)	(0,051)	(0,050)	(0,048)

[1] Estimation unique fournie par la cohorte NWTS ; [2] Modèle d'imputation (6.7) ;
[3] Modèle d'imputation (6.8) ; [4] Modèle d'imputation (6.9) ; [5] Ecart-type (ET) des estimations ;
HC, indicatrice d'histologie centrale défavorable ; Age0 et Age1, termes linéaires de l'âge
au diagnostic (années) avant et après 1 an ; stade, indicatrice de stade III/IV ;
diamètre (cm) de la tumeur ; Std*Diam, interactions Stade*diamètre ; ET, écart-type.

6.4 Discussion

Les comparaisons de méthodes d'analyse des enquêtes cas-cohorte à partir des données entièrement simulées nous ont permis d'apprécier l'existence de biais éventuels pour les scénarios considérés. Comme attendu, l'estimateur du maximum de vraisemblance partielle sur la cohorte entière et l'estimateur pondéré classique étaient sans biais. C'était aussi le cas de l'estimateur de l'imputation multiple sous H_0 quel que soit l'échantillonnage et en présence des risques relatifs égaux à 2 avec un échantillonnage simple de la sous-cohorte. De légers biais (-2 à -5%) ont cependant été observés avec un échantillonnage stratifié. Comme prévu, l'estimateur de l'imputation multiple des effets des variables de phase-1 était plus précis que l'estimateur pondéré classique, et presque aussi précis que l'analyse de la cohorte entière. Dans les douze scénarios considérés, l'estimation de l'effet de la variable de phase-2 était également plus précise avec l'imputation multiple qu'avec l'analyse pondérée.

Globalement, l'estimateur de l'imputation multiple est apparu satisfaisant mais on peut se demander dans quelle mesure ces résultats obtenus avec des modèles d'imputation en accord avec le processus de génération des données et un petit nombre de variables sont généralisables à d'autres situations. D'autres simulations ont été réalisées pour répondre à cette question.

Les simulations d'enquêtes cas-cohorte à partir de la cohorte PRIME impliquaient sept variables explicatives et le processus de génération des données était évidemment inconnu. Il faut noter que nous ne disposions pas d'une variable fortement prédictive de la variable de phase-2 et que l'essentiel des gains attendus de l'imputation multiple concernait la précision des effets des variables de phase-1. L'imputa-

tion multiple a été comparée à l'analyse de la cohorte entière, à l'estimateur pondéré classique et à deux estimateurs calibrés où la calibration portait soit sur les résidus *dfbeta* de la variable de phase-2, soit sur les résidus *dfbeta* de toutes les variables du modèle d'analyse. Pour les deux tailles de sous-cohorte considérées, les cinq méthodes fournissaient des estimateurs centrés sur les mêmes valeurs. En ce qui concerne l'estimation de l'effet des variables de phase-1, l'estimateur calibré sur les résidus de toutes les variables était plus précis que celui calibré sur la seule variable de phase-2 ou que l'estimateur pondéré classique, mais un peu moins précis que l'estimateur de l'imputation multiple, qui atteignait presque la précision fournie par la cohorte entière. En ce qui concerne l'estimation de l'effet de la variable de phase-2, l'estimateur de l'imputation multiple était aussi précis que l'estimateur pondéré calibré sur toutes les variables et plus précis que les autres estimateurs pondérés dans les simulations avec 700 sujets par sous-cohorte, mais dans les simulations avec 2100 sujets par sous-cohorte c'était l'estimateur le moins précis (efficacité relative par rapport à la cohorte entière 0,81 versus 0,84).

Les simulations réalisées à partir de la cohorte NWTS concernaient cinq variables et trois termes d'interactions dans le modèle d'analyse impliquant pour deux d'entre elles les variables de phase-2. Le modèle de génération des données était inconnu. Pour l'effet des variables de phase-1, les différentes approches comparées fournissaient des estimateurs en concordance avec l'analyse de la cohorte entière. Leurs comparaisons en termes de précision conduisaient à des conclusions similaires à celles des simulations précédentes. Ce n'était pas le cas des résultats concernant les trois effets qui impliquaient la variable de phase-2. Des différences de l'ordre de 10% étaient observées entre l'imputation multiple et les autres méthodes, en particulier, l'analyse de la cohorte entière. Cela s'explique par la concordance élevée entre

histologie centrale et histologie locale, et par la présence dans le modèle d'analyse d'interactions impliquant la variable de phase-2, ce qui nous a conduit à inclure des interactions dans le modèle d'imputation, et à estimer des modèles d'imputation spécifiques à différentes strates de la cohorte. Par conséquent, certains paramètres ont été estimés dans des sous-échantillons avec un petit nombre de cas, d'où l'estimateur de maximum de vraisemblance n'était pas distribué selon les conditions asymptotiques requises. La mise en œuvre de l'imputation multiple avec un modèle n'incluant que l'histologie locale et le statut n'a pas permis de résoudre ce problème car l'estimateur utilisé imputait des histologies centrales reflétant l'effet marginal de cette variable sur la survenue d'événement et non sur l'effet ajusté sur l'âge, le stade et le diamètre de la tumeur. Les imputations réalisées avec un modèle correspondant au modèle de prédiction utilisé par Breslow *et al.* (2009b) dans sa mise en œuvre de la calibration n'a pas permis non plus d'améliorer les résultats.

Nous avons étudié la robustesse de l'imputation multiple réalisée selon un modèle linéaire gaussien quand la variable de phase-2 suivait une distribution non gaussienne. Quand cette distribution n'était pas symétrique (log-normale versus normale), un biais négatif d'environ 14% était observé sur l'estimation de l'effet correspondant. Lorsque la vraie distribution était symétrique (uniforme ou Student à 5 degrés de liberté), les biais observés étaient plus faibles (5% avec une distribution uniforme et 7,5% avec une distribution de Student à 5 degrés de liberté).

Nous avons vérifié que l'estimation du modèle d'imputation sur les seuls cas, qu'il comprenne ou non le délai de censure conduisait à des estimations sensiblement moins bonnes que son estimation sur l'échantillon cas-cohorte.

A notre connaissance, c'est la première fois que la capacité prédictive d'un modèle ou d'une variable additionnelle a été estimée dans le cadre d'une enquête cas-cohorte. Nos simulations ont montré que l'utilisation naïve des estimateurs proposés pour les cohortes dans les échantillons cas-cohorte fournissait des estimations inférieures à celle de la cohorte entière. Au contraire, l'imputation multiple et la cohorte entière fournissaient, comme attendu, des estimations analogues. La mauvaise performance de l'estimateur naïf peut s'expliquer par le taux d'échantillonnage différent des couples (non-cas, non-cas), (cas, non-cas), (cas, cas) dans l'échantillon cas-cohorte, en particulier les paires (cas, cas) sont toutes conservées alors qu'elles sont plus souvent discordantes que les paires (cas, non-cas), si la prédiction est bonne. Cependant, il faut noter que la capacité prédictive est toujours surestimée car nous utilisons les mêmes données pour estimer le modèle d'analyse et sa capacité prédictive.

Il est évidemment possible d'estimer l'indice C par une approche pondérée. Cela suppose de calculer la probabilité d'observation de chaque paire de sujets dans l'échantillon cas-cohorte, de taille N', donc de calculer $N'(N'-1)$ probabilités. Ce n'est pas une contrainte bien lourde puisque ces probabilités sont utilisées classiquement dans les calculs de la variances des estimateurs pondérés. En revanche, pour obtenir la variance de l'estimateur pondéré de C, c'est la probabilité d'observation de chaque quadruplet de sujets qui est nécessaire d'où la nécessité de calculer $N'(N'-1)(N'-2)(N'-3)$ probabilités.

Les simulations ont montré que les écart-types estimés du C de Harrell et de l'IDI étaient très différents des écart-types observés dans les cohortes entières comme dans les enquêtes cas-cohorte. Nous n'avons pas étudié ce problème et nous igno-

rons s'il provient de la fonction que nous avons utilisée (*rcorr.cens*, bibliothèque *Hmisc*) ou d'une autre cause.

Plusieurs recommandations peuvent être formulées pour mettre en œuvre l'analyse des enquêtes cas-cohorte par imputation multiple dans les meilleures conditions :

- utiliser les outils diagnostiques pour s'assurer que la distribution observée de la variable de phase-2 est en accord avec les hypothèses du modèle d'imputation,
- adapter au besoin le modèle d'imputation à cette distribution en modifiant les hypothèses du modèle d'imputation ou en ajustant le modèle d'imputation à une transformation des valeurs observées de la variable de phase-2 (log, racine,...),
- construire le modèle d'analyse en utilisant un estimateur pondéré classique puis améliorer la précision des estimations en utilisant l'imputation multiple tout en s'assurant que les résultats sont analogues à ceux de l'analyse pondérée.

CHAPITRE

7 Application : Étude des Trois Cités

L'étude des Trois Cités (Étude 3C, http ://www.three-city-study.com) a pour objectif principal de mesurer la relation entre les facteurs de risque vasculaires et la démence dans une population de 9294 personnes (3650 hommes et 5644 femmes) âgées de 65 ans et plus, recrutées entre 1999 et 2001 dans 3 villes en France : 2104 à Bordeaux, 4931 à Dijon et 2259 à Montpellier.

Une sous-étude cas-cohorte, menée à partir de la cohorte 3C, avait été réalisée à la fin des 4 ans de suivi, dans l'intention d'étudier la relation entre des facteurs de risque non-standard et le risque d'événement cardiovasculaire (CHD) ou de démence vasculaire (VaD). La sous-cohorte, de taille n=1254 (13,5% de la cohorte), avait été sélectionnée par un tirage au sort stratifié sur l'âge, le sexe et le centre de recrutement.

Notre but était d'analyser une vraie enquête cas-cohorte par imputation multiple et d'estimer la capacité prédictive des variables de phase-2. En particulier, cette application portait, d'une part, sur l'étude de la relation entre les niveaux de D-dimère et les risques d'événement cardiovasculaire ou de démence vasculaire ; d'autre part, sur l'étude de la relation entre les apolipopotéines A-I (ApoA) et B-100 (ApoB) et le risque d'événement cardiovasculaire.

Pour l'étude du risque cardiovasculaire, 131 cas prévalents et 37 sujets sans information à propos du statut incident de l'événement avaient été exclus. Ainsi, la sous-cohorte comptait 1086 sujets, dont 33 cas incidents de maladie cardiovasculaire. En outre, 166 sujets avaient développé un premier événement cardiovasculaire pendant la période de suivi de 4 ans. Concernant la démence vasculaire, 29 cas prévalents et 109 sans information sur l'incidence de l'événement avaient été exclus. Ainsi, la sous-cohorte comptait 1116 sujets. Cinquante deux sujets avaient présenté un premier événement de démence vasculaire pendant la période de suivi de 4 ans.

Les variables de phase-1 incluaient entre autres : l'âge, le sexe, le centre de recrutement (Montpellier, Bordeaux, Dijon), le niveau d'études (3 niveaux : primaire, secondaire, universitaire), la consommation de tabac (3 niveaux : non-fumeurs, moins de 10 paquets/année, 10 ou plus paquets/année), la tension artérielle, la taille, le poids. Des paramètres biologiques telles que la glycémie, le cholestérol total, le cholestérol HDL et les triglycérides étaient aussi recueillis. Ils ont été mesurés de manière centralisée au centre hospitalier universitaire de Dijon. Les concentrations des LDL étaient estimées par la formule de Friedewald *et al.* (1972) pour les sujets dont la concentration de triglycérides était inférieure à 5,08mmol/L. L'information sur la présence de allèle ϵ 4 l'apolipoprotéine E était aussi disponible. L'hyperten-

sion (HTA) était définie comme une pression artérielle systolique supérieure à 140 mmHg ou une pression artérielle diastolique supérieur à 90 mmHg ou la prise d'un traitement antihypertenseur. Le diabète comme une valeur de glycémie supérieur ou égale à 6,1 mmol/L et l'hypercholestérolemie (Hchol) comme un taux de cholestérol supérieur ou égal à 6,2 mmol/L. L'indice de masse corporelle (IMC) était calculé comme le rapport entre le poids (kg) et la taille (m) au carré. L'obésité était définie comme un IMC > 27.

7.1 D-dimère et risque d'événement cardiovasculaire et démence vasculaire

Les niveaux de D-dimère, un marqueur à la fois de coagulation et de fibrinolyse, étaient disponibles pour tous les cas et pour les non-cas de la sous-cohorte. Les incidences cumulées observées de CHD et VaD étaient respectivement d'environ 2% et 0,6%. Carcaillon *et al.* (2009) avaient auparavant étudié la relation entre les niveaux de D-dimère et le risque d'événement cardiovasculaire et de démence vasculaire en utilisant les quintiles des niveaux de D-dimère dans leurs analyses. Ils avaient rapporté une augmentation linéaire du risque de démence vasculaire en fonction des quintiles de D-dimère.

Nous avons re-analysé les mêmes données par imputation multiple afin d'estimer les risques relatifs associés aux niveaux de D-dimère, puis nous avons estimé la capacité prédictive des niveaux de D-dimère, vis-à-vis des deux risques. Nous avons utilisé les mêmes modèles d'analyse que Carcaillon *et al.* (2009). En revanche, à cause du petit nombre d'événements, nous n'avons pas utilisé les quintiles mais les tertiles de D-dimère.

Pour l'estimation du risque cardiovasculaire, le modèle de Cox était :

$$\begin{aligned}\lambda(t;Z) &= \lambda_0(t)exp(\beta_0 + \beta_1\text{Sexe} + \beta_2\text{Âge} + \beta_3\text{Bordeaux} + \beta_4\text{Dijon} + \beta_5\text{Obésité} + \beta_6\text{HTA} \\ &+ \beta_7\text{Hchol} + \beta_8\text{Diabète} + \beta_9\text{Tabac1} + \beta_{10}\text{Tabac2} + \beta_{11}\text{TraitDiabète} \\ &+ \beta_{12}\text{D-dimèreT2} + \beta_{13}\text{D-dimèreT3}) \qquad (7.1)\end{aligned}$$

où Bordeaux et Dijon sont les indicatrices d'un recrutement dans ces centre, HTA l'indicatrice d'hypertension, Hchol l'indicatrice d'hypercholestérolemie, TraitDiabète l'indicatrice d'un traitement anti-diabète, Tabac1 (respectivement Tabac2) l'indicatrice de moins de 10 paquets/année (respectivement 10 ou plus paquets/année), et D-dimèreT2 (respectivement T3) est l'indicatrice correspondant au $2^{ème}$ tertile (respectivement $3^{ème}$ tertile) des niveaux de D-dimère.

Pour le risque de démence vasculaire, le modèle de Cox était :

$$\begin{aligned}\lambda(t;Z) &= \lambda_0(t)exp(\beta_0 + \beta_1\text{Sexe} + \beta_2\text{Âge} + \beta_3\text{Bordeaux} + \beta_4\text{Dijon} + \beta_5\text{Obésité} + \beta_6\text{Secondaire} \\ &+ \beta_7\text{Universitaire} + \beta_8\text{Apoe4} + \beta_9\text{D-dimèreT2} + \beta_{10}\text{D-dimèreT3}) \qquad (7.2)\end{aligned}$$

où Secondaire (respectivement Universitaire) est l'indicatrice du niveau d'études secondaire (respectivement universitaire), Apoe4 la présence de l'allèle apolipoproteine e4 et D-dimèreT2 (respectivement D-dimèreT3) l'indicatrice correspondant au $2^{ème}$ tertile (respectivement $3^{ème}$ tertile) des niveaux de D-dimère.

Pour chacun des risques, il était nécessaire de refléter les relations entre la variable incomplète, la survenue de l'événement et les variables de confusion. Nous avons donc construit un modèle d'imputation différent pour chaque risque (CHD et VaD) incluant les variables du modèle de Cox les concernant plus l'indicatrice de cas. Ainsi, le modèle d'imputation des tertiles de D-dimère (D-dimèreT) en relation avec le risque d'événement cardiovasculaire était modélisé à l'aide d'une régression polytomique :

$$\begin{aligned}
log\frac{P[\text{D-dimèreT}=2]}{P[\text{D-dimèreT}=1]} &= \alpha_0^2 + \alpha_1^2\text{Statut} + \alpha_2^2\text{Âge} + \alpha_3^2\text{Sexe} + \alpha_4^2\text{Bordeaux} + \alpha_5^2\text{Dijon} + \alpha_6^2\text{Obésité} \\
&+ \alpha_7^2\text{HTA} + \alpha_8^2\text{HChol} + \alpha_9^2\text{Diabète} + \alpha_{10}^2\text{TraitDiabète} + \alpha_{11}^2\text{Tabac1} \\
&+ \alpha_{12}^2\text{Tabac2} + e \\
log\frac{P[\text{D-dimèreT}=3]}{P[\text{D-dimèreT}=1]} &= \alpha_0^3 + \alpha_1^3\text{Statut} + \alpha_2^3\text{Âge} + \alpha_3^3\text{Sexe} + \alpha_4^3\text{Bordeaux} + \alpha_5^3\text{Dijon} + \alpha_6^3\text{Obésité} \\
&+ \alpha_7^3\text{HTA} + \alpha_8^3\text{HChol} + \alpha_9^3\text{Diabète} + \alpha_{10}^3\text{TraitDiabète} + \alpha_{11}^3\text{Tabac1} \\
&+ \alpha_{12}^3\text{Tabac2} + e
\end{aligned} \tag{7.3}$$

Le modèle d'imputation concernant la démence vasculaire était estimé à l'aide de la régression polytomique :

$$\begin{aligned}
log\frac{P[\text{D-dimèreT}=2]}{P[\text{D-dimèreT}=1]} &= \alpha_0^2 + \alpha_1^2\text{Statut} + \alpha_2^2\text{Âge} + \alpha_3^2\text{Sexe} + \alpha_4^2\text{Bordeaux} + \alpha_5^2\text{Dijon} + \alpha_6^2\text{Obésité} \\
&+ \beta_7^2\text{Secondaire} + \beta_8^2\text{Universitaire} + \alpha_9^2\text{Apoe4} + e \\
log\frac{P[\text{D-dimèreT}=3]}{P[\text{D-dimèreT}=1]} &= \alpha_0^3 + \alpha_1^3\text{Statut} + \alpha_2^3\text{Âge} + \alpha_3^3\text{Sexe} + \alpha_4^3\text{Bordeaux} + \alpha_5^3\text{Dijon} + \alpha_6^3\text{Obésité} \\
&+ \beta_7^3\text{Secondaire} + \beta_8^3\text{Universitaire} + \alpha_9^3\text{Apoe4} + e
\end{aligned} \tag{7.4}$$

7.1 D-dimère et risque d'événement cardiovasculaire et démence vasculaire95

Les risques relatifs associés aux tertiles des D-dimère, estimés par imputation multiple et par analyse pondérée, sont donnés au tableau 7.1 ainsi que leurs intervalles de confiance. Les estimations et les précisions fournies par les deux méthodes, pour la variable de phase-2, étaient similaires. L'intervalle de confiance du risque relatif associé à l'effet linéaire d'une différence d'un tertile pour CHD était (0,94-1,38) avec imputation multiple, (0,92-1,38) avec analyse pondérée, et pour VaD (1,13-2,53) avec imputation multiple et (1,13-2,67) avec analyse pondérée. Concernant les variables de phase-1, l'imputation multiple fournissait toujours des estimations plus précises que l'analyse pondérée (tableau 7.2).

Tableau 7.1 – Estimation des risques relatifs (RR) et intervalles de confiance à 95% (IC 95%) associés aux tertiles de D-dimère.

