Traitement des données manquantes en épidémiologie : application de l'imputation multiple à des données de surveillance et d'enquêtes

Vanina Héraud Bousquet

UNIVERSITE PARIS XI
FACULTE DE MEDECINE PARIS SUD

Ecole Doctorale de Santé Publique - ED420

Année 2011/2012

N° attribué par la bibliothèque
|__|__|__|__|__|__|__|__|__|__|__|__|

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ PARIS XI

Champ disciplinaire : Epidémiologie et intervention en santé publique

Présentée et soutenue publiquement par

Vanina HÉRAUD-BOUSQUET

Le 06 avril 2012

Traitement des données manquantes en épidémiologie : Application de l'imputation multiple à des données de surveillance et d'enquêtes.

Thèse dirigée par : **Jean-Claude DESENCLOS**

Composition du jury :

Pr Jean-Christophe THALABARD	Président
Dr Rodolphe THIÉBAUT	Rapporteur
Pr Philippe VANHEMS	Rappporteur
Pr Arnaud FONTANET	Examinateur
Dr Jean-Claude DESENCLOS	Directeur de thèse

Lieu de préparation de cette thèse :

Institut de Veille Sanitaire
Département des Maladies Infectieuses
12 rue du Val d'Osne
94415 Saint Maurice

Financement de cette thèse assurée par :

L'Agence Nationale de Recherche sur le Sida
et les hépatites virales
101 rue de Tolbiac
75013 Paris

REMERCIEMENTS

Mes remerciements s'adressent en premier lieu à mon directeur de thèse, Jean-Claude Desenclos, pour m'avoir permis d'effectuer ce travail de recherche à l'Institut de Veille Sanitaire. Merci d'avoir eu l'idée audacieuse de me proposer un tel sujet, tout d'abord en master puis en thèse, d'avoir pris le risque de m'encadrer pour ce projet, et de m'avoir laissée choisir en toute liberté mes sujets de recherche. Ta rigueur scientifique, tes critiques pertinentes et ta présence attentive tout au long de ce parcours ont été essentiels à la réalisation de ce travail. J'espère que ce travail de thèse est digne de ton attente et de ton investissement personnel.

Je remercie aussi tout particulièrement Yann Le Strat pour avoir accepté de co-encadrer cette thèse, en assurant le soutien statistique de cette recherche. Tu as pris le temps d'acquérir une expertise dans le traitement des données manquantes, de m'enseigner avec patience les nécessaires bases statistiques et de développer avec moi des projets de recherche. Ta forte implication dans ce travail de thèse et ton soutien amical m'ont été précieux.

Je remercie très sincèrement les membres de mon jury de thèse :

Jean-Christophe Thalabard me fait l'honneur de présider ce jury de thèse. Sa double compétence en statisques et en épidémiologie, ainsi que l'interêt qu'il a déjà porté à la thématique du traitement des données manquantes, m'apparaissent très intéressantes pour apporter un regard critique à ce travail.

Rodolphe Thiébaut a accepté de rapporter ce travail. Sa connaissance approfondie des techniques statistiques dans le champ des maladies infectieuses constitue un atout indéniable pour juger ce travail. Ses remarques permettront certainement d'alimenter une discussion sur ce travail de thèse.

Philippe Vanhems a accepté de rapporter ce travail. Je connais l'étendue de ses travaux touchant à des sujets très variés en santé publique, et je suis honorée de l'intérêt qu'il a bien voulu porter à cette thèse, avec une approche certainement originale. Je tiendrai compte de ses remarques afin d'améliorer ce travail.

Arnaud Fontanet a accepté d'examiner ce travail. Lors de son enseignement de master, il m'a permis d'acquérir les bases méthodologiques nécessaires à la réalisation de ce travail, et j'apprécie particulièrement qu'il puisse prendre part au débat sur cette thématique qui lui est familière.

Je remercie aussi tout particulièrement :

L'Agence Nationale de Recherche sur le Sida et les hépatites pour avoir assuré le financement de cette thèse, et tout particulièrement Véronique Doré qui m'a guidée dans la procédure d'obtention de ce financement.

Laurence Meyer pour le prêt des données de la cohorte COPANA et l'intérêt qu'elle a porté à ce projet, ainsi que Remonie Seng pour la préparation de ces données.

Dominique Costagliola pour m'avoir fourni les données de la deuxième enquête Odyssée.

James Carpenter pour son précieux soutien méthodologique lors de la mise en application de sa méthode d'analyse de sensibilité.

Jean Bouyer pour son accompagnement efficace et bienveillant lors de la finalisation de ce travail de thèse.

Sans oublier que :

Ce travail n'aurait pû voir le jour sans l'accueil que j'ai reçu au sein du département des maladies infectieuses de l'InVS au cours de ces quatre années. Je remercie donc vivement Christine Saura de m'avoir permis de mener à bien mon travail de recherche en toute sérénité au sein du pôle biostatistique.

Ce travail doit beaucoup à l'implication de l'unité VIC.

Merci tout d'abord à toute l'équipe DO pour sa participation active au projet d'imputation des données du système de surveillance du VIH, et particulièrement à Françoise Cazein, Josiane Pillonel, Stéphane Le Vu et Florence Lot. Je vous dois la chaleureuse ambiance de nos réunions de travail et les idées et critiques pertinentes qui m'ont permis de faire évoluer ce projet au cours du temps. Je remercie plus particulièrement Josiane de m'avoir également associée à son travail sur les donneurs de sang, et Florence pour son active collaboration dans le cadre du projet capture-recapture chez les enfants.

Un grand merci à Caroline Semaille de m'avoir permis d'appliquer cette nouvelle méthode au système de surveillance du VIH, et ce malgré l'impact pour toute l'équipe en termes de manipulation de la nouvelle base de données, et de m'avoir accordé sa confiance pour traiter ces données sensibles. Je ne peux qu'espérer que ce projet se poursuivra.

Je remercie aussi chaleureusement Christine Larsen d'avoir cru dès le départ en mon projet de recherche, et de m'avoir proposé de traiter les données des pôles de référence de l'hépatite C. Merci également à Elisabeth Delarocque-Astagneau pour sa participation à ce travail sur le VHC.

Un merci tout particulier à Anne Gallay pour la qualité de notre entente professionnelle sur les deux projets auxquels elle m'a associée au fil de mon parcours, mais aussi et surtout pour son soutien amical durant ces années.

Je remercie aussi :

Les membres du pôle biostatistique pour la richesse de nos échanges sur nos projets communs, en souhaitant pouvoir les prolonger à l'avenir.

Virginie Bufkens qui, depuis mon arrivée comme stagiaire, s'est chargée avec efficacité de mes multiples déplacements au sein du DMI pendant ces 5 années. Merci aussi pour notre colocation finale !

Christine Aranda pour sa participation active et tout à fait indispensable à la finalisation de ce travail.

Et pour finir :

Je remercie ma famille et mes amis, qui avez su comprendre et accepter cette bien tardive lubie. J'ai un peu tout fait dans le désordre, mais je m'arrêterai là, peut être … Merci d'avoir tenu bon.

A mon petit Octave enfin, qui a grandi en même temps que cette thèse. Tu sais lire maintenant Bibou, et je sais que tu seras content de voir enfin mon 'livre'.

SOMMAIRE

TABLES DES FIGURES

LISTE DES TABLEAUX

GLOSSAIRE

AIC : Critère d'information d'Akaïke (Akaïke Information Criterion).

ANRS : Agence nationale de recherche sur le sida et les hépatites virales.

ARS : Agence régionale de santé.

B : Variance inter-imputation (between-imputation variability).

BIC : Critère d'information bayésien (Bayesian Information Criterion).

CNR : Centre national de référence.

COPANA : Cohorte ANRS CO9 COPANA coordonnée par l'Inserm, démarrée en 2004 et incluant des patients non traités par antirétroviraux pour l'étude du pronostic à court, moyen et long terme des patients infectés par le VIH récemment diagnostiqués.

DOVIH : Déclaration obligatoire de l'infection à VIH.

CSF : Enquête sur le contexte de la sexualité en France, réalisée en 2006 par l'Inserm et l'Ined.

DIC : Critère d'information bayésien adapté par Draper (Draper Information Criterion).