	Imputation multiple	Analyse pondérée
	RR (95% CI)	RR (95% CI)
Risque d'événement cardiovasculaire et D-Dimère[a]		
T1	1,00 (référence)	1,00 (référence)
T2	1,42 (0,99 - 2,04)	1,40 (0,97 - 2,04)
T3	1,32 (0,89 - 1,97)	1,30 (0,84 - 1,99)
Tendance linéaire	1,14 (0,94 - 1,38)	1,13 (0,92 - 1,38)
Risque de démence vasculaire et D-Dimère[b]		
T1	1,00 (référence)	1,00 (référence)
T2	1,57 (0,63 - 3,93)	1,60 (0,63 - 4,09)
T3	2,77 (1,17 - 6,57)	2,93 (1,22 - 7,06)
Tendance linéaire	1,69 (1,13 - 2,53)	1,74 (1,13 - 2,67)

T1, tertile 1 ; T2, tertile 2 ; T3, tertile 3.
[a] Modèle 7.1
[b] Modèle 7.2

Tableau 7.2 – Estimation des risques relatifs (RR) et intervalles de confiance à 95% (IC 95%) associés aux variables de phase-1.

	Imputation multiple		Analyse pondérée	
	RR	(95% CI)	RR	(95% CI)
Risque d'événement cardiovasculaire et D-Dimère[a]				
Âge	1,03	(1,00-1,06)	1,03	(1,00-1,06)
Sexe	0,36	(0,26- 0,50)	0,38	(0,26-0,53)
Montpellier	1		1	
Bordeaux	0,89	(0,63-1,26)	0,96	(0,66-1,41)
Dijon	1,04	(0,70-1,53)	1,17	(0,77-1,80)
Obésité	0,86	(0,63-1,16)	0,99	(0,71-1,37)
HTA	1,45	(0,96-2,19)	1,53	(0,99-2,35)
Hchol	1,12	(0,84-1,50)	1,25	(0,92-1,71)
Diabète	2,05	(1,13-3,73)	2,61	(1,37-4,97)
TraitDiab	1,44	(0,74-2,78)	1,22	(0,60-2,46)
Non-fumeurs	1		1	
Tabac <10	0,58	(0,36-0,94)	0,62	(0,37-1,03)
Tabac $\geq$10	0,77	(0,54-1,10)	0,82	(0,56-1,19)
Risque de démence vasculaire et D-Dimère[b]				
Âge	0,64	(0,55-0,74)	0,64	(0,58-0,71)
Sexe	0,50	(0,29-0,87)	0,59	(0,32-1,06)
Montpellier	1		1	
Bordeaux	1,01	(0,51-2,01)	0,97	(0,45-2,12)
Dijon	1,28	(0,58-2,85)	1,42	(0,57-3,53)
Obesité	0,64	(0,34-1,21)	0,50	(0,25-1,01)
Primaire	1		1	
Secondaire	0,49	(0,24-1,00)	0,45	(0,21-0,93)
Universitaire	0,47	(0,24-0,90)	0,49	(0,25-0,97)
apoE4	2,49	(1,41-4,40)	2,23	(1,23-4,07)

[a] Modèle 7.1
[b] Modèle 7.2

En ce qui concerne la capacité prédictive, l'indice C de Harrell du modèle incluant uniquement des variables de phase-1 était 0,69 pour le risque d'événement cardiovasculaire et 0,86 pour le risque de démence vasculaire (Tableau 7.3). Donc, les deux risques étaient bien prédits par les facteurs de risque classiques. L'inclusion du D-dimère n'améliorait pas de façon sensible la capacité prédictive du modèle en dépit du fait qu'un niveau élevé de D-dimère augmentait significativement le risque de démence vasculaire.

Tableau 7.3 – Capacité prédictive et intervalle de confiance à 95% (IC 95%) des tertiles de D-dimère sur le risque d'événement cardiovasculaire et sur le risque de démence vasculaire.

	Événement cardiovasculaire		Démence vasculaire	
	Estimation	IC 95%	Estimation	IC 95%
C_1	0,693	(0,622 - 0,764)	0,865	(0,787 - 0,943)
C_2	0,694	(0,621 - 0,767)	0,874	(0,798 - 0,950)
Δ	0,002	(-0,004 - 0,008)	0,009	(-0,011 - 0,029)
NRI	0,009	(-0,049 - 0,066)	-	-
IDI	0,001	(-0,001 - 0,003)	0,0004	(-0,0002 - 0,0010)

C_1, C de Harrell du modèle de Cox sans les tertiles de D-dimère.
C_2, C de Harrell du modèle de Cox avec les tertiles de D-dimère.
Δ, Capacité prédictive des tertiles de D-dimère.
NRI, *net reclassification improvement* apporté par les tertiles de D-dimère.
IDI, *integrated discrimination index* apporté par les tertiles de D-dimère.

7.2 Apolipoprotéines A-I (ApoA) et B-100 (ApoB) et risque d'événement cardio-vasculaire

Les apolipoprotéines représentent la partie protéique des particules qui contiennent et transportent les lipides dans le sang. Plus précisément, l'apolipoprotéine AI (apoA) est principalement située dans les particules contenant du cholestérol HDL. L'apolipoprotéine B100 (apoB) se trouve dans les particules contenant le cholestérol LDL. Lorsque le taux d'apoA diminue et que le taux d'apoB augmente, le risque cardio-vasculaire augmente.

Nous avons analysé la relation entre les apolipoprotéines apoA et apoB et le risque d'événement cardiovasculaire. Puis, nous avons évalué si l'inclusion de ces facteurs de risque dits non-standard permettaient d'améliorer la prédiction du risque cardio-vasculaire.

Nous avons estimé par imputation multiple les risques relatifs d'événement cardiovasculaire associés à chacun des paramètres lipidiques apoA, apoB, rapport apoB/apoA, cholestérol LDL et HDL, à l'aide d'un modèle de Cox ajusté sur l'âge, le sexe, le centre, l'indice de masse corporelle, le diabète, la tension artérielle systolique, le traitement de l'hypertension et le traitement du cholestérol. Nous avons utilisé le même modèle d'imputation pour les apolipoprotéines apoA et apoB, un modèle incluant les variables d'ajustement du modèle de Cox et l'indicatrice de cas. Les résultats des estimations des risques relatifs sont donnés au tableau 7.4. Après ajustement sur les facteurs de risque classiques, les paramètres considérés étaient significativement associés au risque cardio-vasculaire.

Nous avons évalué si cette relation significative avec le risque d'événement cardio-vasculaire était associée à une amélioration sensible de la prédiction. En particulier, nous avons évalué si l'addition du cholestérol LDL et HDL ou des apolipoproteines apoA et apoB, ou le rapport d'apoA et apoB améliorait la capacité prédictive d'un modèle de base (M_0) qui incluait l'âge, le sexe, le centre, l'indice de masse corporelle, le diabète, la pression artérielle systolique, le traitement anti-hypertensif et le traitement du diabète.

Le risque d'événement cardiovasculaire était bien expliqué par les variables de phase-1 introduites au modèle de base (C=0,708). Cependant, l'inclusion des facteurs de risque non-standards (cholestérol LDL et HDL ou apolipoproteines apoA et apoB, ou le rapport d'apoA et apoB) permettait une amélioration sensible de la prédiction du risque (tableau 7.5).

Tableau 7.4 – Estimation par imputation multiple des risques relatifs (RR) et intervalles de confiance à 95% (IC 95%) associés à différents marqueurs lipidiques.

		RR	(95% CI)
M_0 +	ApoA	0,47	(0,26-0,86)
	ApoB	5,56	(2,92-10,62)
M_0 +	ApoB/ApoA	2,62	(1,81-3,77)
M_0 +	cholestérol LDL	1,47	(1,26-1,72)
	cholestérol HDL	0,61	(0,40-0,94)

M_0 Ajusté sur âge, sexe, centre, indice de masse corporelle,diabète, tension artérielle systolique, traitement hypertension et traitement cholestérol.

Tableau 7.5 – Estimation de la capacité prédictive (C de Harrel, Δ) et NRI.

	C	Δ	IC 95%	% NRI	IC 95%
M_0	0,708	-	-	-	
M_0 + ApoA et ApoB	0,733	0,025	0,003-0,047	8,6	1,8-15,3
M_0 + ApoB/ApoA	0,728	0,020	0,008-0,032	5,9	0,4-11,4
M_0 + cholestérol LDL et HDL	0,733	0,024	0,004-0,044	7,7	0,9-14,4

M_0 : âge, sexe, centre, indice de masse corporelle, diabète, tension artérielle systolique, traitement hypertension, traitement cholestérol.

7.3 Discussion

Le modèle d'imputation doit refléter les relations entre la variable incomplète, la survenue de l'événement et les variables de confusion, ce qui nous a conduit à utiliser des modèles d'imputation du D-dimère différents en fonction du risque étudié car l'inclusion de variables non ou peu prédictives peut augmenter la variance inter-imputations. Cependant, nous avons vérifié que l'inclusion des variables utilisées pour modéliser le risque d'événement cardio-vasculaire dans le modèle d'imputation du D-Dimère sur le risque de démence ne modifiait pas les résultats observés avec le modèle d'imputation plus parcimonieux utilisé.

Les résultats des données 3C étaient en accord avec ceux obtenus dans le chapitre 6 (Validation). En ce qui concerne la relation entre les niveaux de D-dimère et les risques d'événement cardio-vasculaire, l'imputation multiple fournissait des estimations proches de celles fournies par l'analyse pondérée (différence relative entre les deux estimations toujours inférieure à 2%). Les estimations obtenues par imputation multiple étaient plus précises que celles obtenues par analyse pondérée pour les variables de phase-1 et légèrement plus précises pour la variable de phase-2. Concernant la relation entre D-dimères et démence vasculaire, les deux méthodes fournissaient, globalement, des estimations du même ordre de grandeur et une précision similaire. Cependant, nous avons observé une estimation légèrement plus petite pour le log du risque relatif comparant le $3^{ème}$ et le 1^{er} tertile (2,77 versus 2,93, c-à-d. une différence relative de 8% entre l'estimation de l'imputation multiple et celle de l'analyse pondérée). Cela s'explique par le petit nombre de cas de démence vasculaire (51) et l'imputation qualitative des tertiles de D-dimères impliquant l'utilisation d'un modèle d'imputation multinomial et l'estimation des paramètres dans des strates

séparées, définies par les tertiles de la concentration de D-dimère, n'incluant parfois qu'un petit nombre d'événement. Il en résultait un estimateur du maximum de vraisemblance qui ne vérifiait pas les conditions asymptotiques.

Le D-dimère n'augmentait pas significativement le risque d'événement cardio-vasculaire, par conséquent, sa capacité prédictive était négligeable. Les résultats coïncidaient avec les trois indices utilisés. En revanche, une concentration élevée de D-dimère ($3^{ème}$ tertile) augmentait le risque de démence vasculaire. Cependant, l'inclusion du D-dimère n'améliorait pas significativement la capacité prédictive du modèle. Tant l'indice C que l'indice IDI conduisaient à la même conclusion. L'indice NRI n'a pas pu été estimé car nous n'avons pas trouvé d'étude concernant la capacité prédictive des D-dimère sur le risque de démence vasculaire ou présentant des niveaux de risque pertinents cliniquement.

Wang *et al.* (2006) et Tzoulaki *et al.* (2007) ont rapporté que l'inclusion de 10 et 4 biomarqueurs respectivement, n'améliorait que légèrement la prédiction du risque cardio-vasculaire fournie par les facteurs de risque classiques.

L'étude de la relation entre les apolipoproteines A-I (ApoA) et B-100 (ApoB) et risque d'événement cardiovasculaire par imputation multiple est issue d'une collaboration avec l'équipe 4 « Épidémiologie cardiovasculaire et mort subite » du PARCC. A notre connaissance, c'est la première fois que la valeur prédictive d'une variable additionnelle est évaluée dans une enquête cas-cohorte.

CHAPITRE

8 Conclusion et perspectives

Le travail réalisé au cours de cette thèse nous a permis de constater que l'imputation multiple est une alternative intéressante pour analyser les enquêtes cas-cohorte. Les méthodes comparées, les estimateurs pondérés classiques, les approches pondérées de Breslow *et al.* et l'imputation multiple, fournissent en général des estimations non biaisées mais peuvent différer en termes de précision.

Ces simulations ont montré que l'imputation multiple fournissait généralement des estimateurs sans biais des risques relatifs. Pour les variables de phase-1, ils approchaient la précision obtenue par l'analyse de la cohorte complète, ils étaient légèrement plus précis que l'estimateur calibré de Breslow *et al.* (2009b) et surtout que les estimateurs pondérés classiques. Pour les variables de phase-2, l'estimateur de l'imputation multiple était généralement sans biais et d'une précision supérieure à celle des estimateurs pondérés classiques et analogue à celle de l'estimateur calibré de Breslow *et al.* (2009b).

En conclusion, l'imputation multiple est une alternative simple aux approches pondérées pour l'analyse des enquêtes cas-cohorte, qui permet d'obtenir des estimations non biaisées des risques relatifs et de leurs variances, tout en améliorant la précision des effets des variables de phase-1 et en fournissant au moins la même précision pour l'effet de la variable de phase-2 que les estimateurs pondérés. Par ailleurs, elle permet d'estimer facilement la capacité prédictive. Plus généralement, n'importe quelle mesure que l'on sait estimer pour une cohorte pourrait être estimé dans les cadre d'une enquête cas-cohorte à l'aide de l'imputation multiple.

Cependant, les nouvelles méthodes ne sont pas toujours facilement adoptées par les utilisateurs, parfois par la complexité de la méthode, parfois par manque d'une implémentation simple vis-à-vis non seulement des statisticiens mais aussi des épidémiologistes. Nous avons rédigé deux fonctions de simple utilisation sur R afin de faciliter l'analyse d'enquêtes cas-cohorte par imputation multiple.

Perspectives

Nous nous sommes limités au cas de données manquantes dues au plan d'échantillonnage, car alors l'hypothèse de données MA est vérifiée et l'utilisation de *mice* est particulièrement pertinente. Mais les données de l'échantillon cas-cohorte peuvent être elles-mêmes incomplètes (non-réponse, refus ou perte de prélèvement,...). L'une des perspectives est d'utiliser l'imputation multiple afin d'exploiter l'intégralité des observations disponibles et de réaliser des analyses sous l'hypothèse de données MAR puis de vérifier la validité des conclusions par une analyse de sensibilité.

Nous nous sommes placés sous les hypothèses d'une maladie rare et d'une variable de phase-2 suivant une loi de la famille exponentielle. Des extensions pour des

maladies non-rares ou une variable de phase-2 ne suivant pas une loi de la famille exponentielle pourraient être intéressantes : distribution bimodale de la variable de phase-2, ce qui impliquerait d'estimer la distribution de la variable concernée comme un mélange de distributions. Pour une distribution de Poisson sur-dispersée, il est possible de travailler dans le cadre de la loi binomiale négative qui appartient à la famille exponentielle.

Bibliographie

BARLOW, W. (1994). Robust variance estimation for the case-cohort design. *Biometrics*, 50(4):1064–1072.

BELSLEY, D., KUH, E. et WELSCH, R. (1980). *Regression diagnostics : identifying influential data and sources of collinearity*. Wiley, New York :.

BORGAN, O., LANGHOLZ, B., SAMUELSEN, S., GOLDSTEIN, L. et POGODA, J. (2000). Exposure stratified case-cohort designs. *Lifetime Data Anal*, 6(1):39–58.

BRESLOW, N., LUMLEY, T., BALLANTYNE, C., CHAMBLESS, L. et KULICH, M. (2009a). Improved Horvitz-Thompson Estimation of Model Parameters from Two-phase Stratified Samples : Applications in Epidemiology. *Statistics in Biosciences*, 1(1):32–49.

BRESLOW, N., LUMLEY, T., BALLANTYNE, C., CHAMBLESS, L. et KULICH, M. (2009b). Using the whole cohort in the analysis of case-cohort data. *American journal of epidemiology*, 169(11):1398–1405.

BRIER, G. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

CAI, J. et ZENG, D. (2004). Sample size/power calculation for case-cohort studies. *Biometrics*, 60(4):1015–1024.

CAI, J. et ZENG, D. (2007). Power calculation for case-cohort studies with nonrare events. *Biometrics*, 63(4):1288–1295.

CARCAILLON, L., GAUSSEM, P., DUCIMETIERE, P., GIROUD, M., RITCHIE, K., DARTIGUES, J. F. et SCARABIN, P. Y. (2009). Elevated plasma fibrin D-dimer as a risk factor for vascular dementia : the Three-City cohort study. *Journal of Thrombosis and Haemostasis*, 7(12):1972–1978.

CHAVANCE, M. et MANFREDI, R. (2000). Modélisation d'observations incomplètes. *Rev Epidemiol Sante Publique*, 48(4):389–400.

CHEN, K. (2001). Generalized case-cohort sampling. *Journal Of The Royal Statistical Society Series B*, 63(4):791–809.

COTTRELL, G., COT, M. et MARY, J. (2009). L'imputation multiple des données manquantes aléatoirement : concepts généraux et présentation d'une méthode Monte-Carlo = Multiple imputation of missing at random data : general points and presentation of a Monte-Carlo method. *Revue d'Epidémiologie et de Santé Publique*, 57:361–372.

D'ANGIO, G., BRESLOW, N., BECKWITH, J., EVANS, A., BAUM, H., DELORIMIER, A., FERNBACH, D., HRABOVSKY, E., JONES, B. et KELALIS, P. (1989). Treatment of wilms' tumor. results of the third national wilms' tumor study. *Cancer*, 64(2):349–60.

DEVILLE, J. et SARNDAL, C. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382.

FRIEDEWALD, W. T., LEVY, R. I. et FREDRICKSON, D. S. (1972). Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clinical Chemistry*, 18(6):499–502.

GREENLAND, S. et FINKLE, W. (1995). A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American Journal of Epidemiology*, 142(12):1255–1264.

HARRELL, F., CALIFF, R., PRYOR, D., LEE, K. et ROSATI, R. (1982). Evaluating the yield of medical tests. *Journal of the American Medical Association*, 247(18):2543–2546.

HARRELL, F., LEE, K. et MARK, D. (1996). Multivariable prognostic models : Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4):361–387.

HENDERSON, R. (1995). Problems and prediction in survival data analysis. *Statistics in Medicine*, 14(2):161–184.

HOSMER, D. et LEMESHOW, S. (2000). *Applied logistic regression (Wiley Series in probability and statistics)*. Wiley-Interscience Publication, 2 édition.

KALBFLEISCH, J. et LAWLESS, J. (1988). Likelihood analysis of multi-state models for disease incidence and mortality. *Statistics in Medicine*, 7:149–160.

KENT, J. et O'QUIGLEY, J. (1988). Measures of dependence for censored survival data. *Biometrika*, 75(3):525–534.

KREMERS, W. (2007). Concordance for survival time data : fixed and time-dependent covariates and possible ties in predictor and time. Technical Report Series No. 80, Departement of Health Science Research, Mayo Clinic, Rochester, Minnesota.

KUBOTA, K. et WAKANA, A. (2011). Sample-size formula for case-cohort studies. *Epidemiology*, 22(2):279.

KULICH, M. et LIN, D. (2004). Improving the efficiency of relative-risk estimation in case-cohort studies. *Journal of the American Statistical Association*, 99:832–844.

LANGHOLZ, B. (1998). *Case-cohort study*. In P Armetage and D colton (eds), Encyclopedia of Biostatistics. Wiley Chichester.

LIN, D. et YING, Z. (1993). Cox regression with incomplete covariate measurements. *Journal of the American Statistical Association*, 88(424):1341–1349.

LITTLE, R. et RUBIN, D. (1987). *Statistical analysis with missing data*. New York : Wiley.

MACKIE, I., KITCHEN, S., MACHIN, S. et LOWE, G. (2003). Guidelines on fibrinogen assays. *British Journal of Haematology*, 121(3):396–404.

MANN, H. et WHITNEY, D. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50–60.

METZ, C. (1978). Basic principles of ROC analysis. *Seminars in nuclear medicine*, 8(4):283–298.

ONLAND-MORET, N., van der A, D., van der SCHOUW, Y., BUSCHERS, W., ELIAS, S., van GILS, C., KOERSELMAN, J., ROEST, M., GROBBEE, D. et PEETERS, P. (2007). Analysis of case-cohort data : a comparison of different methods. *J Clin Epidemiol*, 60: 350–355.

PENCINA, M. et D'AGOSTINO, R. (2004). Overall C as a measure of discrimination in survival analysis : model specific population value and confidence interval estimation. *Statistics in Medicine*, 23(13):2109–2123.

PENCINA, M., D'AGOSTINO, Sr., R., D'AGOSTINO, Jr., R. et VASAN, R. (2008). Evaluating the added predictive ability of a new marker : From area under the ROC curve to reclassification and beyond. *Statistics in Medicine*, 27(2):157–172.

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

R DEVELOPMENT CORE TEAM (2010). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

ROBINS, J., ROTNITZKY, A. et ZHAO, L. (1994). Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association*, 89(427).

ROBINS, J., ROTNITZKY, A. et ZHAO, L. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121.

RUBIN, D. (1987). *Multiple imputation for nonresponse in surveys*. New York : Wiley.

RUBIN, D. et SCHENKER, N. (1991). Multiple imputation in health-care databases : an overview and some applications. *Statistics in Medicine*, 10(4):585–598.

SAMUELSEN, S., HALLVARD, A. et SKRONDAL, A. (2007). Stratified case-cohort analysis of general cohort sampling designs. *Scandinavian Journal of Statistics*, 34(1):103–119.

SÄRNDAL, C.-E., SWENSSON, B. et WRETMAN, J. H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76(3):527–537.

SCARABIN, P., ARVEILER, D., AMOUYEL, P., DOS SANTOS, C., EVANS, A., LUC, G., FERRIÈRES, J., JUHAN-VAGUE, I. et of MYOCARDIAL INFARCTION, P. E. S. (2003). Plasma fibrinogen explains much of the difference in risk of coronary heart disease between france and northern ireland. the prime study. *Atherosclerosis*, 166(1):103–109.

SCHEMPER, M. (1990). The explained variation in proportional hazards regression. *Biometrika*, 77(1):216–218.

SELF, S. et PRENTICE, R. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *The Annals of Statistics*, 16:64–81.

STEYERBERG, E., VICKERS, A., COOK, N., GERDS, T., GONEN, M., OBUCHOWSKI, N., PENCINA, M. et KATTAN, M. (2010). Assessing the performance of prediction models : a framework for traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 21(1):128–138.

THERNEAU, T. et GRAMBSCH, P. (2000). *Modeling Survival Data : Extending the Cox Model*. Springer. ISBN : 0-387-98784-3.

THERNEAU, T. et LI, H. (1999). Computing the cox model for case cohort designs. *Lifetime Data Analysis*, 5(2):99–112.

TZOULAKI, I., MURRAY, G., LEE, A., RUMLEY, A., LOWE, G. et FOWKES, F. (2007). Relative value of inflammatory, hemostatic, and rheological factors for incident myocardial infarction and stroke - The Edinburgh Artery Study. *Circulation*, 115(16):2119–2127.

VACH, W. et BLETTNER, M. (1991). Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. *American Journal of Epidemiology*, 134:895–907.

van BUUREN, S. et GROOTHUIS-OUDSHOORN, K. (2011). mice : Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67.

WACHOLDER, S. et BOIVIN, J. (1987). External comparisons with the case-cohort design. *J Am J Epidemiol*, 6:1198–209.

WANG, T., GONA, P., LARSON, M., TOFLER, G., LEVY, D., NEWTON-CHEH, C., JACQUES, P., RIFAI, N., SELHUB, J., ROBINS, S., BENJAMIN, E., D'AGOSTINO, R. et VASAN, R. (2006). Multiple biomarkers for the prediction of first major cardiovascular events and death. *New England Journal of Medicine*, 355(25):2631–2639.

YARNELL, J. (1998). The prime study : classical risk factors do not explain the severalfold differences in risk of coronary heart disease between france and northern ireland. prospective epidemiological study of myocardial infarction. *QJM*, 91(10):667–76.

ZWEIG, M. et CAMPBELL, G. (1993). Receiver-operating characteristic (ROC) plots : a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39(4):561–577.

Annexes

Imputation multiple dans les enquêtes cas-cohorte « cch.MI »

Description

Analyse d'enquêtes cas-cohorte par imputation multiple

Bibliothèques requises : mice, survival.
Fonctions requises : « cox.mice.r », « mat.pred.MI.r ».
Télécharger les données « sim.txt ».

Usage

```
cch.MI (data, analysis.model, phase2.var, imp.model, nimp, graine)
```

Arguments

`data` Données
`analysis.model` Modèle d'analyse
`phase2.var` Vecteur alphanumérique des noms des variables de phase-2
`imp.model` Liste des modèles d'imputation, faisant appel aux fonctions `glm` ou `lm`. Peut inclure des interactions (voir exemple ci-dessous)
`nimp` Nombre d'imputations
`graine` Graine pour initialiser le générateur de nombres pseudo-aléatoires

Résultats

La function cch.MI.r retourne comme résultat :
`data.mipo` Résultats de l'analyse combinée

Les objets « data.mira » et « matrice.pred » sont accessibles en dehors de la fonction

Références

Marti H, Chavance M. Multiple imputation analysis of case-cohort studies. Stat Med. 2011 Jun 15;30(13):1595-607

Voir aussi

cch

Exemples

Données « sim.txt ». Ce fichier contient 10000 sujets et 10 variables.

Z1 Variable binaire de phase-1

Z2 Variable normale de phase-2 (dans cet exemple, observée pour tous les sujets)
newZ2 Variable normale de phase-2 (manquante pour les sujets n'appartenant pas à l'échantillon cas-cohorte. Variable qui sera imputée)
Z2sur Variable de phase-1 prédictive de la variable de phase-2
Z3 Variable normale de phase-1
Z4 Variable de phase-2
Zevent Variable indicatrice d'événement : 0- non événement, 1- événement
delai Délai jusqu'à l'événement ou la censure
in.cchsample Variable indicatrice d'appartenance à l'échantillon cas-cohorte
in.subcohort Variable indicatrice d'appartenance à la sous -cohorte

```
source("cox.mice.r ")
source("mat.pred.MI.r ")
source("cch.MI.r ")
sim<-read.table("sim.txt",header=T)
names(sim)


# Analyse de la cohorte complète (nous utilisons ici Z2 au lieu de newZ2)
est.full.data<-coxph(Surv(delai,Zevent)~ Z1+ Z2+ Z3,data=sim)
summary(est.full.data)

# Analyse pondérée (avec « cch » de la bibliothèque « survival »)
cch_data<-subset(sim,in.cchsample)
cch_data$id<-1:nrow(cch_data)
est.LY<-cch(Surv(delai,Zevent)~Z1+newZ2+Z3,id=~id,data=cch_data,
subcoh=~in.subcohort, cohort.size=dim(sim)[1], method="LinYing")
est.LY

# Analyse par imputation multiple
est.IM<-cch.MI(data=sim,  analysis.model = coxph(Surv(delai,Zevent) ~ Z1 +
newZ2 + Z3, data=sim), phase2.var = "newZ2", imp.model = list(glm(newZ2 ~
Zevent  + Z1 + Z2sur + Z3, family=gaussian, data=sim)), nimp=5, graine =
758)
```

Résultats : Estimations (écart-type) fournies par l'analyse de la cohorte complète (est.full.data), l'analyse pondérée (est.LY) et l'imputation multiple (est.IM)

	est.full.data	est.LY	est.IM
Z1	0.641 (0.114)	0.572 (0.134)	0.602 (0.114)
Z2	0.697 (0.057)	0.694 (0.069)	0.686 (0.061)
Z3	0.634 (0.055)	0.691 (0.067)	0.642 (0.056)

```
# Exemple de syntaxe pour plusieurs variables de phase-2
est.IM2<-cch.MI(data=sim, analysis.model = coxph(Surv(delai,Zevent) ~ Z1 +
newZ2 + Z3 + Z4 + Z1*newZ2, data=sim), phase2.var = c("newZ2","Z4"),
imp.model = list(glm(newZ2 ~ Zevent + Z1 + Z2sur + Zevent*Z1,
family=gaussian, data=sim), glm(Z4 ~ Zevent + Z1 + newZ2 + Z3 + Z1*newZ2,
family=gaussian, data=sim)), nimp=5, graine = 758)
```

Fonction « cch. MI.r »

```
cch.MI<-function(data,analysis.model,phase2.var,imp.model,nimp,graine)
  {
  # data : donnéés
  # analysis.model : modèle d'analyse
  # phase2.var : vecteur alphanumérique des noms des variables de phase-2
  # imp.model : list des modèle(s) d'imputation, faisant appel aux
fonctions "glm" ou "lm".
  #             Peut inclure des interactions
  # nimp : nombre d'imputations souhaitées
  # graine : graine pour effectuer les imputations

  # Création de la matrice de prédiction
  pred.MI<-mat.pred.MI(data=data,analysis.model = analysis.model,
  phase2.var = phase2.var,imp.model = imp.model)

  # Imputation multiple selon la matrice de prédiction "matrice.pred"
définie ci-dessus
  data.mids<-mice(dataaux, m = nimp,
                  method = meth,
                  predictorMatrix = matrice.pred,
                  visitSequence =
(1:ncol(dataaux))[apply(is.na(dataaux),2,any)],
                  defaultMethod = c("norm","logreg","polyreg"),
                  maxit = 5,
                  diagnostics = TRUE,
                  printFlag = TRUE,
                  seed = graine)
  data.mira<-coxph.mids(analysis.model$formula,data=data.mids)
  data.mipo<-pool(data.mira)
  data.mipo<<-data.mipo
  matrice.pred<<-matrice.pred
  data.mira<<-data.mira #
  xxx<<-complete(data.mids,1,inc=T)
  return(data.mipo)
 }
```

Création de la matrice de prédiction pour l'imputation multiple « mat.pred.MI »

Description

Création des fichiers auxiliaires « matrice.pred » et « dataaux » pour l'analyse des enquêtes cas-cohorte par imputation multiple.

Usage

```
mat.pred.MI(data, analysis.model, phase2.var, imp.model)
```

Arguments

`data` Données
`analysis.model` Modèle d'analyse
`phase2.var` Vecteur alphanumérique des noms des variables de phase-2
`imp.model` Liste des modèles d'imputation, faisant appel aux fonctions `glm` ou `lm`. Peut inclure des interactions (voir exemple ci-dessous)

Résultats

La function mat.pred.MI.r retourne comme résultat :
`matrice.pred` Matrice de prédiction définissant les variables qui interviennent dans chacun des modèles d'imputation
`dataaux` Matrice de données contenant les variables initiales et, au besoin, les interactions du modèle d'imputation

Références

Marti H, Chavance M. Multiple imputation analysis of case-cohort studies. Stat Med. 2011 Jun 15;30(13):1595-607

Voir aussi

`cch.MI`

Exemples

Données « sim.txt ». Ce fichier contient 10000 sujets et 10 variables.

Z1 Variable binaire de phase-1
Z2 Variable normale de phase-2 (dans cet exemple, observée pour tous les sujets)
newZ2 Variable normale de phase-2 (manquante pour les sujets n'appartenant pas à l'échantillon cas-cohorte. Variable qui sera imputée)
Z2sur Variable de phase-1 prédictive de la variable de phase-2
Z3 Variable normale de phase-1

Z4 Variable de phase-2
Zevent Variable indicatrice d'événement : 0- non événement, 1- événement
*delai D*élai jusqu'à l'événement ou la censure
in.cchsample Variable indicatrice d'appartenance à l'échantillon cas-cohorte
in.subcohort Variable indicatrice d'appartenance à la sous -cohorte

```
source("cox.mice.r ")
source("mat.pred.MI.r ")
sim<-read.table("sim.txt",header=T)

mat.pred.MI(data=sim,
analysis.model = coxph(Surv(delai,Zevent) ~ Z1 + newZ2 + Z3, data=sim),
phase2.var = "newZ2",
imp.model = list(glm(newZ2 ~ Zevent  + Z1 + Z2sur + Z3, family=gaussian,
data=sim)))
```

Fonction « mat.pred.MI.r »

```
mat.pred.MI<-function(data,analysis.model,phase2.var,imp.model)
  {
  # data : données
  # analysis.model : modèle d'analyse
  # phase2.var : vecteur alphanumérique des noms des variables de phase-2
  # imp.model : list des modèle(s) d'imputation, faisant appel aux
fonctions "glm" ou "lm".
  #             Peut inclure des interactions
  attach(data,warn.conflicts = FALSE)
  data$event=cbind(subset(data,
select=c(paste(analysis.model$formula[[2]][[3]]))))
  phase2<-subset(data,select=c(phase2.var),colnames=phase2.var)
  i=length(c(phase2.var))
  in2=0
  # Une seule variable de phase-2
  if (i==1)
  {
    modelmat<-model.matrix(imp.model[[i]],data)
    name.imp.cov<-cbind(colnames(modelmat)[-1])
    modelmat<-model.matrix(imp.model[[i]]$formula[-2],data)
            if (dim(data)[1]==dim(modelmat)[1])
            {
            dataaux<-cbind(phase2[i],modelmat[,-1])
            }
            else
            {
            modelmat<-modelmat[match(rownames(data),rownames(modelmat)),]
            dataaux<-cbind(data,modelmat)
            }
    dataaux<-cbind(phase2[i],modelmat[,-1])
    dataaux<-subset(dataaux,select=c(names(dataaux)))
    aux<-unique(c(names(data),names(dataaux)))
    data<-cbind(data,dataaux)
    data<-subset(data,select=c(aux))
    data<-data
    matrice.pred<-matrix(data=0, nrow=dim(dataaux)[2],ncol=dim(dataaux)[2])
    colnames(matrice.pred)<-rownames(matrice.pred)<-c(names(dataaux))
    matrice.pred[c(phase2.var)[i],name.imp.cov]=1
    ini<-mice(dataaux,max=0,print=F,defaultMethod =
c("norm","logreg","polyreg"))
    meth<-ini$meth
    }
  # Plusieurs variables de phase-2
  else
  {
    aux<-c(attr(imp.model[[1]]$terms,"term.labels"))
      for (k in 2:(i))
      {
      aux2=c(aux,attr(imp.model[[k]]$terms,"term.labels"))
      aux=aux2
      }
    aux<-c(phase2.var,aux)
    aux<-unique(aux)
    # Traitement des interactions dans les modèles d'imputation
    data2=data
    aux3=c(rep(NA,length(imp.model)))
    for (i in 1:length(imp.model))
    {
```

```
        if
(length(attr(terms(imp.model[[i]]),"term.labels")[attr(terms(imp.model[[i]]
),"order")>1])==0)
        {
        aux3[i]<-attr(terms(imp.model[[i]]),"term.labels")
        auxform<-formula(paste("~",paste(aux3[i],collapse="+")))
        modelmataux2<- model.matrix(auxform,data2)
        }
        else
        {
        aux3[i]<-
attr(terms(imp.model[[i]]),"term.labels")[attr(terms(imp.model[[i]]),"order
")>1]
        auxform<-formula(paste("~",paste(aux3[i],collapse="+")))
        modelmataux2<- model.matrix(auxform,data2)
            if (dim(data2)[1]==dim(modelmataux2)[1])
            {
            data2<-cbind(data2,modelmataux2)
            }
            else
            {
            modelmataux2<-
modelmataux2[match(rownames(data2),rownames(modelmataux2)),]
            data2<-cbind(data2,modelmataux2)
            }
        }
    }
    names(data2)
    dataaux<-data2[aux]
  # Définition de la matrice de prédiction correspondant aux modèles
d'imputation
  matrice.pred<-matrix(data=0, nrow=dim(dataaux)[2],ncol=dim(dataaux)[2])
  colnames(matrice.pred)<-rownames(matrice.pred)<-c(names(dataaux))
  in2=0
  for (i in 1:length(c(phase2.var)))
   {

matrice.pred[c(phase2.var)[i],attr(imp.model[[i]]$terms,"term.labels")]=1
   }
    for (j in (1:length(aux3)))
    {
      vecteur.variables<-aux3
      nom.vect<-c()
      for (i in 1:nchar(aux3[j])){
      lettre=substr(aux3[j],i,i)
      nom.vect<-c(nom.vect,lettre)
      }
      inter=which(nom.vect==":")
      a=substr(aux3[j],1,inter-1)
      which(vecteur.variables==a)
      b=substr(aux3[j],inter+1,nchar(aux3[j]))
      which(vecteur.variables==b)
      a
      b
      in2=0
      for (k in (length(phase2.var)-1)){
      if (a==phase2.var[k]| a==phase2.var[k+1]) in2=1
      if (b==phase2.var[k]| b==phase2.var[k+1]) in2=1
      }
      ini<-mice(dataaux,max=0,print=F,defaultMethod =
c("norm","logreg","polyreg"))
```

```
      meth<-ini$meth
      if (in2>0) {
      meth[aux3[j]]<-paste("~I(",a,"*",b,")")
      }
      meth
    }
  }
  meth<<-meth
  matrice.pred<<-matrice.pred
  dataaux<<-dataaux
  return(matrice.pred)
}
```

```
coxph.mids <- function (formula, data, ...) {
   call <- match.call()
   if (!is.mids(data)) stop("The data must have class mids")
   analyses <- as.list(1:data$m)
   for (i in 1:data$m) {
     data.i        <- complete(data, i)
     analyses[[i]] <- coxph(formula, data = data.i, ...)
   }
   object <- list(call = call, call1 = data$call, nmis = data$nmis, analyses =
analyses)
#   oldClass(object) <- if (.SV4.) "mira" else c("mira", "coxph")
   oldClass(object) <- c("mira", "coxph")
return(object)
}

pool <- function (object, method = "smallsample") {
   call <- match.call()
   if (!is.mira(object)) stop("The object must have class 'mira'")
   if ((m <- length(object$analyses)) < 2)
         stop("At least two imputations are needed for pooling.\n")
   analyses <- object$analyses
   k     <- length(coef(analyses[[1]]))
   names <- names(coef(analyses[[1]]))
   qhat  <- matrix(NA, nrow = m, ncol = k, dimnames = list(1:m, names))
   u     <- array(NA, dim = c(m, k, k),dimnames = list(1:m, names, names))
   for (i in 1:m) {
     fit       <- analyses[[i]]
     qhat[i, ] <- coef(fit)
     u[i, , ]  <- vcov(fit)
   }
   qbar <- apply(qhat, 2, mean)
   ubar <- apply(u, c(2, 3), mean)
   e <- qhat - matrix(qbar, nrow = m, ncol = k, byrow = TRUE)
   b <- (t(e) %*% e)/(m - 1)
   t <- ubar + (1 + 1/m) * b
   r <- (1 + 1/m) * diag(b/ubar)
   f <- (1 + 1/m) * diag(b/t)
   df <- (m - 1) * (1 + 1/r)/2
   if (method == "smallsample") {
         if( any( class(fit) == "coxph" ) ){
       ### this loop is the hack for survival analysis ###
       status   <- fit$y[ , 2]
#       status   <- fit$y[ , 3]
       n.events <- sum(status == max(status))
       p        <- length( coefficients( fit )  )
       dfc      <- n.events - p
     }
     else {
       dfc <- fit$df.residual
     }
     df <- dfc/((1 - (f/(m + 1)))/(1 - f) + dfc/df)
   }
   names(r) <- names(df) <- names(f) <- names
   fit <- list(call = call, call1 = object$call, call2 = object$call1,
               nmis = object$nmis, m = m, qhat = qhat, u = u,
               qbar = qbar, ubar = ubar, b = b, t = t, r = r, df = df,
               f = f)
```

```
   oldClass(fit) <- c("mipo", oldClass(object))
   return(fit)
}

summary.mipo <- function(object){
    if (!is.null(object$call1)){
      cat("Call: ")
      dput(object$call1)
    }
    est  <- object$qbar
    se   <- sqrt(diag(object$t))
    tval <- est/se
    df   <- object$df
    pval <- 2 * pt(abs(tval), df, lower.tail = FALSE)
    coefmat <- cbind(est, se, tval, pval)
    colnames(coefmat) <- c("Estimate", "Std. Error","t value", "Pr(>|t|)")
    cat("\nCoefficients:\n")
    printCoefmat( coefmat, P.values=T, has.Pvalue=T, signif.legend=T )
    cat("\nFraction of information about the coefficients missing due to
nonresponse:", "\n")
    print(object$f)
    ans <- list( coefficients=coefmat, df=df,
                 call=object$call1, fracinfo.miss=object$f )
    invisible( ans )
}
```