DOVIH : Déclaration obligatoire du VIH.

EPF : Enquête Périnatale Française, cohorte financée par l'ANRS (ANRS-EPF CO1/CO10/CO11).

ECDC : Centre européen pour la prévention et le contrôle des maladies (European Center for Disease Prevention and Control).

EIA-RI : Test d'infection récente développé par le centre national de référence du VIH (Enzyme Immuno Assay for Recent Infection).

FMI : Fraction d'information manquante (Fraction of Missing Information).

HSH : Homme ayant des relations sexuelles avec des hommes.

InVS : Institut de veille sanitaire.

LaboVIH : Enquête évaluant l'activité de dépistage du VIH en France auprès de l'ensemble des laboratoires d'analyses biologiques et médicales.

MAR : Manquant au hasard (Missing At Random).

MCAR : Manquant complètement au hasard (Missing Completely At Random).

MCMC : Algorithme de type Monte Carlo par chaîne de Markov (Markov chain Monte Carlo).

MNAR : Manquant non au hasard (Missing non At Random).

Odyssée : Enquête nationale réalisée en 2006-2007, coordonnée par l'INSERM, avec comme objectif d'estimer la prévalence de la résistance primaire aux antirétroviraux chez des patients chroniquement infectés et naïfs de traitement antirétroviral.

OMS : Organisation mondiale de la Santé, World Health Organisation (WHO).

p : Degré de signification (p-valeur).

PCR : Réaction en chaîne par polymérase (Polymerase Chain Reaction), méthode de biologie moléculaire d'amplification d'ADN in vitro.

RE : Efficacité statistique relative (Relative Efficiency).

TME : Transmission mère enfant.

UDI : Usagers de drogues par voie intraveineuse.

VHB : Virus de l'Hépatite B.

VHC : Virus de l'Hépatite C.

VIH : Virus de l'Immunodéficience Humaine.

W : Variance intra-imputation (within-imputation variability).

RESUME

Le traitement des données manquantes est un sujet en pleine expansion en épidémiologie. La méthode la plus souvent utilisée restreint les analyses aux sujets ayant des données complètes pour les variables d'intérêt, ce qui peut réduire la puissance et la précision et induire des biais dans les estimations.

L'objectif de ce travail a été d'investiguer et d'appliquer une méthode d'imputation multiple à des données transversales d'enquêtes épidémiologiques et de systèmes de surveillance de maladies infectieuses. Nous avons présenté l'application d'une méthode d'imputation multiple à des études de schémas différents : une analyse de risque de transmission du VIH par transfusion, une étude cas-témoins sur les facteurs de risque de l'infection à Campylobacter et une étude capture-recapture estimant le nombre de nouveaux diagnostics VIH chez les enfants. A partir d'une base de données de surveillance de l'hépatite C chronique (VHC), nous avons réalisé une imputation des données manquantes afin d'identifier les facteurs de risque de complications hépatiques graves chez des usagers de drogue. A partir des mêmes données, nous avons proposé des critères d'application d'une analyse de sensibilité aux hypothèses sous-jacentes à l'imputation multiple. Enfin, nous avons décrit l'élaboration d'un processus d'imputation pérenne appliqué aux données du système de surveillance du VIH et son évolution au cours du temps, ainsi que les procédures d'évaluation et de validation.

Les applications pratiques présentées nous ont permis d'élaborer une stratégie de traitement des données manquantes, incluant l'examen approfondi de la base de données incomplète, la construction du modèle d'imputation multiple, ainsi que les étapes de validation des modèles et de vérification des hypothèses.

Mots clés : données manquantes, imputation multiple, analyse de sensibilité, enquêtes, systèmes de surveillance, VIH, hépatite C chronique.

ABSTRACT

Missing data management in epidemiology: Application of multiple imputation to data from surveillance systems and surveys.

The management of missing values is a common and widespread problem in epidemiology. The most common technique used restricts the data analysis to subjects with complete information on variables of interest, which can reduce substantially statistical power and precision and may also result in biased estimates.

This thesis investigates the application of multiple imputation methods to manage missing values in epidemiological studies and surveillance systems for infectious diseases. Study designs to which multiple imputation was applied were diverse: a risk analysis of HIV transmission through blood transfusion, a case-control study on risk factors for Campylobacter infection, and a capture-recapture study to estimate the number of new HIV diagnoses among children. We then performed multiple imputation analysis on data of a surveillance system for chronic hepatitis C (HCV) to assess risk factors of severe liver disease among HCV infected patients who reported drug use. Within this study on HCV, we proposed guidelines to apply a sensitivity analysis in order to test the multiple imputation underlying hypotheses. Finally, we describe how we elaborated and applied an ongoing multiple imputation process of the French national HIV surveillance data base, evaluated and attempted to validate multiple imputation procedures.

Based on these practical applications, we worked out a strategy to handle missing data in surveillance data base, including the thorough examination of the incomplete database, the building of the imputation model, and the procedure to validate imputation models and examine underlying multiple imputation hypotheses.

Keywords : missing data, multiple imputation, sensitivity analysis, surveillance systems, surveys, HIV, chronic hepatitis C.

Lieu de préparation de cette thèse :

Institut de Veille Sanitaire
Département des Maladies Infectieuses
12 rue du Val d'Osne - 94415 Saint Maurice

PRODUCTION SCIENTIFIQUE

PUBLICATIONS DANS DES REVUES INTERNATIONALES AVEC COMITE DE LECTURE

Articles scientifiques issus du travail de thèse

Articles publiés

Pillonel J, Bousquet V, Pelletier B, Semaille C, Velter A, Saura C, Desenclos JC, Danic B. Deferral from donating blood of men who have sex with men: impact on the risk of HIV transmission by transfusion in France. *Vox Sang*. 2012 ;102 :13-21.

Larsen C*, Bousquet V*, Delarocque-Astagneau E*, Pioche C, Roudot-Thoraval F, Desenclos JC. Hepatitis C virus genotype 3 and the risk of severe liver disease in a large population of drug users in France. *J Med Virol.* 2010;82(10):1647-54.

Gallay A, Bousquet V, Siret V, Prouzet-Mauléon V, Valk H, Vaillant V, Simon F, Le Strat Y, Mégraud F, Desenclos JC. Risk factors for acquiring sporadic Campylobacter infection in France: results from a national case-control study. *J Infect Dis.* 2008 ;15;197(10):1477-84.

Article accepté sous réserve de modification mineure

Bousquet V, Lot F, Esvan M, Cazein F, Warszawski J, Laurent C, Bernillon P, Gallay A. A three-source capture-recapture estimate of the number of new HIV diagnoses in children in France during 2003-2006 with multiple imputation of a variable of heterogeneous catchability. *BMC Infect Dis*.

Article en révision

Bousquet V, Larsen C, Carpenter J, Desenclos JC, Le Strat Y. Practical considerations for sensitivity analysis after multiple imputation applied to epidemiological studies with incomplete data. *BMC Med Res Methodol*.

Article en préparation

Bousquet V, Le Strat Y, Cazein F, Le Vu S, Pillonel J, Semaille C, Desenclos J-C. A two-stage multiple imputation process applied to the National HIV surveillance in France.

* Contribution égale des 3 premiers auteurs

Articles scientifiques en lien avec le travail de thèse

Crémieux AC, Wilson d'Almeida K, de Truchis P, Simon F, Le Strat Y, Bousquet V, Semaille C, Le Vu S, Lert F. Undiagnosed HIV prevalence in France based on non-targeted screening in emergency departments. *Soumis.*

Le Strat Y, Pillonel J, Le Vu S, Bousquet V, Semaille C, Lavallée P, Cazein F (2011). Assessing the completeness of reporting of a notifiable disease by a survey-based approach: illustration for the Human Immunodeficiency Virus diagnoses in France. *Soumis.*

Wilson d'Almeida K, Kierzek G, de Truchis P, Le Vu S, Pateron D, Renaud B, Semaille C, Bousquet V, Simon F, Guillemot D, Lert F, Crémieux AC. Widespread routine HIV testing in emergency departments: is it time to shift back to targeted screening ? *Arch Intern Med*. 2012;172:12-20.