Research Article

Statistics in Medicine

Received 23 December 2009, Accepted 3 October 2010 Published online 24 February 2011 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4130

Multiple imputation analysis of case–cohort studies

Helena Marti*† and Michel Chavance

The usual methods for analyzing case–cohort studies rely on sometimes not fully efficient weighted estimators. Multiple imputation might be a good alternative because it uses all the data available and approximates the maximum partial likelihood estimator. This method is based on the generation of several plausible complete data sets, taking into account uncertainty about missing values. When the imputation model is correctly defined, the multiple imputation estimator is asymptotically unbiased and its variance is correctly estimated. We show that a correct imputation model must be estimated from the fully observed data (cases and controls), using the case status among the explanatory variable. To validate the approach, we analyzed case–cohort studies first with completely simulated data and then with case–cohort data sampled from two real cohorts. The analyses of simulated data showed that, when the imputation model was correct, the multiple imputation estimator was unbiased and efficient. The observed gain in precision ranged from 8 to 37 per cent for phase-1 variables and from 5 to 19 per cent for the phase-2 variable. When the imputation model was misspecified, the multiple imputation estimator was still more efficient than the weighted estimators but it was also slightly biased. The analyses of case–cohort data sampled from complete cohorts showed that even when no strong predictor of the phase-2 variable was available, the multiple imputation was unbiased, as precised as the weighted estimator for the phase-2 variable and slightly more precise than the weighted estimators for the phase-1 variables. However, the multiple imputation estimator was found to be biased when, because of interaction terms, some coefficients of the imputation model had to be estimated from small samples. Multiple imputation is an efficient technique for analyzing case–cohort data. Practically, we suggest building the analysis model using only the case–cohort data and weighted estimators. Multiple imputation can eventually be used to reanalyze the data using the selected model in order to improve the precision of the results. **Copyright © 2011 John Wiley & Sons, Ltd.**

Keywords: case–cohort design; multiple imputation

1. Introduction

Cohort studies, which facilitate causal interpretations, are popular but expensive. Because precision is mainly limited by the number of cases, it is not essential to collect complete information for all the controls. Thus, case–cohort studies and nested case–control studies enable cost reduction with a minimal loss of efficiency [1]. Case–cohort studies were initially proposed by Prentice [2]. When using this approach, the information collected for incompletely observed controls is ignored and inefficient estimators for the effect of phase-1 variables are obtained. In addition, the weighted estimators used to analyze case–cohort data are not fully efficient and this could affect the estimate of the effect of phase-2 variable(s). Alternatively, Breslow *et al.* [3] suggested optimizing the sampling weights using all the available data. But case–cohort studies can also be viewed as a particular example of incomplete data, in which the observation process is controlled by the study organizers. Paik and Tsai [4] proposed a simple imputation approach to model censored observations with missing covariates. In the framework of case–cohort studies, it would imply the simple imputation of the expected value of the

Inserm, CESP Centre for Research in Epidemiology and Population Health, U1018, Biostatistics team, F-94807 Villejuif, France

**Correspondence to: Helena Marti, Inserm, CESP Centre for Research in Epidemiology and Population Health, U1018, Biostatistics team, F-94807 Villejuif, France.*

†E-mail: helena.marti-soler@inserm.fr

Copyright © 2011 John Wiley & Sons, Ltd.

phase-2 variable(s) for incomplete controls. However, that approach ignores the uncertainty concerning imputation model parameters and the values to impute according to a given model.

Multiple imputation is a simple and efficient method for analyzing incomplete observations, while taking into account all the levels of uncertainty regarding missing values. For case–cohort data analysis, the multiple imputation estimator may provide improved precision, compared to weighted estimators, because it integrally uses the available information and approximates the partial likelihood estimator, which can be more efficient than the weighted estimators.

The objective of this study was to establish multiple imputation as an alternative to weighted analysis of case–cohort data. Below, we present the multiple imputation analysis of case–cohort studies and validate this approach by comparing its results to those obtained with weighted estimators. First, we used entirely simulated data. Then, we simulated case–cohort surveys from two cohorts, the Prospective Epidemiological Study of Myocardial Infarction (PRIME) study, in which no strong surrogate of the chosen phase-2 variable was available, and the National Wilms' Tumor Study (NWTS) data, for which a surrogate was available. For simplicity, we only consider time-constant covariates. Extension to time-varying covariates is considered in the discussion.

2. Weighted analysis of case–cohort studies

Case–cohort surveys are examples of two-phase designs. First, the cohort is randomly selected from a general population and the phase-1 information is collected for all the subjects. A subcohort is randomly selected and the entire cohort is followed so as to identify the date of occurrence of the event(s) of interest. Then, the phase-2 information, more expensive, is collected for the subcohort subjects and for all the cases, whether or not they belong to the subcohort. Thus, the phase-2 information is not available for controls not belonging to the subcohort. In cohort surveys, where data are available for the whole cohort, the effect of risks factors on the occurrence of events is generally measured by fitting a proportional hazards model. In case–cohort surveys this model is based on phase-1 and phase-2 variables and the parameters must be estimated from the available incomplete data. In simulations we will consider two phase-1 variables, Z_1 and Z_3, and one phase-2 variable, Z_2.

What is lost in terms of efficiency, when using a case–cohort design rather than a full cohort analysis, can be quantified by the asymptotic relative efficiency (ARE). For a case–cohort design with simple random sampling it was shown to be [5]:

$$\begin{aligned} \text{ARE} &\approx \left\{1+2\frac{1-\alpha}{\alpha}\left[1+\frac{1-d}{d}\log(1-d)\right]\right\}^{-1} \\ &\approx 1-\gamma, \end{aligned} \tag{1}$$

where α is the proportion of the cohort in the subcohort sample, d is the probability of event occurrence, and γ is the fraction of missing information. However, when a phase-1 variable is strongly predictive of the phase-2 variable, stratified sampling of the subcohort can improve efficiency as compared to simple random selection [6].

Weighted estimators of the log-relative risks maximize a weighted pseudo-likelihood ($\widetilde{L}(\beta)$):

$$\widetilde{L}(\beta)=\prod_j\left(\frac{\exp\{\beta' Z_i\}w_{i_j}}{\sum_{k\varepsilon\widetilde{C}\cup D}Y_k(t_j)\exp\{\beta' Z_k\}w_{k_j}}\right) \tag{2}$$

where event j occurs at time t_j, $\widetilde{C}$ is the subcohort of size n_{sc}, D is the set of cases, $Y_k(t_j)$ indicates whether subject k is at risk at time t_j, β is the vector of log relative risks, Z_k the vector of covariates for subject k, w_{k_j} the weight of subject k at time t_j, i_j the index of the subject whose event occurs at t_j, and the symbol $'$ denotes transposition. Barlow [7] proposed weighting each complete observation by the inverse of its probability of being included (1 for the cases). Other authors have proposed variable weights, as a function of time, to slightly improve efficiency [6].

The variance of this estimator must take into account the increased uncertainty associated with the randomized selection of the subcohort. This requirement can be achieved by using the sandwich

Copyright © 2011 John Wiley & Sons, Ltd.

variance, which can be estimated as

$$\mathrm{Var}(\hat{\beta}) = \widehat{I}^{-1} + \frac{n_{\mathrm{sc}}(n - n_{\mathrm{sc}})}{n} \mathrm{Cov}\, D_C \tag{3}$$

where I is the Fisher information matrix, n and n_{sc} are the respective sizes of the cohort and subcohort, and $\mathrm{Cov}\, D_C$ is the empirical covariance matrix of *dfbeta* residuals from subcohort members defined as [8]

$$dfbeta_{ji} = \frac{\beta_j - \beta_{j(i)}}{s_{(i)}\sqrt{(Z'Z)^{-1}_{jj}}} \tag{4}$$

with $\beta_{j(i)}$ the parameter j estimate obtained after deletion of subject i and $s_{(i)}$ the standard error of this estimate.

Stratified sampling of the subcohort considers some information obtained during phase 1, but the information provided by the phase-1 variables is generally ignored in the analysis. Kulich and Lin [9] proposed a family of doubly weighted estimators intended to more efficiently account for the information provided by the initial variables. Qi *et al.* [10] developed nonparametric methods to estimate selection probabilities and nonparametric kernel-smoothing techniques to estimate conditional expectation in fully augmented weighted estimating functions. Breslow *et al.* [3] suggested calibrating or estimating the weights using all the phase-1 information in order to improve precision: (1) with calibration, the weights are subjected to the constraint that the cohort totals of some auxiliary variables are equal to their weighted sum among all phase-2 subjects. Practically, one builds a prediction model for the phase-2 variable to perform a simple imputation of the predicted values among the controls not belonging to the subcohort, fits the model of interest to the completed data set and uses the influence function from the model of interest to calibrate. Eventually, the model of interest is fitted to the calibrated case–cohort data; (2) with estimation, the weights are the reciprocals of the inclusion probabilities, as estimated from a logistic model fitted to the full cohort.

3. Incomplete observations and multiple imputation

Little and Rubin [11] distinguished three observation processes: data missing completely at random (MCAR), when the probability of incomplete observation is constant; data missing at random (MAR), when this probability depends only on observed values; and data missing not at random (MNAR), when this probability depends on unobserved values. The distinction between MAR and MNAR is of utmost importance, because with the former, unbiased estimators of the parameters of interest are available. Case–cohort data are MAR, because the probability of being completely observed depends only on case status and, under stratified sampling, on some phase-1 variables.

The multiple imputation method developed by Little and Rubin [11] provides an approximation of the maximum likelihood estimator and thus enables the potential selection bias to be corrected. This method relies on the generation of several plausibly completed data sets ($M \geqslant 2$), accounting for all the levels of uncertainty concerning the missing values. A prediction model must be built, taking into consideration the relationships between the incomplete variable and the other variables, as observed in the complete part of the data. The missing data are not replaced by their expectation but by a value drawn from the distribution posited by the model. To take into account the uncertainty concerning the parameters of the imputation model, several imputations are performed with parameters drawn from the asymptotic distribution of their estimator. An estimate of the parameter of interest, $\hat{\theta}_m$, $m = \{1, \ldots, M\}$, and an estimate of the variance of the estimator, $\widehat{V}(\hat{\theta}_m)$, are obtained from each completed data set. If the imputation model is correct, these estimators are not biased. The multiple imputation estimate, also unbiased, is the mean of these M estimates:

$$\hat{\theta}_{\mathrm{IM}} = \frac{1}{M}\sum_{m=1}^{M} \hat{\theta}_m \tag{5}$$

The multiplicity of imputations enables a correct estimation of the variance of this single estimator. The variance is the sum of two components: the *within-imputations* component (W_{IM}), estimated as the

Copyright © 2011 John Wiley & Sons, Ltd.

mean of the M asymptotic variances, $\widehat{W}_{\mathrm{IM}}$, and the *between-imputations* component (B_{IM}), estimated from the observed variance of the M estimates, $\widehat{B}_{\mathrm{IM}}$:

$$\begin{aligned}\widehat{V}(\hat{\theta}_{\mathrm{IM}}) &= \widehat{W}_{\mathrm{IM}} + \widehat{B}_{\mathrm{IM}} \\ &= \frac{1}{M}\sum_{m=1}^{M}\widehat{V}(\hat{\theta}_m) + (1+M^{-1})\frac{\sum_{m=1}^{M}(\hat{\theta}_m - \hat{\theta}_{\mathrm{IM}})(\hat{\theta}_m - \hat{\theta}_{\mathrm{IM}})'}{M-1}\end{aligned} \tag{6}$$

where the factor $1+M^{-1}$ is an adjustment for using a finite number of imputations [12].

Rubin [13] showed that the relative efficiency of multiple imputations with a finite number M, as compared to an infinite number of imputations, is

$$\mathrm{ARE} \approx \sqrt{1+\frac{\gamma}{M}} \tag{7}$$

where γ is the fraction of missing information:

$$\gamma \approx \frac{B_{\mathrm{IM}}}{B_{\mathrm{IM}} + W_{\mathrm{IM}}} \tag{8}$$

A nonexhaustive review of published case–cohort studies showed that, among 25 studies, the fraction of missing information ranged from 0.05 to 0.5, with a median around 0.3. With as much as 40 per cent information missing, $M=5$ imputations provides an ARE$=0.97$, and, with 50 per cent missing information, $M=10$ provides an ARE$=0.98$. Thus, in most instances, there is not much to gain by using more than 5 or 10 imputations to analyze case–cohort data.

Multiple imputation requests a correct model of the relationships between the incomplete variable(s) and the variables that are linked to the former. When the statistician doing the multiple imputation is independent of the statistician conducting the analyses: '… it is important to include as predictors as many of the variables that are likely to be used in subsequent analyses as possible. Leaving out such variables, even when they are weak predictors, implies that it is known with certainty that they have no relation with the missing values. The result is that correct uncertainty is not reflected [12]'. In case–cohort studies, the imputation model and the analysis are in the hands of the same statistician, thus only variables useful for the analysis of interest have to be included.

We need to impute missing phase-2 variable values for the controls who do not belong to the subcohort. This requires an imputation model taking into account the differences between cases and controls: otherwise, the multiple imputation estimator of the effect of the phase-2 variable would be biased. Under the rare disease assumption, it can be shown that a simple generalized linear model using all the case–cohort data and including the status indicator among the explanatory variables has to be considered.

Let us assume that the distribution of Z_2 belongs to the exponential family and depends on a phase-1 variable, possibly multidimensional $\tilde{z}_2$ through

$$f(z_2 \mid \tilde{z}_2) = \exp\left(\frac{\theta z_2 - b(\theta) + c(z_2)}{a(\phi)}\right), \tag{9}$$

where ϕ is the dispersion parameter and the canonical parameter θ is a linear function of unknown parameters:

$$\theta = \alpha_0 + \alpha_1' \tilde{z}_2. \tag{10}$$

Under the rare disease assumption, the distribution of Z_2 is approximately the same for the whole population and among controls

$$f(z_2 \mid \tilde{z}_2, \Delta = 0) \simeq f(z_2 \mid \tilde{z}_2), \tag{11}$$

where Δ is the case indicator. Let $\pi(\tilde{z}_2, \mu_2, t_c)$ be the probability of being a case at the end of the observation time, for a subject with $\tilde{Z}_2 = \tilde{z}_2$, $Z_2 = \mu_2 = E[Z_2 \mid \tilde{Z}_2 = \tilde{z}_2]$ and censoring time $T_C = t_c$; let $\pi(\tilde{z}_2, z_2, t_c)$ be the probability of being a case at the end of the observation time, for a subject with $\tilde{Z}_2 = \tilde{z}_2$, $Z_2 = z_2$, and $T_C = t_c$. According to the proportional hazards model and using the rare disease assumption

$$\pi(\tilde{z}_2, z_2, t_c) = \pi(\tilde{z}_2, \mu_2, t_c)\exp[\beta(z_2 - \mu_2)]. \tag{12}$$

Copyright © 2011 John Wiley & Sons, Ltd.

We also assume that the distribution of the censoring time is independent of Z_2, so integrating over the censoring time

$$\pi(\tilde{z}_2, z_2) = \pi(\tilde{z}_2, \mu_2)\exp[\beta(z_2-\mu_2)] \tag{13}$$

and the distribution of Z_2 conditionally on being a case and on $\tilde{Z}_2$, can be obtained as

$$\begin{aligned} f(z_2 \mid \Delta=1, \tilde{z}_2) &= \frac{P[\Delta=1 \mid \tilde{z}_2, z_2]}{P[\Delta=1 \mid \tilde{z}_2]} f(z_2 \mid \tilde{z}_2) \\ &= \frac{\pi(\tilde{z}_2, \mu_2)\exp[\beta(z_2-\mu_2)]}{\int \pi(\tilde{z}_2, \mu_2)\exp[\beta(z-\mu_2)] f(z \mid \tilde{z}_2)\mathrm{d}z} f(z_2 \mid \tilde{z}_2) \\ &= \frac{\exp(\beta z_2)}{\int \exp(\beta z) f(z \mid \tilde{z}_2)\mathrm{d}z} f(z_2 \mid \tilde{z}_2). \end{aligned} \tag{14}$$

The denominator can be developed as

$$\begin{aligned} \int \exp(\beta z) f(z \mid \tilde{z}_2)\mathrm{d}z &= \int \exp\left(\beta z + \frac{\theta z_2 - b(\theta) + c(z_2)}{a(\phi)}\right)\mathrm{d}z \\ &= \int \exp\left(\frac{(\theta + a(\phi)\beta)z - b(\theta + a(\phi)\beta) + b(\theta + a(\phi)\beta) - b(\theta) + c(z)}{a(\phi)}\right)\mathrm{d}z \\ &= \exp\left[\frac{b(\theta + a(\phi)\beta) - b(\theta)}{a(\phi)}\right] \end{aligned} \tag{15}$$

with θ given by (10). The results of (15) and (11) lead to

$$\begin{aligned} f(z_2 \mid \Delta=1, \tilde{z}_2) &\simeq \exp\left(\frac{a(\phi)\beta z_2 - b(\theta + a(\phi)\beta) + b(\theta)}{a(\phi)}\right) f(z_2 \mid \Delta=0, \tilde{z}_2) \\ &= \exp\left(\frac{[\theta + a(\phi)\beta]z_2 - b[\theta + a(\phi)\beta] + c(z_2)}{a(\phi)}\right) \end{aligned} \tag{16}$$

Thus, under the rare disease assumption, the distributions of Z_2, given $\Delta=0$ or $=1$, differ by a shift of $a(\phi)\beta$ on the scale of the canonical link, i.e. a shift of $\sigma^2\beta$ for a linear model, a shift of β on the logit scale for a logistic model, or a shift of β on the log scale for a log-linear model, where β is the coefficient associated with Z_2 in the proportional hazards model of interest.

The building of the prediction model, particularly the choice of the variable(s), $\tilde{z}_2$, is crucial to perform multiple imputation. Practically, in addition to the status indicator and stratification variables, it is necessary to adjust for the confounding variables included in the Cox model and for other predictive variables, which could be available.

The analyses were performed with R software (version 2.9.0, The R Foundation for Statistical Computing), using the mice (Multivariate Imputation by Chained Equations) package http://web.inter.nl.net/users/S.van.Buuren/mi/hmtl/mice.htm), which generates multiple imputations, and the *survival* package, which enables case–cohort designs to be carried out and analyzed by means of weighted estimators.

4. Validation of the method

To validate the method, first, we used entirely simulated data and compared the estimates and their standard errors to the true values. Then, using the PRIME cohort data, for which no strong surrogate for the chosen phase-2 variable was available, and the NWTS cohort data, which had an available surrogate, we compared the multiple imputation estimator to three weighted estimators: inverse probability weights (the most popular one), calibrated weights and re-estimated weights.

Copyright © 2011 John Wiley & Sons, Ltd.