Le Vu S, Le Strat Y, Barin F, Pillonel J, Cazein F, Bousquet V, Brunet S, Thierry D, Semaille C, Meyer L, Desenclos JC. Population-based HIV-1 incidence in France, 2003-08: a modelling analysis. *Lancet Infect Dis*. 2010;10(10):682-7.

Semaille C, Cazein F, Lot F, Pillonel J, Le Vu S, Le Strat Y, Bousquet V, Velter A, Barin F. Recently acquired HIV infection in men who have sex with men (MSM) in France, 2003-2008. *Euro Surveill.* 2009;14(48).

PUBLICATIONS DANS DES REVUES NATIONALES AVEC COMITE DE LECTURE

Cazein F, Lot F, Pillonel J, Pinget R, Bousquet V, Le Strat Y, Le Vu S, Leclerc M, Benyelles L, Haguy H, Brunet C, Thierry D, Barin F, Semaille C (2010) Surveillance de l'infection à VIH-sida en France, 2009. *Bulletin Epidémiologique Hebdomadaire*, n° 45-46 : 467-472.

Le Vu S, Le Strat Y, Barin F, Pillonel J, Cazein F, Bousquet V, Brunet C, Thierry D, Semaille C, Meyer L, Desenclos JC (2010) Incidence de l'infection par le VIH en France, 2003–2008. *Bulletin Epidémiologique Hebdomadaire*, n° 45-46 : 473-475.

Cazein F, Pillonel J, Imounga L, Le Strat Y, Bousquet V, Lot F, Leclerc M, Benyelles L, Haguy H, Brunet S, Thierry D, Barin F, Semaille C. (2009) Caractéristiques des personnes diagnostiquées avec une infection à VIH ou un sida, France, 2008. *Bulletin Epidémiologique Hebdomadaire Web.*

Cazein F, Pillonel J, Imounga L, Le Strat Y, Bousquet V, Lot F, Leclerc M, Couturier S, Benyelles L, Haguy H, Semaille C. (2009) Dépistage et diagnostic de l'infection VIH et du sida, France, 2008. *Bulletin Epidémiologique Hebdomadaire Web.*

COMMUNICATIONS ORALES DANS DES CONFERENCES NATIONALES ET INTERNATIONALES

Bousquet V, Le Strat Y, Cazein F, Le Vu S, Pillonel J, Semaille C, Desenclos J-C. A two-stage multiple imputation process applied to the National HIV surveillance in France. 31th Annual Conference of the International Society for Clinical Biostatistics, Montpellier 2010.

Bousquet V, Le Strat Y, Larsen C, Desenclos J-C. Sensitivity Analysis after multiple imputation : application of a weighting approach to epidemiological data. 30th Annual Conference of the International Society for Clinical Biostatistics, Prague 2009.

Bousquet V, Gallay A, Le Strat Y, Desenclos JC. Prise en compte des données manquantes dans une enquête épidémiologique: Intérêt de l'imputation multiple et application à une enquête Cas-Témoins nationale. Communication aux Journées Scientifiques de l'InVS, 29-30 nov 2007.

INTRODUCTION

1. Contexte

Le mode de prise en compte des données manquantes constitue un problème récurrent en épidémiologie, dans le cadre des différents types d'enquêtes, mais aussi, et c'est un sujet plus rarement abordé, des systèmes de surveillance, notamment des maladies infectieuses. Pour ces différents cas de figure, le mécanisme à l'origine des données manquantes peut varier, ce qui conditionne à la fois l'impact et le mode de prise en compte de ces données manquantes ainsi que la nature des biais qui peuvent en découler.

Si, en amont, la maîtrise du processus de collecte des données est essentielle pour limiter au maximum l'impact des données manquantes, des variables collectées sont parfois incomplètes et leur mode de gestion le plus courant consiste à restreindre les analyses aux sujets pour lesquels l'ensemble des variables est entièrement renseigné (analyse dite cas-complet). Cette méthode, généralement appliquée par défaut en épidémiologie, a deux effets majeurs selon le type de données considérées. Lors d'analyses descriptives de données issues d'enquêtes ou de données de surveillance, les résultats obtenus sans tenir compte des données manquantes peuvent ne pas représenter la situation réelle. Mais l'impact des données manquantes est plus marqué lors d'analyses étiologiques ou d'estimations d'indicateurs épidémiologiques (prévalence, incidence). En effet, une analyse étiologique de type cas-complet, ne retenant que les individus pour lesquels les informations sont complètes, induit (i) une perte systématique de puissance statistique, (ii) un risque potentiel de biais des estimations (odds ratios, risques relatifs, rapports de prévalences, etc…), et (iii) un processus de sélection des variables faussé dans une analyse multivariée, au profit des variables les mieux renseignées. De même, l'estimation d'une prévalence ou d'une incidence peut être biaisée si elle est réalisée à partir de variables incomplètes sans prise en compte des données manquantes.

De nombreuses méthodes alternatives de traitement des données manquantes ont été proposées et appliquées en épidémiologie, et les méthodes de maximisation de la vraisemblance et

d'imputation multiple sont les plus performantes. Pour notre travail de recherche, nous avons fait le choix d'appliquer une méthode d'imputation multiple selon des critères théoriques et pratiques. Cette méthode permet d'estimer les valeurs manquantes d'une base de données à partir des valeurs observées, et de redresser les biais potentiels dus aux données manquantes. Issue des travaux de Rubin et basée sur les statistiques bayésiennes, l'imputation multiple consiste à remplacer chaque donnée manquante par un jeu de données estimées à partir des données observées, ce qui permet de prendre en compte l'incertitude associée à chaque étape du processus d'imputation. Chacune des bases complètes ainsi générées fournit alors une estimation du paramètre d'intérêt, puis un estimateur unique est obtenu en calculant la moyenne de ces estimations.

2. Objectifs de la thèse

L'objectif de ce travail est d'investiguer et d'appliquer une méthode d'imputation multiple à des données d'enquêtes et de surveillance dans le domaine des maladies infectieuses.

Cet objectif général se décompose en objectifs plus spécifiques :

- Modéliser les mécanismes de données manquantes des variables d'intérêt dans le cadre d'enquêtes et de systèmes de surveillance. Estimer les données manquantes par une méthode d'imputation multiple avec comme objectif un gain de puissance lors des analyses descriptives et/ou étiologiques et l'obtention d'estimations non-biaisées.

- Aborder à partir d'applications pratiques variées les différentes questions méthodologiques soulevées lors de l'estimation des données manquantes par imputation multiple, selon le type de données traitées et l'objectif de l'étude. Elaborer une procédure standardisée pour la mise en œuvre de cette méthode d'imputation multiple dans le traitement de données de type transversal.

- Proposer des méthodes de validation des résultats obtenus par imputation multiple : élaborer un processus de validation interne (diagnostic interne, validation croisée), appliquer une analyse de sensibilité après imputation, et proposer une validation par des données externes.

3. Démarche méthodologique

L'impact des données manquantes dans le traitement de données d'enquêtes ou de surveillance est souvent ignoré, au mieux discuté. Cependant, l'utilisation d'outils performants permettant de gérer ces données manquantes devient plus fréquente en épidémiologie, avec une évolution nette ces dernières années en termes de publications et de programmes dédiés.

Le contexte méthodologique du traitement des données manquantes est exposé dans le chapitre 1. Nous présentons tout d'abord l'impact des méthodes de traitement standards des bases de données incomplètes sur les estimations en considérant un schéma d'étude de type étiologique. Puis nous décrivons la méthode d'imputation multiple retenue, en abordant ses fondements théoriques puis le processus de mise en application pour des données transversales.

Le chapitre 2 reprend, pour trois études avec des schémas différents, le cheminement méthodologique décrit dans le chapitre 1, tout en soulevant des points spécifiques à chacune des applications de l'imputation multiple. Les illustrations abordent successivement une analyse de risque de transmission du VIH par les dons de sang, une étude cas-témoins exploratoire sur les facteurs de risque de l'infection à Campylobacter, et une étude de type capture-recapture pour l'estimation du nombre de nouveaux diagnostics de l'infection par le VIH chez les enfants, pour laquelle une variable de stratification incomplète a été imputée.

Dans le chapitre 3, nous présentons une étude réalisée à partir d'un système de surveillance de l'hépatite C chronique. Dans un premier temps, une étude étiologique visant à identifier les facteurs de risque de complications hépatiques graves est réalisée pour un sous-échantillon d'usagers de drogue, après imputation des données manquantes. Puis, dans un deuxième temps, nous détaillons la mise en application d'une analyse de sensibilité par pondération à trois facteurs de risque identifiés lors de l'analyse étiologique initiale.

Dans le chapitre 4, nous abordons une application pérenne d'imputation multiple à l'ensemble du système de surveillance du VIH. Nous décrivons l'élaboration d'un modèle d'imputation complexe et son évolution au cours du temps, la mise en place de procédures de validation interne ainsi qu'un essai de validation à partir de deux sources de données externes.

Nous présentons dans le chapitre 5 une synthèse de ce travail et nous proposons des perspectives de recherche.

CHAPITRE 1

TRAITEMENT DES DONNEES MANQUANTES EN EPIDEMIOLOGIE : REVUE DES METHODES

1. Origine des données manquantes

Malgré des collectes de données qui se veulent aussi performantes que possible et pour des raisons souvent hors de tout contrôle, les données manquantes sont fréquemment rencontrées en épidémiologie. C'est le cas pour tous les types d'enquêtes épidémiologiques : enquêtes de cohorte avec des patients perdus de vue à l'origine du phénomène d'attrition, enquêtes cas-témoins et enquêtes transversales avec des problèmes de non-réponse aux questionnaires. Ce phénomène est également prégnant dans le cadre des essais cliniques avec des sorties d'étude ou des problèmes de non-compliance au traitement. On observe le même phénomène dans des systèmes de surveillance lié à des défauts de déclaration ou bien à des déclarations incomplètes. Les données manquantes peuvent découler soit d'une non-réponse réelle, soit d'une réponse inexploitable.

Dans le cas d'une non-réponse réelle, la non-réponse est dite totale lorsque toutes les variables d'intérêt sont manquantes ou lorsque la quantité d'information utilisable est jugée insuffisante. C'est le cas lorsqu'une personne refuse de répondre à une enquête dans sa globalité, ou lorsque la variable d'intérêt (par exemple un test biologique) est manquante alors que les autres variables ont été recueillies. Face à une non-réponse totale, il est important de rechercher si les répondants et les non-répondants diffèrent selon les variables recueillies. On dispose le plus souvent de quelques informations concernant la personne qui ne souhaite pas répondre à l'enquête, soit recueillies lors du contact en même temps que la raison du refus, par exemple à partir de la feuille contact ou de données sociodémographiques récoltées au préalable, soit provenant de bases médico-administratives. Nous n'aborderons pas dans ce travail le traitement de la non-réponse totale, habituellement effectué au moyen d'une méthode d'ajustement des poids de sondage.

Ce travail concerne le traitement de la non-réponse partielle. Les causes de non-réponse partielles sont variées et il est important de pouvoir appréhender le mécanisme sous-jacent à cette non-réponse pour permettre une prise en compte adéquate des données manquantes [1]. La non-réponse peut ainsi être complètement involontaire lorsque la personne enquêtée omet une question ou ne sait pas comment répondre. C'est le cas également lorsque les conditions d'entretien sont instables et que le recueil de données peut être interrompu. Il peut d'agir par exemple d'enquêtes téléphoniques ou bien d'enquêtes réalisées auprès de populations marginalisées telles que les sans domicile fixe, les usagers de drogues ou les populations carcérales. Un autre cas particulier de non-réponse involontaire est celui des prélèvements biologiques. Ainsi, la personne enquêtée peut ne pas être prélevée pour des raisons médicales, ou bien l'analyse biologique peut se révéler impossible car l'échantillon est inutilisable, par exemple si l'échantillon est de volume insuffisant ou de mauvaise qualité.

La non-réponse peut également découler d'une inconsistance des réponses dans un même questionnaire. Il est ainsi courant d'avoir des réponses plus convenues sur des sujets sensibles socialement en début de questionnaire. Il vaut donc mieux prévoir de situer ces questions sensibles dans le questionnaire après une première série de questions standards. Il est également classique d'observer dans un même questionnaire des contradictions évidentes entre deux réponses, comme par exemple entre une vaccination déclarative et le nombre de doses vaccinales indiqué dans le dossier médical. Des inconsistances peuvent également être relevées entre les réponses après recoupement entre deux sources d'information telles que le médecin et son patient, ou encore entre deux sources de fichiers. Un phénomène comparable est dû à des problèmes de lisibilité des questionnaires (écriture illisible, mauvais état des questionnaires), tout particulièrement quand il s'agit d'auto-questionnaires, ainsi qu'à des erreurs de saisie. Les réponses pour lesquelles une inconsistance est relevée doivent faire l'objet d'un recodage en données manquantes.

La non-réponse peut également dépendre d'un mécanisme non-contrôlé lorsqu'elle découle d'une méconnaissance de la réponse. Il s'agit alors d'un manque d'information si la personne enquêtée ne sait pas répondre à certaines questions, ou si le médecin ne peut renseigner une variable lorsque l'information est manquante dans le dossier médical. Une cause fréquente de manque d'information découle de problèmes de mémorisation du fait de la nature de la question ou de l'ancienneté des informations. Il peut s'agir par exemple de questions sur des consommations

alimentaires plus ou moins récentes, sur des expositions professionnelles précises au fil du temps, sur un passé médical ou sur des antécédents médicaux familiaux. Il est alors flagrant que les expositions marquantes vont être mémorisées de façon différentielle, comme par exemple pour une maladie héréditaire grave de type diabète dans un historique médical familial, au détriment de pathologies moins connues ou plus bénignes.

Les mécanismes de non-réponse évoqués ont en commun le fait qu'ils sont dus à des phénomènes involontaires, c'est-à-dire qu'il n'existe pas de lien entre la non-réponse et le mécanisme à l'origine de cette non-réponse. Par contre, cela n'est pas le cas lorsque la non-réponse dépend de la nature des questions abordées qui peuvent être considérées par le répondant comme trop sensibles comme pour certaines consommations telles que la consommation d'alcool ou de tabac, certains comportements liés à l'hygiène ou à la sexualité, ainsi que certains sujets d'ordre médical ou intime. La non-réponse sera alors liée à un phénomène volontaire dû au répondant et répercutée comme un refus direct ou comme une réponse de type "ne sait pas".

Dans l'idéal, il faudrait tenir compte du risque de non-réponse lors du recueil de données en construisant des questionnaires adaptés, c'est-à-dire les plus courts possible, avec une durée de passation suffisante avant d'aborder les questions sensibles, ou encore de fiches de déclaration obligatoire synthétiques et claires. Une modalité "ne sait pas" devrait être prévue pour chaque question afin de discriminer les vraies non-réponses. Enfin, lorsque le recueil de données est terminé et qu'aucun retour à l'enquêté n'est possible, il est important de pouvoir identifier le mécanisme à l'origine de la non-réponse, d'un point de vue épidémiologique mais aussi statistique, car le traitement des données manquantes reposera en partie sur la connaissance a priori de ce mécanisme.

2. Typologie des données manquantes

2.1. Notations

Pour chaque individu $i = 1, \cdots, n$ de l'enquête, on note Y_i^{obs} l'ensemble des valeurs renseignées et Y_i^{miss} l'ensemble des valeurs manquantes. Pour chaque individu *i*, on peut donc constituer l'ensemble $Y_i = \left(Y_i^{obs}, Y_i^{miss}\right)$ auquel on associe un vecteur, noté $R_i = (R_{ij}\,; j = 1, \cdots, k)$, composé de variables indicatrices égales à 1 si la variable n'est pas renseignée pour l'individu *i* et 0 sinon. Soit $X = \left(X_1, ..., X_l\right)$ un vecteur de *l* covariables complètes pour les mêmes individus.

2.2. Structure des données manquantes

La structure des données manquantes doit être explorée avant le traitement d'une base de données incomplète, car le choix de la méthode d'estimation des données manquantes en dépend [2].

Structure univariée

La structure des données manquantes est dite univariée lorsqu'une seule variable contient des données manquantes comme illustré en Figure 1.1. C'est le cas le plus simple et il est rarement observé en pratique. Nous y ferons référence dans une illustration simple des mécanismes de données manquantes (paragraphe 2.3.3).