4.1. Simulations

4.1.1. Completely simulated data. Two phase-1 variables were simulated: a binary variable, Z_1, and a Gaussian variable, Z_3, observed in the entire cohort. Also simulated was a phase-2 standard Gaussian variable, Z_2, which was independent of Z_1, but had a correlation of 0.2 with Z_3. The time to the event of interest had an exponential distribution, with $\lambda = \exp(\beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3)$. β_1, β_2, and β_3 were fixed at the same value of 0 or log(2). The censoring time followed a uniform distribution over the interval $[0, \tau]$, where τ was defined so that the probability of the event was approximately 0.01 ($\tau = 0.008$). The cohort size was 25 000 subjects. We also simulated a phase-1 variable predictive of the variable Z_2, $\widetilde{Z}_2 = Z_2 + \varepsilon$ with $\varepsilon \sim N(0, 1)$ independent of Z_2 (the correlation between Z_2 and $\widetilde{Z}_2$ was $\sqrt{2}/2 \simeq 0.7$).

We wanted to estimate the effect of Z_2 on the occurrence of the event, adjusting for Z_3 within the framework of stratified sampling of the subcohort. The cohort was divided into nine strata based on $\widetilde{Z}_2$ and Z_3 tertiles, and the controls were chosen by stratified sampling. Case–cohort sampling was simulated with 1000 subjects in each subcohort (Table I).

We built a linear prediction model for Z_2 based on the stratum indicators (phase-1 variables) and the status indicator. Z_3 was not directly included in the imputation model to predict Z_2, because it was used to define the strata, included in the model, and weakly correlated with Z_2.

The imputation model was

$$Z_2 = \alpha_0 + \alpha_1 \text{status} + \alpha_2' I_{\text{strata}} + e$$

where I_{strata} is the vector of stratum indicators. The mean multiple R^2 was 0.40.

Five plausible sets of complete data were generated for each cohort. Using weighted analysis, the variance estimator II proposed by Borgan *et al.* [6] was used. One thousand cohorts were simulated.

To assess the consequences of a misspecification of the imputation model, we compared the multiple imputation estimates, obtained when a predictive variable was omitted or included, to weighted estimates. Two scenarios were considered. In the first, the omitted variable was a confounder. Data were simulated according to the previously used conditions. The cohort was stratified only according to $\widetilde{Z}_2$ tertiles, ignoring Z_3. We used two imputation models, including the status and strata indicators with or without Z_3. The same two models were used in the simple imputation stage of the calibrated weighted analysis. For the second scenario, the omitted variable was not a confounder. Data were still simulated according to the previous conditions. The subcohort was selected by simple random sampling so that the imputation model included no strata indicator; $\widetilde{Z}_2$, which was related to the phase-2 variable (correlation $\rho = 0.7$), was omitted or included in the model.

4.1.2. Results. In Table II, for phase-1 (Z_1 and Z_3) and phase-2 (Z_2) variables, all estimates of the log relative risks were unbiased. Likewise, the multiple imputation variance and Borgan's variance estimator II (BII) agreed with the observed dispersion of estimates. For phase-1 variables, this observed dispersion was close to those of the entire cohort and multiple imputation but larger with the weighted estimator. For the estimated effect of the phase-2 variable, the observed standard errors were obviously larger for case–cohort than full cohort analyses, but they were slightly smaller with multiple imputation than with weighted estimators. For phase-1 variables, the relative increases of standard errors for the

Table I. Distribution of subcohort by stratum.

		Correlation $(Z_2, \widetilde{Z}_2) = 0.7$	
Stratum	Subcohort size	Cohort size*	Cases*
Tertile 1 $\widetilde{Z}_2$, tertile 1 Z_3	80	3241	7
Tertile 1 $\widetilde{Z}_2$, tertile 2 Z_3	80	2768	14
Tertile 2 $\widetilde{Z}_2$, tertile 1 Z_3	80	2765	10
Tertile 3 $\widetilde{Z}_2$, tertile 1 Z_3	80	2327	15
Tertile 1 $\widetilde{Z}_2$, tertile 3 Z_3	100	2323	27
Tertile 2 $\widetilde{Z}_2$, tertile 2 Z_3	100	2802	23
Tertile 2 $\widetilde{Z}_2$, tertile 3 Z_3	150	2766	55
Tertile 3 $\widetilde{Z}_2$, tertile 2 Z_3	150	2762	38
Tertile 3 $\widetilde{Z}_2$, tertile 3 Z_3	180	3244	117
Total	1000	25 000	308

*Rounded mean of the 1000 replications.

Copyright © 2011 John Wiley & Sons, Ltd.

Table II. Parameter estimates, stratified sampling of the subcohort (mean results from 1000 simulations).

	Correlation $(Z_2, \widetilde{Z}_2)=0.7$				
Parameter	Est	$\widehat{SE}$	SE	PC	Ratio
$\beta_1=0$					
Cohort	0.0143	0.1553	0.1568	95.0	
IM	0.0144	0.1553	0.1568	95.0	
BII	0.0151	0.1713	0.1763	95.3	1.10
$\beta_2=0$					
Cohort	0.0002	0.0618	0.0613	95.4	
IM	0.0004	0.0667	0.0681	94.2	
BII	0.0014	0.0703	0.0701	95.1	1.05
$\beta_3=0$					
Cohort	0.0022	0.0606	0.0624	93.9	
IM	0.0021	0.0609	0.0627	93.4	
BII	0.0015	0.0658	0.0682	94.4	1.08
$\beta_1=0.6931$					
Cohort	0.7133	0.1737	0.1744	95.2	
IM	0.7016	0.1742	0.1747	95.1	
BII	0.7177	0.2011	0.2011	95.7	1.15
$\beta_2=0.6931$					
Cohort	0.6940	0.0588	0.0589	95.0	
IM	0.6890	0.0707	0.0718	94.6	
BII	0.7040	0.0844	0.0835	95.8	1.19
$\beta_3=0.6931$					
Cohort	0.6955	0.0576	0.0621	92.7	
IM	0.6831	0.0601	0.0656	91.9	
BII	0.7069	0.0824	0.0894	94.1	1.37

Est, mean of the estimates; $\widehat{SE}$, mean of the standard error estimates; SE, standard error of the estimates; PC, Per cent coverage; Ratio, weighted to multiple imputation SE estimator; BII, Borgan's variance estimator II. IM, imputation model: $Z_{2i}=\alpha_0+\alpha_1 \mathrm{Ind}_{\mathrm{cas_}i}+\alpha_2 \mathrm{Strata}+e_i$.

weighted analysis estimators compared to those of multiple imputation, ranged from 8 to 37 per cent, whereas they were slightly smaller (5 or 19 per cent) for phase-2 variables.

Table III gives the results obtained with the calibrated weighted estimator and the multiple imputation estimator using a correct imputation model and a misspecified imputation model omitting a confounder variable. All calibrated weighted estimates were unbiased, for phase-1 and phase-2 variables, using a correct or a misspecified simple imputation model. All multiple imputation estimates were unbiased when a correct imputation model was used. However, the misspecified model yielded biased multiple imputation estimates for the phase-2 variable and the omitted phase-1 variable. All multiple imputation estimates were more precise than calibrated weighted estimates, for phase-1 and phase-2 variables effects. The relative increase of the standard error of the calibrated weighted estimate, as compared to multiple imputation estimate, exceeded 10 per cent.

Table IV reports the results obtained with the weighted estimator, a correct imputation model and a misspecified imputation model omitting a predictive variable that was not a confounder. All the estimates of phase-1 variable effects were unbiased. The precision of the two multiple imputation estimates, obtained with the correct or the misspecified models, were similar. The weighted estimate was less precise than multiple imputation estimates, more specifically, the relative increase of the standard error for the weighted estimator compared to multiple imputation estimators was 20 per cent for Z_1 and over 46 per cent for Z_3.

For the phase-2 variable, the multiple imputation estimate with the correct imputation model was unbiased, while the misspecified imputation model and the weighted analysis led to estimates slightly biased in opposite directions, respectively, 2.3 and 1.9 per cent. As expected, ignoring Z_3 in the imputation model decreased the precision of the multiple imputation estimator. The relative increase of the standard error with the misspecified model compared to that with the correct model was 7.6 per cent. The standard error of the weighted estimator was larger than that of the multiple imputation estimator. The relative increase of the standard error of the weighted estimator, compared to multiple

Copyright © 2011 John Wiley & Sons, Ltd.

Table III. Consequences of a misspecification of the imputation model. Stratified sampling of the subcohort. Results from 1000 simulations.

	Cohort	IM1	IM2	Calibrated1	Calibrated2
Parameter	β $(\widehat{SE})$	β $(\widehat{SE})$	β $(\widehat{SE})$	β $(\widehat{SE})$	β $(\widehat{SE})$
$\beta_1=0.6931$	0.7133 (0.1737)	0.7017 (0.1742)	0.6947 (0.1743)	0.7126 (0.1901)	0.7123 (0.1933)
$\beta_2=0.6931$	0.6940 (0.0588)	0.6867 (0.0718)	0.7554 (0.0694)	0.6921 (0.0853)	0.6902 (0.0865)
$\beta_3=0.6931$	0.6955 (0.0576)	0.6911 (0.0596)	0.7543 (0.0573)	0.6975 (0.0702)	0.6970 (0.0772)

Mean of the 1000 estimates (mean of the 1000 standard error estimates).
IM1: $Z_{2i}=\alpha_0+\alpha_1\text{Ind}_{\text{cas}}+\alpha_2\text{Strata}+\alpha_3 Z_3+e_i$ (correct model).
IM2: $Z_{2i}=\alpha_0+\alpha_1\text{Ind}_{\text{cas}}+\alpha_2\text{Strata}+e_i$ (misspecified model).
Calibrated1: $Z_{2i}=\alpha_0+\alpha_1\text{Ind}_{\text{cas}}+\alpha_2\text{Strata}+\alpha_3 Z_3+e_i$ (correct model).
Calibrated2: $Z_{2i}=\alpha_0+\alpha_1\text{Ind}_{\text{cas}}+\alpha_2\text{Strata}+e_i$ (misspecified model).

Table IV. Simple random sampling of the subcohort. Parameter estimates with correct and misspecified imputation model. Mean results from 1000 simulations.

	Correlation $(Z_2, \widetilde{Z}_2)=0.7$		
Parameter	Est	$\widehat{SE}$	SE
$\beta_1=0.6931$			
Cohort	0.7133	0.1737	0.1744
IM1	0.7045	0.1741	0.1745
IM2	0.6969	0.1744	0.1742
Weighted	0.7120	0.2088	0.2032
$\beta_2=0.6931$			
Cohort	0.6940	0.0588	0.0589
IM1	0.6853	0.0697	0.0706
IM2	0.6770	0.0750	0.0764
Weighted	0.7066	0.0909	0.0928
$\beta_3=0.6931$			
Cohort	0.6955	0.0576	0.0621
IM1	0.6934	0.0592	0.0632
IM2	0.6911	0.0603	0.0642
Weighted	0.7113	0.0881	0.1011

Est, mean of the estimates; $\widehat{SE}$, mean of the standard error estimates; SE, standard error of the estimates.
IM1: $Z_{2i}=\alpha_0+\alpha_1\text{Ind}_{\text{cas_}i}+\alpha_2\widetilde{Z}_2+\alpha_3 Z_3+e_i$ (correct model).
IM2: $Z_{2i}=\alpha_0+\alpha_1\text{Ind}_{\text{cas_}i}+\alpha_3 Z_3+e_i$ (misspecified model).

imputation estimator, was 30 per cent with the correct model and 21 per cent with the misspecified model.

4.2. *PRIME data*

4.2.1. Description of the data. The PRIME survey [14] was a multicenter cohort investigation studying risk factors of ischemic heart disease (IHD) and other cardiovascular end points. Among the 9 520 male subjects, 642 (6.7 per cent) experienced a cardiovascular event. The median follow-up time was 10 years and for those who suffered an event, the median time to its occurrence was 5 years.

We chose to estimate the effect of fibrinogen (phase-2 variable) on the occurrence of the event, adjusting for age, center, total cholesterol (CH), high-density lipoprotein cholesterol (HDL), systolic blood pressure (SBP), and tobacco use. Case–cohort data were simulated based on the entire PRIME cohort, and the results obtained with the different estimators were compared to each other and to the results obtained from the full cohort. The validation procedure used 1000 simulated subcohorts of size 2100.

Copyright © 2011 John Wiley & Sons, Ltd.

4.2.2. Sampling of the subcohort. Fibrinogen is a protein whose circulating concentration increases during inflammatory conditions. It also plays an important role in normal and pathological blood coagulation, and elevated fibrinogen levels are associated with a higher risk of cardiovascular events. Because smoking is one of the main factors determining the fibrinogen level [15], the cohort was stratified according to tobacco use, treated as a phase-1 variable available for the entire cohort. The strata were created as follows: stratum 1: non-smokers; stratum 2: ex-smokers; stratum 3: smokers (1–9 g/day); stratum 4: smokers (10–19 g/day); and stratum 5: smokers ($\geqslant$20 g/day). Stratified sampling was used to select subcohorts and the number of events in each stratum is given in Table V. The sampling probabilities were approximately 0.16, 0.20, 0.21, 0.25, and 0.46, respectively, for strata 1–5.

The linear model for fibrinogen imputation used as explanatory variables the status indicator and the variables included in the Cox model: tobacco, age, center, CH, HDL, and SBP (mean multiple R^2: 0.07). Five imputations were performed for each subcohort.

4.2.3. Results. The multiple imputation estimator was compared to the standard weighted estimator [6], calibrated weights estimator [3], and re-estimated weights estimator [3]. The mean of the 1000 log relative risk estimates, corresponding to the 1000 subcohorts, are given in Table VI. The median fraction of missing information about the fibrinogen effect was 0.22, so five imputations can be considered sufficient. For the phase-2 variable, the estimates were similar but, as expected, the standard error was larger in the case–cohort analysis than the full cohort analysis. Similar mean standard errors were obtained with multiple imputation and standard weighted estimation (BII), they were very close to those obtained with calibrated and re-estimated weights. For phase-1 variables, the multiple imputation estimates were more precise than those obtained by weighted, calibrated, and re-estimated weights analyses, and were nearly the same as those obtained from the entire cohort.

Table V. Distribution of cases by stratum.

Stratum	Cases (per cent)	Stratum size	Subcohort size
1 Non-smokers	153 (5.3)	2890	475
2 Ex-smokers	261 (6.4)	4078	800
3 Smokers 1–9 g/day	51 (6.3)	816	175
4 Smokers 10–19 g/day	55 (7.8)	707	175
5 Smokers $\geqslant$ 20 g/day	122 (11.9)	1029	475
Total	642 (6.7)	9520	2100

Table VI. Estimates of the log relative risks.

			Weighted estimator		
Variable	Whole cohort* β (SE)	Multiple imputation† β (SE‡)	Standard† β (SE‡)	Calibrated† β (SE‡)	Re-estimated† β (SE‡)
Fibrinogen	0.1312	0.1381	0.1346	0.1338	0.1337
	(0.0346)	(0.0421)	(0.0425)	(0.0414)	(0.0413)
Tobacco	0.0219	0.0218	0.0220	0.0214	0.0217
	(0.0036)	(0.0037)	(0.0045)	(0.0040)	(0.0039)
Age	0.0573	0.0566	0.0573	0.0566	0.0570
	(0.0138)	(0.0138)	(0.0160)	(0.0162)	(0.0160)
Center	0.2788	0.2764	0.2795	0.2749	0.2770
	(0.0857)	(0.0863)	(0.1006)	(0.1001)	(0.0990)
CH	0.0078	0.0077	0.0078	0.0078	0.0078
	(0.0010)	(0.0010)	(0.0012)	(0.0012)	(0.0012)
HDL	−0.0526	−0.0528	−0.0528	−0.0529	−0.0529
	(0.0070)	(0.0070)	(0.0080)	(0.0084)	(0.0082)
SBP	0.0104	0.0104	0.0105	0.0105	0.0105
	(0.0016)	(0.0016)	(0.0020)	(0.0019)	(0.0019)

*Unique estimates provided by the PRIME cohort.
†Mean estimations of the 1000 subcohorts.
‡Asymptotic standard error (SE) of the estimate.
FN, fibrinogen; CH, cholesterol; HDL, high-density lipoprotein; SBP, systolic blood pressure.
Imputation model: $\mathrm{FN} = \beta_0 + \beta_1 \mathrm{status} + \beta_2 \mathrm{tobacco} + \beta_3 \mathrm{age} + \beta_4 \mathrm{center} + \beta_5 \mathrm{CH} + \beta_6 \mathrm{HDL} + \beta_7 \mathrm{SBP} + \varepsilon$.

Copyright © 2011 John Wiley & Sons, Ltd.

Table VII. Distribution of central histology among cases and controls in the sampled strata.

	Controls			Cases		
	Central histology			Central histology		
Stratum	Favorable	Unfavorable	Subcohort fraction	Favorable	Unfavorable	Subcohort fraction
1	450	2	0.27	53	4	1
2	1569	51	0.10	216	16	1
4	897	17	0.13	188	20	1

4.3. NWTS data

The NWTS cohort [16] consisted of 3915 patients with Wilms' tumor diagnosed during 1989–1994 and followed until the earliest sign of disease progression or death for event-free survival. Baseline covariates included stage (I–IV), age at diagnosis, tumor diameter, and two binary histological evaluations (favorable vs unfavorable): the local hospital histology and central histology evaluated in a centralized reference laboratory. The former was strongly predictive of the latter (specificity 98 per cent, sensitivity 74 per cent) and both were available for all the patients. However, like Breslow *et al.* [3], we simulated case–cohort studies using central histology as the phase-2 variable. For the NWTS analysis, the Cox model included central histology (phase-2 variable), age as a piecewise linear variable with change point at 1 year, stage III/IV vs I/II, tumor diameter and the interactions local histology*age and stage*diameter. Breslow *et al.* [3] defined 16 strata according to event-free survival, stage (I/II or III/IV), favorable local histology (Yes or No), and age <1 year (Yes or No). They sampled controls from only three strata, defined by favorable local histology, and 'age <1 year + stage I/II' ($n=120$), 'age $\geqslant 1+$ stage I/II' ($n=160$), and 'age $\geqslant 1$ + stage III/IV' ($n=120$), while including all the subjects in the 13 other strata. They predicted unfavorable central histology, according to local histology, stage, age, tumor diameter, and the interaction local histology*stage.

These data present two specific features: first, a phase-1 variable, strongly predictive of the phase-2 variable, is available and, second, an interaction exists between the phase-2 variable and a phase-1 variable. Among these real data, in the sampled strata (all with favorable local histology) only a few controls had unfavorable central histology, especially in stratum 1 ($n=2$) and stratum 4 ($n=17$) (Table VII). Thus, for some subcohorts, the specific imputation model for these strata could not be estimated because of observed infinite odds ratio. Moreover, even when the subcohorts included some controls with unfavorable central histology in these strata, their number was necessarily small, and the estimator of the imputation model parameters might not have been distributed as assumed according to the asymptotic results.

Because of both particularities of these data, we considered two imputation models. The first, based on the proportional hazards model of interest and local histology, included the status indicator, local histology, stage, age, tumor diameter, and the interaction local histology*stage; the second was limited to the status indicator and local histology. When the imputation model included the interaction local histology*age, biased estimates and large standard errors were observed due to the small number of unfavorable central histologies in some subgroups used for the estimation of the interaction terms (Table VIII). For the imputation model including only the status indicator and the surrogate variable, the estimates were slightly biased. The simple effect of central histology, which represented the expected difference at age 1 between children with favorable or unfavorable central histology, differed slightly from the effect estimated for the full cohort (respectively, 4.11 vs 4.04); the same held true for the age effect after age 1 for the children with favorable central histology (respectively, 0.10 vs 0.11), and for the age effect after age 1 for the children with unfavorable histology (respectively, -3.63 vs -3.30). With the second imputation model, including only local histology and status indicator, we observed a small bias and good precision: although the multiple imputation was slightly biased, in particular for the age*histology interaction, it was always more precise than the standard weighted estimator and slightly more precise than the calibrated and re-estimated weighted estimators.