Structure monotone

La structure des données manquantes est dite monotone lorsque les variables incomplètes peuvent être ordonnées en fonction de la proportion de données manquantes qu'elles contiennent. Ainsi, on peut dire que les variables $Y_1, ..., Y_k$ sont ordonnées selon une structure monotone si, pour $j = 1, ..., k-1$, tous les cas contenant des données manquantes pour Y_j présentent également des données manquantes pour $Y_{>j}$ (Figure 1.1).

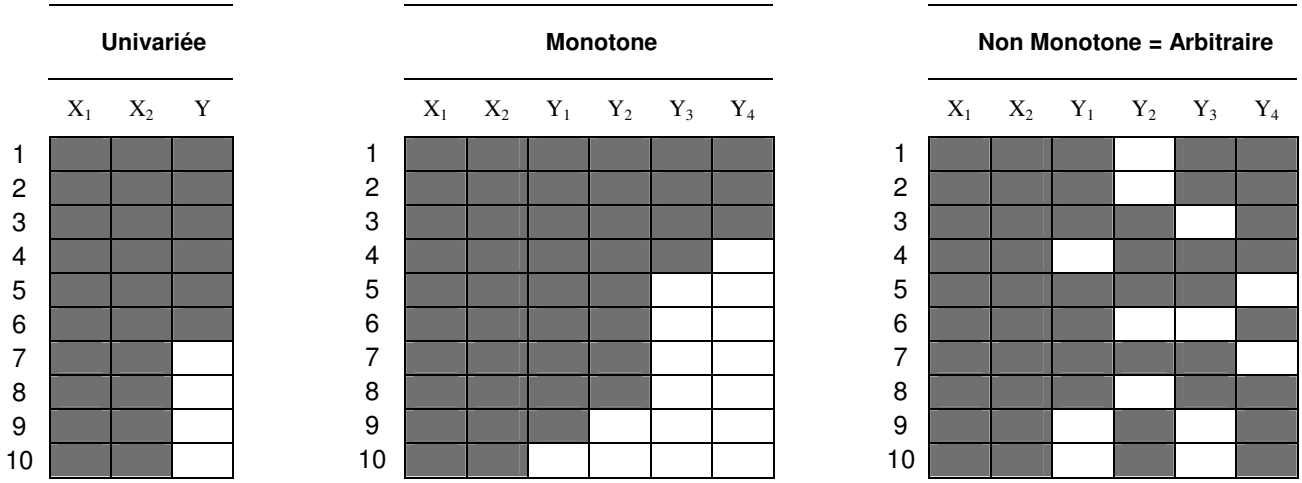
Une structure de type monotone est observée par exemple sur des données d'étude longitudinale lorsqu'un évènement cause la sortie d'étude d'un sujet. On parle alors de phénomène d'attrition.

Lorsqu'un sujet ne se présente pas à une visite ou qu'un examen médical ne peut être effectué, on parle alors de structure monotone intermittente.

Structure arbitraire

La structure des données manquantes est dite arbitraire lorsque les variables incomplètes ne peuvent pas être ordonnées selon leur proportion de données manquantes. Les données manquantes suivent alors une structure non-monotone puisqu'elles sont réparties uniformément dans l'ensemble de la base de données (Figure 1.1). Une structure de type arbitraire est souvent observée en épidémiologie pour les données de type transversal ainsi que pour les données produites par les systèmes de surveillance.

Figure 1.1 – Représentation des structures de données manquantes (d'après [2])



Les zones blanches représentent les données manquantes.

2.3. Mécanismes de données manquantes

2.3.1. Terminologie

Lorsque l'on souhaite analyser un jeu de données présentant des données manquantes, il est nécessaire d'appréhender la loi de probabilité à l'origine de ces données manquantes. De manière formelle, cette loi est souvent nommée mécanisme de données manquantes. Ce mécanisme est défini indépendamment de la structure de données manquantes précédemment décrite.

En reprenant les notations précédentes, un mécanisme peut être formulé comme la distribution conditionnelle de *R* (vecteur d'indicatrices de données manquantes) sachant *Y* (vecteur de covariables incomplètes) et *X* (vecteur de covariables complètes), soit $P(R=1)=f(R|Y,X,\phi)$ où ϕ désigne les paramètres du modèle.

Une terminologie proposée par Little & Rubin [3] permet de distinguer trois catégories de données manquantes en fonction du type de relation statistique entre les données et le mécanisme de données manquantes. Nous présentons cette typologie dans le cas général, puis nous la développons à partir d'un exemple simple permettant de préciser des sous-catégories de ces mécanismes, correspondant à des situations concrètes en épidémiologie.

2.3.2. Typologie générale

Les données sont dites manquantes complètement au hasard, ou MCAR pour "Missing Completely At Random", si la probabilité qu'une donnée soit manquante est une constante : $P(R=1)=p$. Dans ce cas, cette probabilité ne dépend pas des valeurs observées ou manquantes des autres variables. Les données manquantes représentent alors un sous-échantillon aléatoire de l'ensemble des données.

Les données sont dites manquantes au hasard, ou MAR pour "Missing At Random", si la probabilité qu'une donnée soit manquante ne dépend que des valeurs observées des variables (Y^{obs}) et des valeurs des variables complètes (X): $P(R=1)=f(R|Y^{obs},X,\phi)$. Cette probabilité ne dépend donc pas des valeurs non-observées des variables, conditionnellement aux valeurs observées.

Les données sont dites manquantes non au hasard, ou MNAR pour "Missing Non At Random", si la probabilité qu'une donnée soit manquante dépend de valeurs non-observées des variables, ou bien de variables qui n'auraient pas été recueillies : $P(R=1)=f\left(R\middle|Y^{miss},Y^{obs},X,\phi\right)$. Cette probabilité peut donc dépendre des valeurs observées des variables $\left(Y^{obs}\right)$, des valeurs des variables complètes $\left(X\right)$, mais aussi des valeurs non-observées des variables $\left(Y^{mis}\right)$.

Cette terminologie déjà ancienne est conservée telle quelle dans la littérature mais il faut noter qu'elle est ambigüe. En effet, dans le cas de données manquantes au hasard (MAR), le mécanisme est dit aléatoire alors qu'il dépend de certaines valeurs observées. Le terme aléatoire signifie alors qu'une analyse correcte peut être réalisée sans qu'une modélisation préalable du mécanisme de données manquantes soit nécessaire. Nous verrons que, pour les analyses standards, cela n'est vrai que sous certaines conditions.

2.3.3. Illustration en épidémiologie

On propose de s'intéresser à une situation fictive où, pour chaque individu, on dispose de trois variables binaires [1;4]: une variable d'intérêt qui serait le statut de la maladie (M), une variable d'exposition (E) et une variable de confusion binaire incomplète (C). On construit une variable indicatrice de données manquantes (R) qui vaut 1 si la variable C n'est pas renseignée et 0 sinon.

- ***Mécanisme MCAR***

La probabilité de données manquantes peut s'écrire : $P(R=1|M=m,E=e,C=c)=p$. C'est l'hypothèse la plus stricte concernant les mécanismes de données manquantes et elle est rarement rencontrée en pratique. Outre les cas de figures où des données sont manquantes de façon non-planifiée, comme par exemple si un échantillon biologique est perdu ou inutilisable, un mécanisme MCAR est plausible lorsque les données manquantes sont planifiées. Ainsi, un examen complémentaire coûteux peut être effectué seulement pour un sous-échantillon aléatoire de patients, comme par exemple le génotypage du VIH dans un système de surveillance du VIH, ou certains dosages biologiques tels que le dosage de plombémie ou des dioxines, dans une enquête de terrain en santé environnementale. Ces données manquantes sont alors planifiées dès l'élaboration du protocole et l'analyse se fait en deux temps en tenant compte de ce mode de recueil différentiel [1;5].

- ***Mécanisme MAR***

Trois scénarios sont possibles :

(1) La probabilité de données manquantes du facteur de confusion C ne dépend que du statut de la maladie M : $P(R=1|M=m,E=e,C=c)=p(m)$ et nous noterons alors le mécanisme de données manquantes MAR(M).