5. Discussion

The aim of the weighted analysis and of multiple imputation is to reconstitute the whole cohort. The former weights the subjects in the case–cohort sample by the inverse of the probability of being

Copyright © 2011 John Wiley & Sons, Ltd.

Table VIII. Mean results from 1000 simulated phase-2 samples based on the NWTS data.

		Multiple imputation		Weighted estimator		
Parameter	Whole cohort* β (SE)	Model 1† β (SE§)	Model 2‡ β (SE§)	Standard β (SE§)	Calibrated β (SE§)	Re-estimated β (SE§)
UCH	4.042	3.596	4.106	4.046	4.046	4.050
	(0.413)	(0.555)	(0.469)	(0.537)	(0.520)	(0.518)
Age0	−0.661	−0.702	−0.537	−0.669	−0.663	−0.676
	(0.326)	(0.329)	(0.329)	(0.359)	(0.324)	(0.324)
Age1	0.104	0.102	0.106	0.106	0.104	0.107
	(0.017)	(0.017)	(0.016)	(0.026)	(0.017)	(0.017)
Stage	1.346	1.441	1.353	1.344	1.345	1.344
	(0.244)	(0.257)	(0.251)	(0.346)	(0.273)	(0.272)
Diameter	0.069	0.073	0.072	0.070	0.070	0.070
	(0.014)	(0.014)	(0.014)	(0.021)	(0.015)	(0.015)
Stage*diameter	−0.076	−0.082	−0.083	−0.076	−0.076	−0.076
	(0.019)	(0.020)	(0.020)	(0.029)	(0.021)	(0.021)
UCH*age0	−2.635	−2.239	−3.097	−2.648	−2.655	−2.651
	(0.464)	(0.611)	(0.525)	(0.612)	(0.592)	(0.590)
UCH*age1	−0.058	−0.041	−0.065	−0.051	−0.050	−0.052
	(0.034)	(0.041)	(0.040)	(0.051)	(0.050)	(0.048)

*Unique estimates provided by the NWTS cohort.

†Imputation model for unfavorable central histology: status indicator, local histology, stage, age, tumor, diameter, and the interaction local histology*stage.

‡Imputation model for unfavorable central histology: status indicator and local histology.

§Asymptotic standard error (SE) of the estimate.

UCH, binary indicator of unfavorable central histology; Age0 and Age1, piecewise linear terms for age at diagnosis (years) before and after 1 year; stage, binary indicator of stage III/IV disease; diameter, diameter (cm) of excised tumor; SE, standard error.

observed during phase-2, but the phase-1 data observed for the other subjects are generally ignored. Alternatively, Breslow *et al.* [3] proposed two approaches using all the phase-1 information. Multiple imputation uses all the available data supplementing them with plausible values agreeing with what is observed in the complete data.

A key aspect of multiple imputation is the construction of the prediction model. It is necessary to reproduce correctly the relationship between the outcome and the incomplete variables, adjusting for the confounders included in the Cox model. With case–cohort data, the problem is complicated by the censoring process. One might think that it useful to include censoring time in the imputation model because, when the phase-2 variable is predictive of the event, its distribution among controls might not be the same at the beginning and the end of follow-up. However, in section 3, we demonstrated that, for a phase-2 variable with an exponential family distribution, a generalized linear model, including the case-status indicator as an exploratory variable, provides an approximately unbiased multiple imputation estimate. The proof relies on several assumptions: first, the studied event has to be rare; and second, the phase-2 variable has to be independent of the censoring time. The first assumption is precisely what justifies the use of a case–cohort design, while the second is required by the proportional hazards model. Thus, they do not represent new limitations. Our simulations showed that using all the complete subjects, cases as well as controls, and with a correct imputation model, including the case-status indicator, the multiple imputation estimator was unbiased. We also performed some simulations that confirmed no improvement of the multiple imputation estimator by adding the censoring time to the imputation model (data not shown). Confounding variables appearing in the analysis model also have to be included in the imputation model, if they remain predictive of the phase-2 variable, adjusting for the other predictors. On the other hand, misspecification of the imputation model can affect the phase-2 variable estimates but also the estimate of the omitted phase-1 variable. Omitting variables that improve the prediction can yield biased estimates and increase uncertainty about the parameter. Including variables that do not improve the prediction increases the uncertainty of the model coefficients and thus the between-imputation variance. Calibrated weighted estimates were found to be less sensitive to a misspecification of the imputation model than multiple imputation estimates concerning bias. However, multiple imputation estimates were more precise.

Copyright © 2011 John Wiley & Sons, Ltd.

The completely simulated data showed that, when the imputation model was correct, the multiple imputation approach provided unbiased and efficient estimators. Both the weighted and multiple imputation estimators were centered on the true values. An important result of this simulation study was that multiple imputation correctly estimated the variance of its estimators (just as BII correctly estimated the variance of the standard weighted estimator). This finding allowed us to compare the variance estimators of case–cohort data simulated from cohort surveys, for which the variance of the estimators cannot be compared to the true value because it is unknown. As expected, the multiple imputation approach was more precise than the usual weighted estimators for the parameters associated with phase-1 variables. The former also was slightly more precise than the latter for the phase-2 variable, despite the fact that it only used a categorized transformation of the explanatory variables (the stratum indicators used for stratified sampling). One explanation might be that multiple imputation approximates the maximum partial likelihood estimate, which is more efficient than weighted estimators. The simulations implying misspecified imputation models revealed, as expected, that the omission of a confounding variable from the imputation model had consequences in terms of bias and precision. We insist that it is essential to include the stratification variable(s) and the variables included in the analysis model in the imputation model. The consequences of omitting variable(s) related to the phase-2 variable, but not to the event of interest, mainly concern precision. In these simulations, the weighted estimators did not suffer from serious bias problems, not even the calibrated weighted estimator using a misspecified model in its imputation phase. However, they suffered from appreciable losses of precision. It should be underlined that most case–cohort surveys are generally performed to answer several scientific goals dealing with different diseases and exposures. Thus, it can be difficult to define a stratified sampling efficient for all the analyses and to optimize the weighted estimators. By contrast, multiple imputation can be adapted to each phase-2 variable and each substudy to improve the precision of the estimates of interest. When a misspecified imputation model was used, multiple imputation estimators for the phase-1 variable effects were not biased and their precision was similar to that obtained with a correct imputation model. For the phase-2 variable effect, slightly biased estimates were observed with the misspecified imputation model and weighted analyses. The effect of the misspecification on the imputation model was more noticeable in terms of precision. The loss of precision using a misspecified model as compared to a correct model was 7.6 per cent. This loss was greater for the weighted estimate than the multiple imputation estimate, and exceeded 21 per cent. We did not include the Z_3 variable, which was weakly correlated to the phase-2 variable. Results were less satisfactory when the correlation between the two variables was stronger, in particular concerning the precision of the weighted estimators (data not shown).

The case–cohort data simulated from the PRIME cohort study showed that multiple imputation can be used, even when no strong predictor of the phase-2 variable is available. Using the variables of the analysis model plus the case indicator in the imputation model, the multiple imputation estimator was more precise than the weighted estimators, particularly the standard estimator, for the effects of phase-1 variables. For the phase-2 variable, the multiple imputation estimator had the same precision as the standard weighted estimators and it was slightly less precise than calibrated and re-estimated weighted estimators.

The NWTS cohort data represent one of the worst possible situations for using multiple imputation. Inclusion of an interaction term between the phase-2 variable and the two age-effect components required an imputation model including similar interactions between the indicator status and the age effects. The corresponding coefficients had to be estimated in separate strata, with very low numbers of patients presenting unfavorable central histology or even no such patient at all. As a consequence, the maximum likelihood estimator of the imputation model could be expected to be biased [17, Chapter 8.4] and the multiple imputation estimator reflected this bias. Although local histology was strongly predictive of central histology, imputation model 2, using only the former and the case indicator, also yielded biased estimates of the Cox model parameters. The imputed values correctly reflected the global proportion of patients with unfavorable central histology, but not always the proportional difference between cases and controls, as was the case for the interaction terms age*local histology or stage*diameter.

It is reasonable to wonder how many imputations are needed. The number of requested imputations increases with the proportion of missing information. However, in case–cohort studies, the proportion of missing information is considerably smaller than the percentage of incompletely observed subjects, and a small number of imputations, 5–10, should suffice. With the PRIME data and using a small sampling rate (700/25 000), the results obtained with five imputations did not differ appreciably from those obtained with 10 (data not shown).

Copyright © 2011 John Wiley & Sons, Ltd.

Herein, we presented simulations with only one phase-2 variable. However, the approach can easily be extended to several phase-2 variables. If several covariates are incomplete, we suggest imputing them using a multivariate distribution, which takes into account the correlation structure between the covariates, for instance with the mice package, which generates multivariate imputations via Markov Chain Monte Carlo (MCMC) algorithm according to their joint distribution.

We focused on situations in which covariates were time-fixed. However, the multiple imputation approach can be extended to time-varying covariates using mixed models to account for the within-subject correlation structure [18].

When the phase-2 data allow consistent estimation of the imputation model, multiple imputation is an efficient technique to analyze the data from case–cohort surveys. For phase-1 variables, multiple imputation has better efficiency than weighted estimators, a valuable improvement for prediction studies or when the effect of some phase-1 variables is a focus of interest. A large number of imputations is not required to obtain good quality estimates. Software that simply implements this procedure is available: under R, the mice library; under Stata, the Imputation by Chained Equations library, or under SAS, the PROC MI, and PROC MIANALYZE.

To gain time and determine whether multiple imputation provides estimates similar to weighted analysis, we suggest building the analysis model using only the case–cohort data and weighted estimators. Multiple imputation can eventually be used to reanalyze the data with the selected final model to improve the precision of the results.

Acknowledgements

This study was supported by a grant from the Région Île-de-France. The authors are grateful to Pierre-Yves Scarabin and Pierre Ducimetière for providing the PRIME data.

References

1. Langholz B, Thomas DC. Nested case–control and case–cohort methods of sampling from a cohort: a critical comparison. *American Journal of Epidemiology* 1990; **131**:169–176.
2. Prentice RL. A case–cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* 1986; **73**:1–11.
3. Breslow NE, Lumley T, Ballantyne CM, Chambless LE, Kulich M. Using the whole cohort in the analysis of case–cohort data. *American Journal of Epidemiology* 2009; **169**:1398–1405. DOI: 10.1093/aje/kwp055.
4. Paik MC, Tsai WY. On using the Cox proportional hazards model with missing covariates. *Biometrika* 1997; **84**:579--593. DOI: 10.1093/biomet/84.3.579.
5. Self SG, Prentice RL. Asymptotic distribution theory and efficiency results for case–cohort studies. *The Annals of Statistics* 1988; **16**:64–81. DOI: 10.1214/aos/1176350691.
6. Borgan O, Langholz B, Samuelsen SO, Goldstein L, Pogoda J. Exposure stratified case–cohort designs. *Lifetime Data Analysis* 2000; **6**:39–58. DOI: 10.1023/A:1009661900674.
7. Barlow WE. Analysis of case–cohort designs. *Journal of Clinical Epidemiology* 1999; **52**:1165–1172. DOI: 10.1016/S0895-4356(99)00102-X.
8. Langholz B, Jiao J. Computational methods for case–cohort studies. *Computational Statistics and Data Analysis* 2007; **51**:3737–3748.
9. Kulich K, Lin DY. Improving de efficiency of relative-risk estimation in case–cohort studies. *Journal of the American Statistical Association* 2004; **99**:832–844. DOI: 10.1198/016214504000000584.
10. Qi L, Wang CY, Prentice RL. Weighted estimators for proportional hazards regression with missing covariates. *Journal of the American Statistical Association* 2005; **100**:1250–1263. DOI: 10.1198/016214505000000295.
11. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. Wiley: New York, 1987.
12. Rubin DB, Schenker N. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association* 1986; **81**:366–374.
13. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Wiley: New York, 1987.
14. Yarnell JWG. The PRIME study: classical risk factors do not explain the severalfold differences in risk of coronary heart disease between France and Northern Ireland. *The Quarterly Journal of Medicine* 1998; **91**:667–676.
15. Scarabin PY, Jiao J. Plasma fibrinogen explains much of the difference in risk of coronary heart disease between France and Northern Ireland. The PRIME study. *Atherosclerosis* 2003; **166**:103–109.
16. d'Angio GJ, Breslow N, Beckwith JB, Evans A, Baum H, deLorimier A, Fernbach D, Hrabovsky E, Jones B, Kelalis P, Othersen HB, Tefft M, Thomas PRM. Treatment of Wilms' tumor. Results of the Third National Wilms' Tumor Study. *Cancer* 1989; **64**:349–360.
17. Cox DR, Hinkley DV. *Theoretical Statistics*. Chapman and Hall: London, 1974.
18. Yucel RM. Multiple imputation inference for multivariate multilevel continuous data with ignorable non-response. *Philosophical Transactions of the Royal Society A* 2008; **366**:2389–2403. DOI: 10.1098/rsta.2008.0038.

Copyright © 2011 John Wiley & Sons, Ltd.

BMC Medical Research Methodology

BioMed Central
The Open Access Publisher

This Provisional PDF corresponds to the article as it appeared upon acceptance. Fully formatted PDF and full text (HTML) versions will be made available soon.

Multiple imputation for estimating hazard ratios and predictive abilities in case-cohort surveys

BMC Medical Research Methodology 2012, **12**:24 doi:10.1186/1471-2288-12-24

Helena Marti (helena.marti-soler@inserm.fr)
Laure Carcaillon (laure.carcaillon@inserm.fr)
Michel Chavance (michel.chavance@inserm.fr)

ISSN 1471-2288

Article type Research article

Submission date 29 June 2011

Acceptance date 9 March 2012

Publication date 9 March 2012

Article URL http://www.biomedcentral.com/1471-2288/12/24

Like all articles in BMC journals, this peer-reviewed article was published immediately upon acceptance. It can be downloaded, printed and distributed freely for any purposes (see copyright notice below).

Articles in BMC journals are listed in PubMed and archived at PubMed Central.

For information about publishing your research in BMC journals or any BioMed Central journal, go to

http://www.biomedcentral.com/info/authors/

© 2012 Marti *et al.* ; licensee BioMed Central Ltd.
This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Multiple imputation for estimating hazard ratios and predictive abilities in case-cohort surveys

Helena Marti,[Aff1]
Corresponding Affiliation: Aff1
Email: helena.marti-soler@inserm.fr
Laure Carcaillon,[Aff2]
Email: laure.carcaillon@inserm.fr
Michel Chavance,[Aff1]
Email: michel.chavance@inserm.fr

Aff1 Inserm, CESP Centre for Research in Epidemiology and Population Health, U1018, Biostatistics team, F-94807 Villejuif, France

Aff2 Inserm, CESP Centre for Research in Epidemiology and Population Health, U1018, Hormones and Cardiovascular Disease team, F-94807 Villejuif, France

Abstract

Background

The weighted estimators generally used for analyzing case-cohort studies are not fully efficient and naive estimates of the predictive ability of a model from case-cohort data depend on the subcohort size. However, case-cohort studies represent a special type of incomplete data, and methods for analyzing incomplete data should be appropriate, in particular multiple imputation (MI).

Methods

We performed simulations to validate the MI approach for estimating hazard ratios and the predictive ability of a model or of an additional variable in case-cohort surveys. As an illustration, we analyzed a case-cohort survey from the Three-City study to estimate the predictive ability of D-dimer plasma concentration on coronary heart disease (CHD) and on vascular dementia (VaD) risks.

Results

When the imputation model of the phase-2 variable was correctly specified, MI estimates of hazard ratios and predictive abilities were similar to those obtained with full data. When the imputation model was misspecified, MI could provide biased estimates of hazard ratios and predictive abilities. In the Three-City case-cohort study, elevated D-dimer levels increased the risk of VaD (hazard ratio for two consecutive tertiles = 1.69, 95%CI: 1.63–1.74). However, D-dimer levels did not improve the predictive ability of the model.

Conclusions

MI is a simple approach for analyzing case-cohort data and provides an easy evaluation of the predictive ability of a model or of an additional variable.

Background

Case-cohort surveys produce incomplete data by design. A subcohort is selected by simple or stratified random sampling, all subjects are followed up and the events of interest are recorded. The phase-1 variables are observed for the entire cohort, whilethe phase-2 variables are only known for the case-cohort sample, i.e., subjects belonging to the subcohort and all those presenting the event of interest [1]. Thus, in case-cohort studies, the non-cases who do not belong to the subcohort are incompletely observed by design, enabling cost reduction with a small loss of efficiency.

Various approaches have been described to estimate the proportional hazard model in case-cohort surveys: Weighted estimators [2-6] are classically used in these surveys, with analysis restricted to the completely observed subsample, so the information collected for incompletely observed non-cases is ignored and inefficient estimators for the effect of phase-1 variables are obtained. One of the most popular is the Borgan II estimator [4]. Scheike and Martinussen [7] proposed a maximum likelihood estimator based on proportional hazards assumption, using the EM algorithm [8], therebyincreasing efficiency as compared to weighted estimators when the relative risk and disease incidence are high. However, in general, the studied disease incidence in case-cohort surveys is low. Breslow et al. [9] suggested calibrating or estimating the weights a posteriori, using all the phase-1 information, to improve precision with respect to classical weighted estimators. Marti and Chavance [10] showed that multiple imputation (MI) is a good alternative to classical weighted methods for the analysis of case-cohort data. When the imputation model was correct, the MI approach provided unbiased estimators of the log hazard ratios and correctly estimated the variance of its estimators. As expected, the MI approach was more precise than the usual weighted estimators for the parameters associated with phase-1 variables. The former was also slightly more precise than the latter for the phase-2 variable. In Marti and Chavance [10] the imputations were performed according to a correctly specified imputation model. However, in practise, the distribution of the phase-2 variable is unknown and onemay wonder how MI compares to weighted estimators when the imputation model is misspecified.

No standard method exists for quantifying the usefulness or predictive ability of a model or an additional variable in the framework of case-cohort surveys. The predictive ability can be measured in terms of calibration, which refers to the ability of a model to match predicted and observed values, when we are interested in individual predictions; or in terms of discrimination, which refers to the ability of a model to distinguish between subjects with or without a binary event, when we are interested in identifying a group of high-risk subjects. In the present work, we focus on discrimination.

As shown below, a naive measurement of predictive ability from case-cohort data often leads to a biased estimate of the predictive ability because it varies with the censoring rate and thus depends on the subcohort size. Alternatively, because MI reconstitutes whole cohorts, any tool developed to estimate the predictive ability in the framework of cohort surveys can be

applied to case-cohort data, so we propose using the MI approach to estimate the predictive ability ofa model or of an additional variable and their standard errors.

The objectives of this study were 1) to evaluate MI for estimating hazard ratios when the distribution of the phase-2 variable is misspecified; and 2) to present an adequate methodology for estimating the predictive ability of a model or of an additional variable in case-cohort surveys. We performed a simulation study to validate the MI approach for estimating the predictive ability of a model or of an additional variable and to assess its potential limits. As an illustration, we analyzed case-cohort data from the Three-City study [11] to estimate the predictive ability of the D-dimer plasma concentration, a marker of coagulation and fibrinolysis, on coronary heart disease (CHD) and on vascular dementia (VaD) risks.

Methods

Incomplete observations and multiple imputation

Case-cohort surveys are a particular type of incomplete observations, in which data are missing at random [12] by design, as the probability of being completely observed depends only on the case status, with simple random sampling, and on some phase-1 variables with stratified sampling. MI is a simple and efficient method for analyzing incomplete observations, while taking into account all the levels of uncertainty regarding missing values. This provides an approximation of the maximumlikelihood estimator and thus enables the potential selection bias to be corrected. This method relies on the generation of several plausibly completed data sets ($M \geq 2$), accounting for all levels of uncertainty concerning the missing values. A prediction model must be built, taking into consideration the relationships between the incomplete variable and the other variables, as observed in the complete part of the data. The missing data are not replaced by their expectation but by a value drawnfrom the distribution posited by the model. To take into account the uncertainty concerning the parameters of the imputation model, several imputations are performed with parameters drawn from the asymptotic distribution of their estimator. An estimate of the parameter of interest, $\hat{\theta}_m$, $m = \{1,\ldots,M\}$, and an estimate of the variance of the estimator, $\hat{V}(\hat{\theta}_m)$, are obtained from each completed data set. If the imputation model is correct, these estimators are not biased. The MI estimate, also unbiased, is the mean of the M estimates:

$$\hat{\theta}_{MI} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m \tag{1}$$

where M is the number of completed data sets and $\hat{\theta}_m$, $m = \{1,\ldots,M\}$ is the estimate of the parameter of interest provided by the m^{th} completed data set. The multiplicity of imputations enables correct estimation of the variance of this single estimator, which is the sum of 2 components: the within-imputations component, W_{MI}, and the between-imputations component, B_{MI}:

$$\begin{aligned}\hat{V}(\hat{\theta}_{MI}) &= \hat{W}_{MI} + \hat{B}_{MI} \\ &= \frac{1}{M}\sum_{m=1}^{M}\hat{V}(\hat{\theta}_m) + (1+M^{-1})\frac{\sum_{m=1}^{M}(\hat{\theta}_m - \hat{\theta}_{MI})(\hat{\theta}_m - \hat{\theta}_{MI})'}{M-1}\end{aligned} \tag{2}$$

where the factor $1+M^{-1}$ is an adjustment for using a finite number of imputations [13].

MI requires a model correctly reflecting the relationship between the incomplete variable and the outcome of interest. In case-cohort surveys, we need to impute phase-2 variable values for the non-cases who do not belong to the subcohort. Under the rare disease assumption, we have shown that a simple generalized linear model, using all the complete data (cases and non-cases) and including the case indicator among the explanatory variables, has to be considered [10]. Practically, in addition to the case indicator and the stratification variables, when the subcohort was selected by stratified sampling, it is necessary to include in the imputation model all the variables appearing in the proportional hazard model. Because imputations are based on asymptotic distributions, caution is necessary, since if too few subjects present the event of interest, the distribution of the estimators can differ from the asymptotic one. As a consequence, the maximum likelihood estimator of the imputation model could be biased or not normally distributed.

Predictive ability of a model and of a supplementary variable

Harrell et al. [14] proposed the C index to measure the predictive ability of a model in cohort studies as the agreement between the order of the predicted and observed survival times in any pair of subjects (the event of interest is assumed to be death, leading to the use of survival terminology). That is, the concordance probability using all pairs of subjects in the population. However, with censored data, it is not possible to consider all the pairs of subjects because survival time is not observed for censored subjects. Let T_i be the survival time for subject i, i = 1,…,N, where N is the cohort size, and C_i the censoring time for subject i. We observe $X_i = \min(T_i, C_i)$. Usable pairs are those for which the order of the predicted survival times can be compared to the order of the true survival times, i.e., pairs formed by 2 uncensored subjects or an uncensored subject and a subject censored after the uncensored subject's death. A pair of censored subjects carries no information about its agreement with the expected survival provided by the model since the order of the survival times is not known. Similarly a pair formed by a subject whose survival time is observed and a subject censored before this survival time provides no information on this agreement since the unknown survival time could be anterior or posterior to the observed one. Harrell et al. [15] showed that, in the common models used for survival analysis, such as the proportional hazard model, the predicted survival times and the predicted survival probabilities at a fixed time t can be interchanged for the comparison. The Harrell's C index is defined as:

$$C = \frac{\pi_c}{\pi_c + \pi_d} \tag{3}$$

where π_c is the probability of concordance for a pair (i,j) and π_d is the probability of discordance. We assume continuous survival times and continuous predicted survival probabilities, so $P(X_i = X_j) = P(Y_i = Y_j) = 0$, thus $\pi_c + \pi_d = 1$. C is estimated by the proportion of concordant pairs among the usable pairs. The estimated variance was given by Kremers [16].

In practice, we are often interested in estimating the predictive ability of an additional phase-2 variable. Let M_1 be a proportional hazard model including only phase-1 variables, and C_1 and SE_{C1} respectively the C indexof M_1 and its standard error. Let M_2 be a proportional hazard model adding the phase-2 variable to M_1, and C_2 and SE_{C2}, respectively, the C index

of M_2 and its standard error. Harrell's predictive ability ofthe added phase-2 variable is $\Delta = C_2 - C_1$. Complementary measures of predictive ability of a new variable, such as the net reclassification improvement (NRI) and the integrated discrimination index (IDI), were proposed by Pencina [17]. NRI needs some a priori meaningful risk categories. It quantifies the correct reclassification introduced by using a model with the added variable as compared to the classification obtained without this variable. The IDI can be viewed as a continuous version of the NRI with probabilities used instead of categories. It can be defined as the discrimination-slope difference between the models with and without a quantitative variable. To estimate the predictive ability of a model or of an additional variable, we reconstructed plausible whole cohorts using MI. For each reconstructed whole cohort, we could then directly obtain C_1, SE_{C1}, C_2, SE_{C2}, Δ, NRI, IDI and their respective variances. Using equations (1) and (2), we obtained the MI estimates of these quantities. Concerning the variance of Δ, the between-imputation component is estimated by the empirical variance of the M estimates of Δ provided by the M completed data sets. However, for the within-imputation component, the asymptotic variance of the estimator provided by a complete data set, does not have an analytical form. With a fully observed cohort, bootstrapping is a way to estimate the variance of the corresponding Δ. Therefore, each whole cohort reconstructed by MI has to be resampled. In the simulations as in the real data analysis, we used 100 bootstrap samples.

Simulation study

Two phase-1 variables were simulated: a binary variable, Z_1, and a Gaussian variable, Z_3, observed for the entire cohort. For the phase-2 variable, Z_2, we considered three different distributions: normal, log-normal and uniform, all of them with unit variance, independent of Z_1, but having a correlation coefficient of 0.2 with Z_3. The survival time had an exponential distribution, with $\lambda = \exp(\beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3)$. β_1, β_2 and β_3 were fixed at the same value and set at 0 or log(2). The censoring time followed a uniform distribution over the interval $[0,\tau]$, where τ was chosen so that the probability of an event was approximately 0.03 τ(= 0.025). The cohort size was 10,000. We also simulated a phase-1 variable predictive of Z_2, $\tilde{Z}_2 \equiv Z_2 + \varepsilon$ with $\varepsilon \sim N(0, \sigma^2)$ independent of Z_2. The variance σ^2 was fixed at 1 which corresponds to a correlation between Z_2 and $\tilde{Z}_2$ of approximately 0.7. We wanted to estimate the effect of Z_2 on survival time and its predictive ability. The cohort was divided into 9 strata based on the tertiles of $\tilde{Z}_2$ and Z_3, and the non-cases were chosen by stratified sampling. Case-cohort sampling was simulated with 1,000 subjects in each subcohort. The phase-2 variable was not available for non-cases not included in the subcohort,so MI was used to complete the data set. Thus, we built the same linear prediction model for each Z_2 based on the stratification phase-1 variable and the case indicator. Z_3 was not directly included in the imputation model to predict Z_2,because it was a stratification variable included in the model and because of the weak correlation between Z_2 and Z_3. The imputation model was: $Z_2 = \alpha_0 + \alpha_1 I_{case} + \alpha_2 \text{Strata} + \varepsilon$, where α_0 and α_1 are scalar, α_2 is a vector coefficient, I_{case} is the case indicator, Strata is the vector of stratum indicators and ϵ is the vector of errors independently and identically distributed $\sim N(0, \sigma)$. Thus, the imputation model was correctly specified for Z_2 normally distributed but misspecified for Z_2 log-normally or uniformly distributed. One thousand cohorts were simulated for each scenario.

Five imputations were performed and 5 complete data sets were generated for each cohort. We estimated the log hazard ratios using MI and the "Borgan II" weighted estimator [4]. We

used MI to estimate the predictive ability of models with and without the phase-2 variable, and the predictive ability of the phase-2 variable, NRI and IDI. We also studied the consistency of the naive estimator of Harrell's C index in case-cohort surveys by varying the subcohort size. Using the above simulation conditions and, exceptionally, a scenario with $\beta_1 = \beta_3 = \log(2)$ and $\beta_2 = \log(1.5)$, we simulated case-cohort samples with the subcohort size set at 300 or 1,000 subjects. We estimated the predictive ability in the case-cohort samples and in the multiply imputed data sets.

Case-cohort survey from Three-City study

Briefly, the 3C-Study was designed to examine the relationship between vascular diseases and dementia in a community housing 9,294 persons aged 65 years and over between 1999 and 2001 in three French cities. The detailed methodology has been previously described [11]. A case-cohort substudy was conducted [18], to investigate the relationship between biomarkers, such as plasma levels of D-dimer (a marker of coagulation and fibrinolysis) and the 4-year incidence of coronary heart disease (CHD),stroke and all subtypes of dementia, including vascular dementia (VaD), in an elderly population. The phase-1 variables provided information on socio-demographic characteristics, education, medical history, diet, alcohol and tobacco use. Blood pressure, height and weight were also available. A subcohort of size n = 1,254, (13.5% of the full cohort) was randomly selected, stratifying on age, sex and recruitment center. Observed cumulated incidences of CHD and VaD were approximately 2% and 0.6%, respectively. Plasma D-dimer levels were only available for phase-2 subjects. Carcaillon et al. [18] treated quintiles of D-dimer level both qualitatively and linearly. They reported a linear increase in the risk of VaD according to D-dimer quintiles.

We re-assessed the relationship between plasma D-dimer levels and the risk of CHD and VaD, using MI and weighted estimators, and evaluated the predictive ability of D-dimer levels on both risks. We included the same explanatory variables as Carcaillon et al. [18] although we used tertiles of D-dimer rather than quintiles, to estimate CHD and VaD risks, due to the small number of events. Therefore, to estimate the risk of CHD, the proportional hazard model includedthe phase-1 variables: age, sex, center, body mass index, hypertension, hypercholesterolemia, diabetes, tobacco use, diabetes drugs, and as phase-2 variables, indicators of D-dimer tertiles. To estimate the risk of VaD, the proportional hazard model included the phase-1 variables: age, sex, centre, educational level, body mass index, the presence or absence of an apolipoprotein $\epsilon 4$ allele and indicators of D-dimer tertiles.

For each outcome (CHD or VaD), it was necessary to reproduce the relationships among the incomplete variable, the outcomes and the confounder variables. For each outcome, we built an imputation model of tertiles of D-dimer levels, including the variables used in the proportional hazard model and the case-indicator. We estimated the predictive ability of proportional hazard models, without (C_1) and with (C_2) D-dimer levels, $\Delta = C_2 - C_1$, and IDI for CHD and VaD risks. The NRI requires that some a priori meaningful risk categories be known. Based on the Third Adult Treatment Panel [ATP III] [19] risk classification for the 10-year risk of CHD, we adapted the cut-offs to 4-year risk. For VaD, we do not know a priori meaningful risk categories and did not compute NRI.

Results

Simulation study

The mean fraction of missing information about the effect of Z_2 ranged from 5 to 14 per cent when $\beta_2 = 0$ and from 23 to 30 per cent when $\beta_2 = \log(2)$ (data not shown). For each estimator (full cohort, case-cohort with MI and case-cohort with weights), we give the mean of the estimated coefficients, the mean of their standard error estimates, the observed standard error of the estimated coefficients and the mean squared errors of 1000 simulations (Table 1). Not surprisingly, the full cohort estimates and the case-cohort weighted estimates of the log hazard ratios were unbiased. Similarly, with a correctly specified normal imputation model, all MI estimates were unbiased. With a misspecified normal imputation model, MI estimate of the effect $\beta_2 = \log(2)$ of Z_2 was biased (−13%) when Z_2 was log-normally distributed. When Z_2 was uniformly distributed, MI estimate of the effect of Z_2 was slightly biased (−5%). With a misspecified normal imputation model and $\beta_2 = 0$, no bias was observed. The MI variance and the weighted estimator variance agreed with the observed dispersions of the estimates. The observed dispersion was always smaller with MI than with the weighted estimator. For the phase-1 variables, this dispersion was similar for the entire cohort and with MI, whatever the distribution of the phase-2 variable. For the estimated effect of the phase-2 variable, the observed standard deviations were smaller with MI than with the Borgan II weighted estimator but, as expected, slightly larger with MI than in the full cohort analyses. Altogether, the mean squared errors were smaller with MI than with the weighted estimator, except for the effect of the phase-2 variable with $\beta_2 = \log(2)$ and Z_2 was log-normally distributed.

Table 1 Mean of the log hazard ratio estimates (Est), mean of the standard error estimates $\hat{SE}$, standard error of the estimates (SE) and mean of the mean square error (MSE). Results of 1,000 simulations

	Full cohort				Multiple imputation[a]				Weighted estimator			
	Est	$\hat{SE}$	SE	MSE	Est	$\hat{SE}$	SE	MSE	Est	$\hat{SE}$	SE	MSE
Z_2 normally distributed												
$\beta_1 = \beta_2 = \beta_3 = 0$												
β_1	−0.003	0.107	0.100	0.010	−0.003	0.107	0.110	0.010	−0.001	0.133	0.128	0.016
β_2	−0.001	0.054	0.058	0.003	−0.001	0.060	0.062	0.004	0.001	0.065	0.068	0.005
β_3	−0.004	0.053	0.056	0.003	−0.004	0.054	0.057	0.003	−0.003	0.058	0.060	0.004
$\beta_1 = \beta_2 = \beta_3 = \log(2)$												
β_1	0.689	0.118	0.113	0.013	0.676	0.119	0.112	0.013	0.696	0.168	0.165	0.027
β_2	0.687	0.058	0.057	0.003	0.679	0.070	0.068	0.005	0.701	0.088	0.097	0.009
β_3	0.683	0.057	0.057	0.003	0.679	0.058	0.058	0.004	0.689	0.080	0.090	0.007
Z_2 log normally distributed												
$\beta_1 = \beta_2 = \beta_3 = 0$												
β_1	−0.003	0.107	0.100	0.010	−0.003	0.107	0.100	0.010	−0.004	0.133	0.128	0.016
β_2	−0.001	0.027	0.034	0.001	0.015	0.031	0.032	0.001	0.002	0.034	0.038	0.001
β_3	−0.004	0.053	0.056	0.003	−0.014	0.054	0.058	0.004	−0.005	0.059	0.062	0.004
$\beta_1 = \beta_2 = \beta_3 = \log(2)$												
β_1	0.686	0.058	0.056	0.003	0.621	0.061	0.055	0.008	0.686	0.112	0.117	0.014
β_2	0.692	0.013	0.015	$2e0^{-4}$	0.602	0.015	0.014	0.008	0.695	0.020	0.023	0.001
β_3	0.685	0.029	0.031	0.001	0.686	0.032	0.031	0.001	0.687	0.049	0.053	0.003
Z_2 uniformly distributed												
$\beta_1 = \beta_2 = \beta_3 = 0$												
β_1	0.007	0.181	0.175	0.031	0.007	0.181	0.175	0.031	0.007	0.197	0.188	0.035
β_2	−0.001	0.092	0.087	0.008	0.004	0.094	0.088	0.008	−0.002	0.098	0.095	0.009
β_3	0.003	0.090	0.090	0.008	0.002	0.090	0.090	0.008	0.004	0.093	0.093	0.009
$\beta_1 = \beta_2 = \beta_3 = \log(2)$												
β_1	0.690	0.120	0.116	0.013	0.680	0.121	0.115	0.013	0.694	0.166	0.169	0.028
β_2	0.695	0.069	0.063	0.004	0.656	0.075	0.066	0.006	0.698	0.087	0.082	0.007
β_3	0.690	0.058	0.054	0.003	0.689	0.059	0.055	0.003	0.698	0.081	0.081	0.007

[a] MI estimates with imputation model: $Z_2 = \alpha_0 + \alpha_1 \text{Ind}_{\text{case}} + \alpha_2 \text{Strata} + \varepsilon, \varepsilon \sim N(0, \sigma)$

The results concerning the consistency of the naive estimator of Harrell's C index are reported in Table 2. In the scenario $\beta_1 = \beta_2 = \beta_3 = 0$, the mean C index was nearly 0.5 for both models, without and with Z_2, whatever the analysis performed. In the scenarios $\beta_1 = \beta_3 =$ log(2) and $\beta_2 =$ log(1.5) or $\beta_2 =$ log(2), the naive computation of C with the case-cohort data led to lower predictive ability than with the full cohort, especially for the smaller subcohort. Bycontrast, the Harrell's C indexes estimated by MI were similar to those computed for the full cohort and did not depend on the subcohort size. The estimated dispersion of the C index was slightly greater than the observed dispersion of the estimates. The rejection percentage of the null hypothesis $\Delta = 0$ was always similar to the full cohort analysis and to MI. As a consequence of the standard error underestimation, the observed first type error rate was slightly lower than 5%. Nevertheless, in the considered scenarios, the observed power was very high. As expected, the loss of power when comparing case-cohort with MI to full cohort analysis was small: with $\beta_2 =$ log(1.5), the observed power was 84.6% with a subsample size of 300, and 90.6% with a subsample size of 1000 versus 91.6% with the full cohort. MI estimates of NRI and IDI indexes were close to those obtained with the full cohort analysis and did not depend on the subcohort size. As compared to the full cohort results, the rejection percentage of the null hypothesis NRI = 0 was smaller with MI analysis when $\beta_2 = 0$, larger when $\beta_2 =$ log(1.5) and similar when $\beta_2 =$ log(2). When the effect of the phase-2 variable was not null, the rejection percentage of the null hypothesis IDI = 0 was similar with MI and with full cohort analysis. By contrast, whatever the effect of the phase-2 variable, the estimation of NRI and IDI in the case-cohort sample provided larger measures of these indexes than the full cohort analysis.