Illustration pour des données d'enquêtes cas-témoins ou transversales.

Dans les études cas-témoins, les données manquantes dépendent souvent du statut de la maladie car les cas et les témoins différent dans leur comportement ainsi que dans leur volonté de participer à l'investigation et de répondre à des questions spécifiques. On s'attend ainsi à ce qu'un cas se souvienne mieux de certaines expositions parce qu'il a réfléchi aux causes possibles de sa maladie. Un effet inverse est attendu pour des maladies stigmatisantes comme le VIH où le cas peut ne pas souhaiter répondre à un questionnaire complémentaire.

Par exemple, une enquête cas-témoins est réalisée pour identifier les facteurs de risque de survenue d'une tumeur cérébrale [1]. Parmi d'autres facteurs de risques, le groupe sanguin des sujets est recueilli. Pour les témoins, seules les données issues du questionnaire sont disponibles alors que, pour les cas, des informations provenant des dossiers hospitaliers peuvent également être utilisées. Pour la variable groupe sanguin, la proportion de données manquantes est de 9% pour les cas et de 46% pour les témoins.

A l'inverse, si l'on considère une enquête transversale visant à construire un modèle prédictif d'une pathologie, par exemple un score basé sur plusieurs variables, le recueil de données est forcément indépendant de la pathologie c'est-à-dire de la variable à expliquer puisque celle-ci est définie après le recueil de données [6].

Illustrations pour des données de cohortes

Il faut noter que, pour les données de cohortes, les données manquantes découlent de mécanismes différents selon que les données sont collectées à l'inclusion ou au cours du suivi. En effet, dans ce type d'étude, la problématique des données manquantes se pose plus particulièrement au cours du suivi des patients inclus, souvent effectué au moyen d'autoquestionnaires. Par ailleurs, le lien entre la probabilité de données manquantes pour certaines variables et le statut de la maladie est lié au mode de recueil des données de la cohorte. En effet, dans les études de cohortes prospectives, on peut a priori exclure une

dépendance entre la probabilité de données manquantes et le statut de la maladie puisque toutes les données sont collectées à l'inclusion. A l'inverse, le recueil de données est souvent différentiel selon le statut de la maladie pour les cohortes rétrospectives.

(2) La probabilité de données manquantes du facteur de confusion C dépend de l'exposition E : $P(R=1|M=m,E=e,C=c)=p(e)$ et le mécanisme de données manquantes est noté MAR(E).

Considérons par exemple une enquête auprès de sujets exposés à des radiations dans le cadre de leur travail, et pour lesquels cette exposition est suivie grâce à des dosimètres individuels. Les sujets pour lesquels une exposition aux radiations est relevée ont des données plus complètes pour d'autres facteurs de risques tels que la consommation de tabac, car ils sont mieux suivis médicalement et donc davantage questionnés.

(3) La probabilité de données manquantes du facteur de confusion C dépend de la maladie et de l'exposition : $P(R=1|M=m,E=e,C=c)=p(e,m)$ et le mécanisme de données manquantes est noté MAR(ME).

Prenons l'exemple d'une enquête cas-témoins dont l'objectif est d'estimer l'excès de risque de leucémie secondaire lié à une chimiothérapie chez des femmes traitées pour un cancer du sein. La variable thérapie hormonale de substitution sera mieux renseignée chez les malades mais également chez les exposées, c'est-à-dire chez les femmes traitées pour un cancer du sein. En effet, l'historique de l'exposition à différents traitements hormonaux sera plus particulièrement exploré chez ces femmes atteintes d'un cancer du sein.

Ces trois mécanismes de type MAR ne sont pas équivalents lors de l'analyse d'une base de données incomplète. Nous verrons que le cas MAR dépendant conjointement de la maladie et de l'exposition, dénoté MAR(ME), est fréquemment rencontré dans les enquêtes épidémiologiques. Il peut induire des biais dans les estimations lors d'analyses par les méthodes classiques.

- ***Mécanisme MNAR***

Un mécanisme MNAR signifie que la probabilité de données manquantes pour le facteur de confusion C dépend des valeurs non-observées de C (et éventuellement d'autres variables non-recueillies). Le mécanisme de données manquantes est noté selon les cas de figure MNAR(C), MNAR(EC), MNAR(MC) ou MNAR(MEC).

Par exemple, dans une base de données de surveillance du VIH, le mode de contamination supposé peut être moins bien renseigné si le sujet est homosexuel ou bisexuel car il s'agit d'une question sensible. Le mécanisme est alors MNAR(C).

De même, dans une enquête transversale sur les risques de complication d'une hépatite C, les personnes ayant une consommation d'alcool importante peuvent donner moins fréquemment le niveau de celle-ci. Les données manquantes dépendent alors de la vraie valeur, non renseignée pour cette variable, et suivent un mécanisme de type MNAR(C). Ce mécanisme peut être amplifié pour les patients souffrant de complications hépatiques graves. Il dépend alors également de la variable d'intérêt et il est donc de type MNAR(MC). Il est par ailleurs possible que des patients, soumis à une autre exposition telle que la co-infection par le VIH, aient tendance à moins bien renseigner la variable consommation d'alcool. On est alors dans le cas d'un mécanisme de type MAR(EC), voire MAR(MEC) si une dépendance additionnelle avec la présence de complications hépatiques graves est observée.

Ces mécanismes de données manquantes, définis à partir de trois variables seulement, permettent d'explorer les principaux cas de figures que l'on peut rencontrer en pratique.

3. Méthodes simples de traitement des données manquantes

3.1. Analyse cas-complet

La méthode d'analyse de bases de données incomplètes la plus répandue consiste à restreindre l'analyse aux individus pour lesquels l'ensemble des variables est entièrement renseigné. Cette méthode, dite analyse cas-complet et appliquée par défaut par la plupart des logiciels d'analyse statistique, a longtemps été considérée comme la manière la plus "propre" de gérer les données manquantes. Mittienen propose ainsi de supprimer de l'analyse tous les sujets pour lesquels des informations sont manquantes, ajoutant que cette approche est la seule qui garantisse qu'aucun biais n'est introduit, et ce quels que soient les mécanismes induisant les données manquantes [7].

Ce sujet a été largement repris dans la littérature et il est à présent établi que l'analyse cas-complet, puisqu'elle n'utilise pas toutes les informations disponibles dans la base de données, induit une perte de puissance et donc de précision. Dans le cas d'une analyse multivariée, ce type d'analyse peut également fausser le processus de sélection des variables puisque celui-ci se fera au profit des variables les mieux renseignées. Enfin, puisque l'analyse cas-complet sélectionne un sous-échantillon de la base de données initiale qui n'est généralement pas aléatoire, elle peut induire des biais dans les estimations en fonction du mécanisme de données manquantes en cause [3].

Selon la proportion et la répartition des données manquantes dans la base de données, il est aisé de prévoir l'étendue de la perte de puissance attendue pour une analyse donnée. Si la répartition des données manquantes est de type aléatoire, et que les variables incluses dans une analyse multivariée sont majoritairement incomplètes avec une proportion d'environ 5 à 10% chacune, la perte d'effectifs dépasse couramment 50%, et peut même empêcher la convergence statistique du modèle souhaité [8;9].

Pour ce qui est des biais attendus pour les estimations en analyse multivariée, la littérature n'est pas toujours claire. Alors qu'il est bien établi qu'une analyse cas complet ne sera pas biaisée si le mécanisme de données manquantes est de type MCAR, il est parfois stipulé que, si les données manquantes sont dues à un mécanisme de type MAR, une analyse cas complet

sera systématiquement biaisée [10]. Même si dans un contexte pratique cette assertion sera souvent vérifiée, il est important de la modérer en spécifiant qu'une analyse cas-complet sera non-biaisée pour un mécanisme MCAR, ou MAR ne dépendant pas de la variable à expliquer [11]. Il faut aussi noter que, comme le souligne Allison [12], une analyse cas-complet peut être valide lorsque le mécanisme de données manquantes est de type MNAR, alors que des analyses plus élaborées telles que l'imputation multiple seront biaisées.