Table 2 Mean of the predictive ability estimates (Est), mean of the standard error estimates $\hat{SE}$ and standard error of the estimates (SE). Results from 1000 simulations

	$\beta_1 = \beta_2 = \beta_3 = 0$				$\beta_1 = \beta_3 =$ log(2), $\beta_2 =$ log(1.5)				$\beta_1 = \beta_2 = \beta_3 = log(2)$			
	Est	$\hat{SE}$	SE	% H_0 rejected	Est	$\hat{SE}$	SE	% H_0 rejected	Est	$\hat{SE}$	SE	% H_0 rejected
Full Cohort												
C_1	0.518	0.033	0.012		0.727	0.032	0.015		0.733	0.029	0.014	
C_2	0.524	0.033	0.013		0.747	0.031	0.015		0.733	0.029	0.014	
$\underline{\Delta}$	0.006	0.010	0.009	3.7	0.020	0.007	0.007	91.6	0.049	0.010	0.010	100
NRI	0.007	0.017	0.019	4.8	0.071	0.030	0.033	52.5	0.167	0.034	0.035	99.9
IDI	$2e^{-4}$	$2e^{-4}$	$3e^{-4}$	6.0	0.014	0.003	0.005	99.9	0.048	0.006	0.009	99.9
MI1000												
C_1	0.518	0.033	0.012		0.724	0.032	0.016		0.733	0.029	0.014	
C_2	0.526	0.033	0.013		0.745	0.031	0.016		0.783	0.027	0.014	
Δ	0.008	0.012	0.010	3.4	0.021	0.008	0.008	90.6	0.049	0.010	0.011	100
NRI	0.009	0.019	0.017	1.5	0.076	0.033	0.033	64.8	0.172	0.037	0.036	100
IDI	$3e^{-4}$	$3e^{-4}$	$4e^{-4}$	3.5	0.014	0.004	0.005	99.0	0.045	0.008	0.010	100
MI300												
C_1	0.518	0.033	0.012		0.724	0.032	0.016		0.733	0.029	0.014	
C_2	0.528	0.033	0.012		0.745	0.031	0.017		0.783	0.027	0.015	
Δ	0.010	0.014	0.011	3.0	0.021	0.008	0.009	84.6	0.050	0.011	0.012	100
NRI	0.013	0.023	0.018	1.3	0.076	0.035	0.035	57.0	0.172	0.039	0.039	99.7
IDI	$4e^{-4}$	$4e^{-4}$	$5e^{-4}$	1.8	0.014	0.005	0.006	87.5	0.046	0.010	0.012	100
CC1000												
C_1	0.528	0.032	0.013		0.667	0.033	0.015		0.670	0.031	0.014	
C_2	0.534	0.033	0.015		0.709	0.032	0.022		0.737	0.029	0.014	
Δ	0.006	0.010	0.010	4.7	0.043	0.011	0.017	100	0.067	0.012	0.012	100
NRI	0.017	0.031	0.033	6.7	0.147	0.039	0.043	96.7	0.261	0.041	0.043	100

IDI	0.002	0.001	0.003	15.2	0.058	0.009	0.014	100	0.114	0.011	0.017	100
CC300												
C_1	0.523	0.034	0.013		0.620	0.037	0.016		0.620	0.034	0.015	
C_2	0.529	0.034	0.015		0.647	0.036	0.016		0.668	0.032	0.015	
Δ	0.006	0.010	0.009	3.6	0.027	0.011	0.011	83.3	0.048	0.013	0.013	99.8
NRI	0.019	0.039	0.043	6.2	0.154	0.043	0.050	94.4	0.257	0.046	0.051	99.9
IDI	0.002	0.001	0.003	13.9	0.040	0.008	0.014	99.8	0.078	0.010	0.017	100

C_1, Harrell's C index of the proportional hazard model without the phase-2 variable
C_2, Harrell's C index of the proportional hazard model with the phase-2 variable
Δ, Harrell's predictive value of the phase-2 variable, H_0: $\Delta = 0$
NRI, Net reclassification index by adding the phase-2 variable, H_0: NRI = 0
IDI, Integrated discrimination index by adding the phase-2 variable, H_0: IDI = 0
Cohort, full cohort estimates; MI300, MI1000: multiple imputation estimates with subcohort sizes set, respectively, at 300 and 1,000; CC300, CC1000, case-cohort estimates with subcohort sizes set, respectively, at 300 and 1,000

Table 3 gives the results of the estimated predictive abilities for the correctly specified and the two misspecified normal imputation models. Full cohort analysis and MI provided similar predictive abilities estimates when the imputation model was correctly specified or when the phase-2 variable had no effect on the studied risk. In the scenario $\beta_1 = \beta_2 = \beta_3 = \log(2)$, when Z_2 was uniformly distributed, MI and full cohort analysis still provided similar estimates. However, when Z_2 was log-normally distributed, the MI estimate was slightly smaller than the full cohort estimate (−15%).

Table 3 Predictive ability of the two models and of the phase-2 variable. Results of 1000 simulations

	Full cohort				Multiple imputation			
	Est	$\widehat{SE}$	SE	% H_0 rejected	Est	$\widehat{SE}$	SE	% H_0 rejected
Z_2 normally distributed								
$\beta_1 = \beta_2 = \beta_3 = 0$								
C_1	0.518	0.033	0.012		0.518	0.033	0.012	
C_2	0.524	0.033	0.013		0.526	0.033	0.013	
Δ	0.006	0.010	0.010	3.7	0.008	0.012	0.010	3.4
$\beta_1 = \beta_2 = \beta_3 = \log(2)$								
C_1	0.733	0.029	0.014		0.733	0.029	0.014	
C_2	0.783	0.027	0.013		0.783	0.027	0.014	
Δ	0.049	0.010	0.010	100	0.049	0.010	0.011	100
Z_2 normally distributed								
$\beta_1 = \beta_2 = \beta_3 = 0$								
C_1	0.518	0.033	0.012		0.518	0.033	0.012	
C_2	0.524	0.033	0.013		0.520	0.031	0.016	
Δ	0.006	0.010	0.009	5.5	0.002	0.013	0.012	4.2
$\beta_1 = \beta_2 = \beta_3 = \log(2)$								
C_1	0.784	0.013	0.006		0.784	0.013	0.006	
C_2	0.881	0.011	0.006		0.866	0.011	0.006	
Δ	0.097	0.005	0.005	100	0.082	0.005	0.004	100
Z_2 uniformly distributed								
$\beta_1 = \beta_2 = \beta_3 = 0$								
C_1	0.532	0.055	0.019		0.532	0.055	0.019	
C_2	0.540	0.055	0.019		0.541	0.055	0.020	
Δ	0.008	0.015	0.013	2.2	0.009	0.017	0.013	4.0
$\beta_1 = \beta_2 = \beta_3 = \log(2)$								
C_1	0.733	0.029	0.014		0.733	0.029	0.014	
C_2	0.781	0.027	0.012		0.785	0.027	0.012	
Δ	0.048	0.009	0.009	100	0.052	0.010	0.010	100

C_1, Harrell's C index of the proportional hazard model without the phase-2 variable
C_2, Harrell's C index of the proportional hazard model with the phase-2 variable
Δ, Harrell's predictive value of the phase-2 variable, H_0: $\Delta = 0$

Mean of the predictive ability estimates (Est), mean of the standard error estimates $\hat{SE}$ and standard error of the estimates (SE), with a correctly specified normal imputation model (Z_2 normally distributed), and with two misspecified normal imputation models (Z_2 log-normally and uniformly distributed)

Application to the Three-City study

The mean fraction of missing information about the effect of D-dimer was 4.9 and 3.7 per cent for CHD and VaD risks, respectively. Table 4 gives the estimated hazard ratios associated with D-dimer tertiles. The MI and the weighted approaches yielded similar estimates and precision. The CI of the hazard ratio associated with the linear effect of a one-tertile difference were respectively (0.94–1.38) versus (0.92–1.38) for CHD and (1.13–2.53) versus (1.13–2.67) for VaD. For phase-1 variables, both estimators provided similar results, but MI was always the more precise (data not shown).

Table 4 Estimates of hazard ratios (HR) and 95% confidence interval (CI) associated with D-dimer tertiles

	Multiple imputation estimates	Weighted estimates
	HR (95% CI)	HR (95% CI)
Risk of CHD and D-Dimer [a]		
T1	1.00 (reference)	1.00 (reference)
T2	1.42 (0.99–2.04)	1.40 (0.97–2.04)
T3	1.32 (0.89–1.97)	1.30 (0.84–1.99)
Linear trend	1.14 (0.94–1.38)	1.13 (0.92–1.38)
Risk of VaD and D-Dimer [b]		
T1	1.00 (reference)	1.00 (reference)
T2	1.57 (0.63–3.93)	1.60 (0.63–4.09)
T3	2.77 (1.17–6.57)	2.93 (1.22–7.06)
Linear trend	1.69 (1.13–2.53)	1.74 (1.13–2.67)

CHD, cardiovascular heart disease; T1, tertile 1; T2, tertile 2; T3, tertile 3; VaD, vascular dementia
[a] Adjusted for age, center, sex, body mass index, hypertension, hypercholesterolemia, diabetes, diabetes drugs, tobacco use
[b] Adjusted for age, center, sex, educational level, body mass index, apolipoprotein $\varepsilon 4$

Harrell's C for the models including only phase-1 variables were above 0.69 for CHD risk and above 0.86 for VaD risk (Table 5). Hence, CHD and VaD risks were largely explained by standard risk factors, and the inclusion of plasma D-dimer levels did not significantly improve the predictive ability of the model, despite the fact that elevated D-dimer levels significantly increased the VaD risk. For CHD as for VaD, the index did not significantly differ from 0.

Table 5 Predictive ability and 95% confidence interval (CI) of D-Dimer tertiles on cardiovascular heart disease (CHD) and vascular dementia (VaD) risks

	CHD		VaD	
	Estimate	95% CI	Estimate	95% CI
C_1	0.693	(0.622–0.764)	0.865	(0.787–0.943)
C_2	0.694	(0.621–0.767)	0.874	(0.798–0.950)
Δ	0.002	(−0.004–0.008)	0.009	(−0.011–0.029)
NRI	0.009	(−0.049–0.066)	-	-
IDI	0.001	(−0.001–0.003)	0.0004	(−0.0002–0.0010)

C_1, Harrell's C index of the proportional hazard model without the phase-2 variable
C_2, Harrell's C index of the proportional hazard model with the phase-2 variable
Δ, Harrell's predictive ability of the phase-2 variable
NRI, net reclassification improvement by adding the phase-2 variable
IDI, integrated discrimination index by adding the phase-2 variable

Discussion

Use of a consistent estimator does not guarantee the absence of any bias for finite sample. We only showed that MI analysis of case-cohort data provides unbiased estimates of the log-hazard ratio when the imputation model and the proportional hazard model are correctly specified. The misspecification of the imputation model can originate from an erroneous choice of the distribution, or from wrongly assuming that the estimator of the imputation model is consistent and normal, or from the omission of some important explanatory variable. Imputations carried out using a misspecified distribution in the imputation model can provide biased estimates of hazard ratios, especially, if the specified distribution of the phase-2 variable differs from the true one in terms of symmetry (log-normal versus normal distribution). The negative bias on a log hazard ratio of 0.69 was noticeable but not large when a log-normal variable was imputed according to a normal distribution (−0.09 or −13%), but it is clearly a type of misspecification easily identified with diagnostic tools [20]. One can then transform the incomplete variable in order to obtain a symmetrical distribution, impute transformed values and apply the inverse transformation to the imputed values. Note that although a normal and a uniform distribution are quite different, both are symmetrical and the observed bias was quite smaller (only 5%). In the 3C study of the relationship between VaD and D-dimer, we observed slightly different estimates of the log hazard ratio when comparing the third to the first tertile (2.77 versus 2.93, i.e. a relative difference of 8% between the MI and the weighted estimates). This is probably because of the qualitative imputation of D-dimer, and thus, the use of a multinomial imputation model, which implied estimation of parameters in separate strata defined by D-dimer concentration tertiles, some of which had a small number of events. Due to these small numbers (only 51 VaD in total), asymptotic conditions might not have been fulfilled in at least some strata, and the estimated coefficients of the imputation model could have been biased and notnormally distributed. We give below some recommendations regarding the choice of explanatory variables in the imputation model. Since the potential bias of MI estimates can be detected by comparing them to weighted estimates, we suggest building the proportional hazard model by using only the case-cohort data and weighted estimators. MI can eventually be used to reanalyze the data with the selected model to improve the precision of the results, while verifying that no bias was introduced.

In simulated data, for the phase-1 variables, the precision of MI and full cohort estimates was similar and smaller than with the weighted estimator. For the phase-2 variable, MI estimates were slightly more precise than weighted estimates. Globally, the mean squared errors were smaller with MI than with the weighted estimator, with one exception implying a normal imputation model for a log-normally distributed phase-2 variable, an error which should easily be avoided.

There is no standard method for estimating the predictive ability of a model in the framework of case-cohort surveys. We showed that the naive application of the C index to case-cohort surveys yielded an underestimation of the predictive ability of the model that depended on the subcohort size when the phase-2 variable had an effect on the risk. Similarly, the naive

estimates of the predictive ability of an added phase-2 variable differed notably from the full cohort values when the effect of the phase-2 variable was not null. Harrell's C index could theoretically be estimated with a weighted approach, but this can be computationally difficult because it requires weighting each pair by the pairwise sampling probabilities, i.e., using a square matrix of size $N'(N'-1)$, where N' is the size of the case-cohort sample. Computing the variance of this Horvitz-Thompson estimator requires either weighting each quadruplet by the quadruple-wise sampling probabilities, i.e., working with a matrix of size $N'(N'-1)(N'-2)(N'-3)$, or bootstrapping the case-cohort data. By contrast, MI easily allows estimation of the predictive ability of a model or of an additional phase-2 variable and their variances in the context of case-cohort data, only requiring bootstrapping to estimate the variance of the predictive ability of the phase-2 variable. MI provided estimates of Harrell C, NRI and IDI indexes similar to those obtained with the full cohort analysis. Note, however, that the predictive abilities were always overestimated because the same data were used to estimate the model and its predictive ability.

Analysis of the Three-City case-cohort study was in agreement with our previous work [10]. The weighted and the MI approaches yielded similar estimates of the hazard ratios and MI was slightly more precise, particularly for phase-1 variables. The relative differences between both estimates was always below 2% for the hazard ratios related to CHD and D-dimer, but as early discussed, they could be slightly higher (8%) for a hazard ratio related to VaD and D-dimer. The precision was similar for both analyses.

The imputation model must reflect the association between the incomplete variable, the outcome and the other explanatory variables. Therefore, variables included in the proportional hazard model as well as the stratification variables must be included in the imputation model. If a surrogate of the phase-2 variable is available, it should also be included in the imputation model. On the other hand, multiple imputation approach can provide unbiased and more efficient estimates than weighted analysis even when no strong predictor of the phase-2 variable is available [10]. The inclusion of additional variables other than strongly predictive variables can lead to an increased inter-imputation variance. This prompted the use of different imputation models for D-dimer levels in the CHD and VaD analyses. However, we verified that adding the variables only used in the CHD analysis to the model used for VaD, did not modify the results observed in the former (data not shown).

The number of requested imputations depends on the proportion of missing information which, in case-cohort studies, is considerably smaller than the percentage of incompletely observed subjects. Rubin showed that with as much as 40 per cent information missing, M = 5 imputations provides an asymptotic relative efficiency was 0.97, and, with 50 per cent missing information, M = 10 provides an asymptotic relative efficiency of 0.98. Thus, a small number of imputations, 5–10, should suffice [21]. In our analyses, we used 5 imputations to limit the computer time of the simulations, a reasonable choice since the proportion of missing information was always smaller than 30 per cent. However, a slightly larger number of imputations (e.g. 10) could have been performed on the 3C study data at a reasonable time cost; it would have provided a more precise estimate of the between imputation variance and of the percentage of missing information.

The VaD risk increased with D-dimer tertiles. However, D-dimer inclusion did not significantly improve the predictive ability of the model for VaD risk. Computations of the C and IDI index yielded the same conclusion. To our knowledge, no other results concerning the predictive ability of D-dimer on the risk of VaD have been published to date. The risk of

CHD did not vary with D-dimer, so, not surprisingly, the predictive ability of this variable was negligible, regardless of the index used. Wang et al. [22] and Tzoulaki [23] reported that the use of 10 and 4 biomarkers respectively added only moderately to the overall risk prediction based on conventional cardiovascular risk factors.

Conclusions

MI is a simple alternative approach to weighted analysis for analyzing case-cohort surveys, obtaining correct estimates of the log hazard ratios and their standard errors, improving precision for the phase-1 variable estimates, and providing at least the same precision as weighted estimators for phase-2 variable estimates. It allows an easy evaluation of the predictive ability of the model and, more generally, any tool proposed in the framework of cohort studies can be applied to case-cohort data using MI.

Abbreviations

MI, Multiple imputation; CHD, Coronary heart disease; VaD, Vascular dementia; NRI, Net reclassification index; IDI, Integrated discrimination index.

Competing interests

The authors declare that they have no competing interests

Authors' contributions

HM conducted the literature review, simulations, data analyses and wrote the manuscript. LC conducted the analysis of the relationship between D-dimer levels and CHD and VaD risks and supervised the epidemiological aspects of the application to the Three-City study. MC conducted and supervised the writing of the manuscript. All authors have read the manuscript, are in agreement that the work is ready for submission to the journal, and accept responsibility for the manuscript's contents.

Acknowledgements

This study was supported by a grant from the Région Île-de-France. It used data from the Three-City study which is conducted under an agreement between the Institut National de la Santé et de la Recherche Médicale and the Université Victor Segalen-Bordeaux 2. This manuscript was not prepared in collaboration with the 3C study Steering Committee and does not necessarily reflect its opinions or views.

References

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

2. Chen K, Lo SH: **Case-cohort and case–control analysis with cox's model.** Biometrika 1999, **86**(**4**):755–764.

3. Therneau TM, Li H: **Computing the cox model for case cohort designs.** Lifetime Data Anal 1999, **5**(2):99–112.

4. Borgan O, Langholz B, Samuelsen SO, Goldstein L, Pogoda J: **Exposure stratified case-cohort designs.** Lifetime Data Anal 2000, **6:**39–58.

5. Kulich M, Lin D: **Improving the efficiency of relative-risk estimation in case-Cohort studies.** J Am Stat Assoc 2004, **99:**832–844.

6. Langholz B, Jiao J: **Computational methods for case-cohort studies**. Comput Stat Data Anal 2007, **51**(**8**):3737–3748.

7. Scheike TH, Martinussen T: **Maximum likelihood estimation for Cox's regression model under case-Cohort sampling.** Scand Stat Theory Appl 2004, **31**(2):283–293.

8. Dempster AP, Laird NM, Rubin DB: **Maximum likelihood from incomplete data via the EM algorithm.** J R Stat Soc Series B Stat Methodol 1977, **39:**1–38.

9. Breslow N, Lumley BCCLT, Kulich M: **Using the whole cohort in the analysis of case-cohort data.** Am J Epidemiol 2009, **169**(11):1398–1405, [http://dx.doi.org/10.1093/aje/kwp055].

10. Marti H, Chavance M: **Multiple imputation analysis of case-cohort studies.** Stat Med 2011, **30**(13):1595–1607.

11. Alperovitch A, 3C Study Grp: **Vascular factors and risk of dementia: Design of the three-city study and baseline characteristics of the study population.** Neuroepidemiology 2003, **22**(6):316–325.

12. Little R, Rubin D: Statistical analysis with missing data. New York: Wiley; 1987.

13. Rubin DB, Schenker N: **Multiple imputation in health-care databases: an overview and some applications.** Stat Med 1991, **10**(4):585–598.

14. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA: **Evaluating the yield of medical tests.** J Am Med Assoc 1982, **247**(18):2543–2546.

15. Harrell F, Lee K, Mark D: **Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors.** Stat Med 1996, **15**(4):361–387.

16. Kremers WK: **Concordance for survival time data: fixed and time-dependent covariates and possible ties in predictor and time.** Tech rep Mayo Fundation 2007.

17. Pencina M, D'Agostino R Sr, D'Agostino R Jr, Vasan R: **Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond.** Stat Med 2008, **27**(2):157–172.

18. Carcaillon L, Gaussem P, Ducimetiere P, Giroud M, Ritchie K, Dartigues JF, Scarabin PY: **Elevated plasma fibrin D-dimer as a risk factor for vascular dementia: the three-city cohort study.** J Thromb Haemost 2009, **7**(12):1972–1978.

19. **Executive Summary of The Third Report of The National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, And Treatment of High Blood Cholesterol In Adults (Adult Treatment Panel III).** J Am Med Assoc 2001, **285**(19):2486–2497.

20. D'Agostino RB, Stephens MA: (Eds): Goodness-of-fit techniques. New York: Marcel Dekker; Inc.; 1986.

21. Rubin DB: Multiple imputation for nonresponse in surveys. New York: Wiley; 1987.

22. Wang TJ, Gona P, Larson MG, Tofler GH, Levy D, Newton-Cheh C, Jacques PF, Rifai N, Selhub J, Robins SJ, Benjamin EJ, D'Agostino RB, Vasan RS: **Multiple biomarkers for the prediction of first major cardiovascular events and death.** N Engl J Med 2006, **355**(25):2631–2639.

23. Tzoulaki I, Murray GD, Lee AJ, Rumley A, Lowe GDO, Fowkes FGR: **Relative value of inflammatory, hemostatic, and rheological factors for incident myocardial infarction and stroke - The Edinburgh artery study.** Circulation 2007, **115**(16):2119–2127.