En pratique, une analyse cas-complet est justifiée si la proportion de cas incomplets est faible, induisant ainsi une perte de puissance et de précision limitée. En pratique, un seuil de 5% de données manquantes est souvent cité dans la littérature. Il est cependant difficile de formuler des recommandations puisqu'il faut également tenir compte du mécanisme à l'origine des données manquantes et du nombre d'individus dans la base de données.

3.2. Analyse de tous les cas disponibles

3.2.1. Création d'une catégorie additionnelle

Une approche alternative à l'analyse cas complet, souvent utilisée en épidémiologie lorsque les variables incluses dans l'analyse sont binaires ou catégorielles, consiste à créer une catégorie additionnelle en remplaçant toutes les données manquantes par une valeur fixe. Cette approche présente l'avantage de conserver l'intégralité des effectifs. Elle est peu discutée dans la littérature, mais Vach et Blettner [1;4] ont démontré à partir d'un exemple simple qu'elle produit toujours des estimations biaisées, et ce quelle que soit la typologie des données manquantes, c'est-à-dire même lorsque les données sont MCAR.

Par ailleurs, d'un point de vue épidémiologique, les estimations produites sont difficilement interprétables puisque la catégorie additionnelle peut regrouper des modalités très différentes de la variable ainsi recodée. Cette méthode ne peut donc être considérée comme valide que dans des cas de figures très particuliers tels que l'utilisation des scores de démence en psychiatrie: les résultats d'un test sont manquants car les patients ne le comprennent pas et la catégorie données manquantes est alors très prédictive du diagnostic [13]. Cette approche reste donc à proscrire en dehors de cas très particuliers.

3.2.2. Indicateur de donnée manquante

Une autre méthode consiste à remplacer dans le modèle d'analyse chaque variable incomplète X_i par une paire de variables. Il s'agit d'associer une variable indicatrice de réponse R_i, codée 1 si la valeur est manquante et 0 sinon, à une variable X_i^* égale à X_i si celle-ci est connue, et égale à 0 sinon. Pour une variable continue, X_i peut prendre pour valeur la moyenne des valeurs observées. L'analyse peut alors porter sur l'intégralité de la base de données en remplaçant X_i par le couple $X_i^* - R_i$. Notons que, dans le cas d'une variable catégorielle, la méthode est identique à l'ajout d'une catégorie supplémentaire. Au final, même si cette méthode présente l'avantage de conserver l'intégralité de la base de données, elle peut induire des biais pour les estimations quel que soit le mécanisme de données manquantes [11].

3.3. Imputation simple

L'imputation simple consiste à remplacer chaque donnée manquante par une estimation (et une seule) de sa valeur et à analyser la base de données ainsi complétée. D'un point de vue statistique, cette procédure de remplacement peut être stochastique ou déterministe, selon qu'elle implique ou non le tirage d'un nombre aléatoire.

Pour les méthodes déterministes, sous l'hypothèse MCAR, les valeurs manquantes peuvent être remplacées par la valeur moyenne des valeurs observées auprès des sujets ayant des données complètes. Sous l'hypothèse MAR, il est fréquent de remplacer les données manquantes par la moyenne des valeurs observées sur les sujets ayant les mêmes caractéristiques ou par la valeur prédite en fonction des covariables à l'aide d'un modèle de régression estimé sur l'échantillon complet [12].

Pour la méthode stochastique la plus simple, la valeur de remplacement est issue d'un tirage aléatoire à partir des réponses complètes. Le choix de cette valeur peut également être effectué par tirage aléatoire parmi les sujets ayant la même probabilité de non-réponse pour la variable à estimer [14;15]. Les modalités de tirage aléatoire peuvent être plus ou moins complexes. Ainsi, des échantillonnages de type bootstrap ou jacknife sont choisis pour des méthodes d'estimation conditionnelle élaborées [16].

Les méthodes déterministes produisent des estimateurs biaisés si les données ne sont pas MCAR. Si le modèle d'imputation est correct et que les données sont MAR, les paramètres estimés à partir des données complétées sont non biaisés. Cependant, ces méthodes d'imputation simple induisent systématiquement une sous estimation de la variance car l'incertitude liée à la présence de valeurs estimées n'est pas prise en compte par les logiciels standards [17].

4. Méthode par maximisation de la vraisemblance

La méthode par maximisation de la vraisemblance est une méthode reconnue de traitement des données manquantes, considérée par certains auteurs comme la méthode de référence dans ce domaine [10;17]. Elle permet de déterminer les valeurs des paramètres qui donnent la densité de probabilité la plus élevée possible en fonction des données observées, et ce pour une famille de modèles paramétriques.

La fonction de vraisemblance et sa maximisation

Reprenons les notations présentées dans le paragraphe 2.1.

Notons $f(Y_i|\theta)$ la distribution de Y_i. L'objectif est d'estimer sans biais le vecteur des paramètres θ, à partir des seules données observées. En notant ψ le vecteur des paramètres décrivant le processus d'observation, la vraisemblance des données observées Y^{obs} et R s'écrit [10;12;14] :

$$L(\theta,\psi|Y^{obs},R)=\prod_{i=1}^{n} f(Y_i^{obs},R_i|\theta,\psi)=\prod_{i=1}^{n}\int f(Y_i^{obs},Y_i^{mis},R_i|\theta,\psi)\,dY_i^{mis}$$

$$=\prod_{i=1}^{n}\int f(Y_i^{obs},Y_i^{mis}|\theta)\times f\left(R_i|Y_i^{obs},Y_i^{mis},\psi\right)dY_i^{mis}$$

Lorsque le mécanisme de données manquantes est MAR, le second terme dans l'intégrale $f\left(R_i|Y_{i,obs},Y_{i,mis},\psi\right)$ devient $f\left(R_i|Y_{i,obs},\psi\right)$ et peut donc sortir de l'intégrale.

La vraisemblance s'écrit alors :

$$L(\theta,\psi|Y^{obs},R)=\prod_{i=1}^{n} f\left(R_i|Y_i^{obs},\psi\right)\int f(Y_i^{obs},Y_i^{mis}|\theta)\,dY_i^{mis}$$

$$=\prod_{i=1}^{n} f\left(R_i|Y_i^{obs},\psi\right)\times f(Y_i^{obs}|\theta)$$

En supposant que θ et ψ sont distincts, seules les informations complètes contribuent à la vraisemblance de θ puisque la contribution d'une observation incomplète s'obtient en faisant la somme sur toutes les modalités possibles de Y^{mis}.

La maximisation de ce type de fonction de vraisemblance peut s'avérer délicate et des méthodes numériques adaptées sont souvent nécessaires (par exemple l'algorithme Estimation Maximisation).

Cette méthode fait l'hypothèse que les données sont MAR, et le calcul de la variance tient compte de la présence de valeurs estimées dans la base de données. Des procédures sont implémentées dans certains logiciels (SAS, S-PLUS, SPSS) mais restent complexes à utiliser [12;14].

5. Méthodes par imputation multiple

Les méthodes d'estimations simples présentées précédemment se révèlent peu efficaces, et peuvent produire des estimations biaisées. L'imputation multiple, par rapport à l'imputation simple, consiste à remplacer chaque valeur manquante par plusieurs valeurs estimées. Ces estimations sont basées sur un modèle qui permet d'utiliser toute l'information disponible dans la base de données, et de préserver ainsi les relations entre les variables. De plus, le fait d'estimer plusieurs valeurs pour chaque donnée manquante permet de prendre en compte la variabilité autour de chaque donnée imputée, et d'obtenir une variance correcte pour les estimations. L'imputation multiple est basée sur l'hypothèse que les données sont MAR, c'est-à-dire que le mécanisme de données manquantes ne dépend pas de données non-observées des variables.

La méthode de maximisation de la vraisemblance est souvent comparée à l'imputation multiple puisqu'elle présente des propriétés proches. Ainsi, les estimations obtenues par la méthode de maximisation de la vraisemblance sont en théorie plus précises que pour l'imputation multiple, qui introduit une composante aléatoire due à la simulation. Cependant, ce bruit résiduel en imputation multiple tend à être négligeable si le nombre de bases imputées est adapté à la proportion de données manquantes. L'imputation multiple est ainsi considérée comme une bonne approximation de la méthode de maximisation de la vraisemblance [10].

Par ailleurs, la méthode de maximisation de la vraisemblance repose comme l'imputation multiple sur l'hypothèse que les données sont MAR. Si le mécanisme de données manquantes est de type MNAR, la maximisation de la vraisemblance présente les mêmes limites que l'imputation multiple, c'est-à-dire qu'elle peut induire des biais dans les estimations. Cependant, puisqu'un même modèle doit être spécifié pour les phases d'estimation et d'analyse, il est plus difficile d'inclure des variables supplémentaires dans le modèle afin de mieux capturer le mécanisme de non réponse et de rendre l'hypothèse MAR plus plausible. Enfin, cette méthode est d'un abord plus complexe et ses propriétés sont plus particulièrement intéressantes dans le traitement de données longitudinales ou à structure hiérarchique (modèles multi-niveaux)[12;14;17].

Du fait des propriétés statistiques de l'imputation multiple, nous avons fait le choix de l'appliquer dans notre travail pour traiter des bases de données incomplètes.

5.1. Historique

La méthode d'imputation multiple a été originellement proposée en 1978 par Rubin dans le domaine des sciences sociales [3]. Par la suite, la mise en application de la méthode a été développée par Rubin, dans le contexte de bases de données importantes issues d'enquêtes complexes et destinées à être exploitées par de nombreux utilisateurs et pour des analyses variées [18]. L'imputation multiple est alors appliquée à l'ensemble de la base de données incomplète, sans tenir compte des analyses ultérieures. A l'époque, les développements de l'imputation multiple sont encore obscurs et son utilisation est limitée à des experts [19]. Les équipes assurant les étapes d'imputation multiple et d'analyse de ces bases de données sont donc des entités distinctes. L'objectif est alors que les équipes d'utilisateurs puissent effectuer toutes les analyses prévues en appliquant les commandes usuelles.

Au cours du temps, la mise en œuvre de l'imputation multiple est devenue plus accessible du fait de l'apparition d'ordinateurs performants et de programmes dédiés [20]. Le champ d'application de l'imputation s'est étendu dans le domaine de la santé publique, avec des applications à d'importantes bases de santé publique [21-23], mais aussi à des enquêtes observationnelles [24], et de façon précoce et élaborée aux essais cliniques [25]. Cependant son utilisation dans le domaine de l'épidémiologie reste encore globalement limitée [26], même si une littérature très récente détaille les aspects théoriques ainsi que la mise en application de la méthode [27;28], ce qui préfigure une utilisation plus fréquente.

L'approche de l'imputation multiple a évolué au cours du temps vers un mode d'imputation dit "in house", puisque c'est souvent la même équipe qui exécute les processus d'imputation et d'analyse des données. L'imputation est réalisée conditionnellement aux analyses ultérieures, et les résultats des analyses peuvent être pris en compte pour une amélioration éventuelle du processus d'imputation. C'est dans cette optique que nous avons choisi d'appliquer la méthode d'imputation multiple à des données de surveillance et d'enquêtes, et nous verrons que, dans certains cas, il est impératif de tenir compte des analyses programmées par les utilisateurs.

Dans la littérature, les livres de Little et Rubin [3] et de Schafer [29] sont les ouvrages de référence, complétés par le livre de Molenberghs [8] dans le domaine clinique. Par ailleurs, des articles synthétisent l'essentiel de la théorie [10;17;19;30] alors que d'autres articles privilégient une approche plus pratique [31-33]. Des articles présentent également des applications de l'imputation multiple à des données d'enquêtes ou de surveillance [21;22;34-37], ou dans le cadre de la construction de modèles prédictifs [6;9]. Les diverses implémentations dédiées à l'imputation multiple sont détaillées par Horton et al. [38;39] et des informations complètes et accessibles sont disponibles sur des sites internet (www.missingdata.org.uk de Carpenter et Kenward ; www.multiple-imputation.com de Van Buuren).

5.2. Bases théoriques

Nous détaillons dans cette partie les fondements statistiques des étapes d'imputation et d'analyse, sachant que la mise en pratique du processus complet d'imputation multiple sera présentée dans le paragraphe 5.3.

5.2.1. Principe

Le principe général de l'imputation multiple consiste à remplacer chaque valeur manquante par un ensemble de M valeurs plausibles, de façon à prendre en compte l'incertitude liée au processus d'estimation des valeurs manquantes [3].

L'imputation multiple se décompose en trois phases (Figure 1.2) :

- ***Phase d'imputation***

Les données manquantes sont estimées M fois à partir d'un modèle spécifique pour obtenir M bases de données complètes et potentiellement différentes.

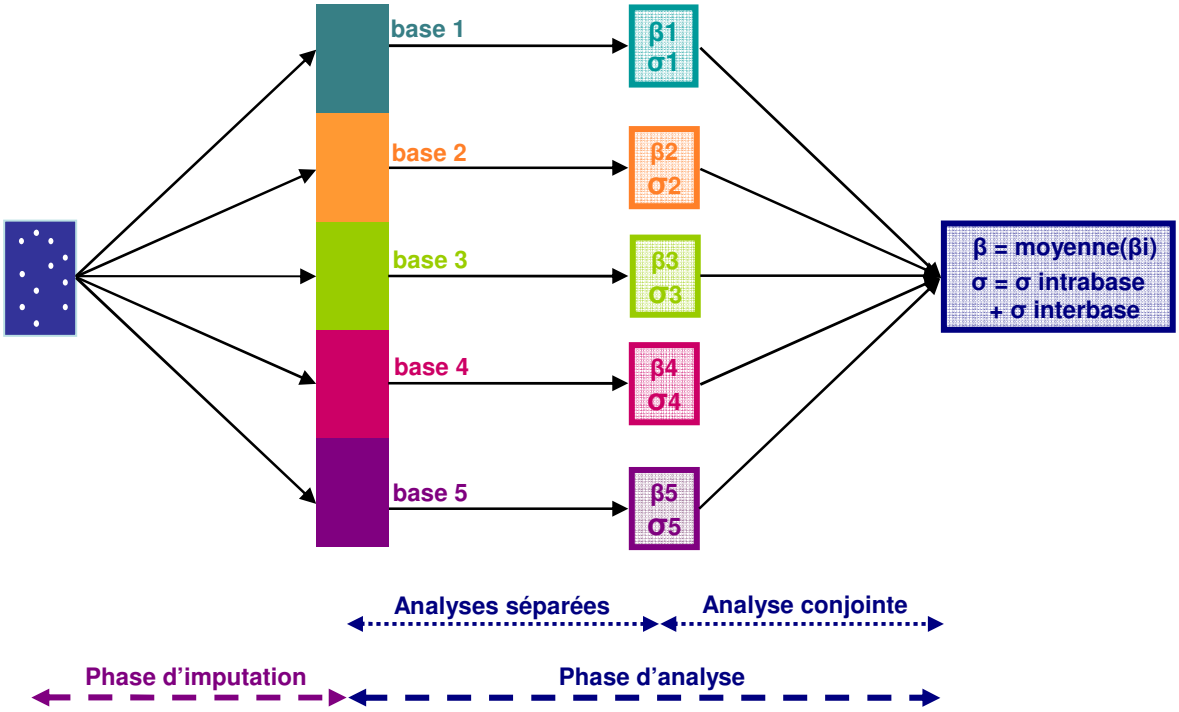
- ***Phase d'analyse séparée***

L'analyse retenue est réalisée séparément sur chacune des $m = 1,...,M$ bases de données imputées pour obtenir *M* estimations (valeur centrale et variance).

- ***Phase d'analyse combinée***

Les résultats obtenus à partir des *M* analyses sont combinés selon des règles établies par Rubin pour obtenir une seule estimation finale.

Figure 1.2 – Représentation graphique des étapes de l'imputation multiple



Alors que l'étape d'analyse de multiples bases de données est simple pour la plupart des estimations, la phase d'imputation constitue une étape complexe, dont dépend la validité des estimations finales.

5.2.2. Phase d'imputation

La construction du modèle d'imputation est une étape cruciale incluant la sélection des variables prédictrices, la spécification du type du modèle et le choix du nombre *M* de bases à imputer.

A partir du cadre théorique général proposé par Rubin, différentes méthodes ont été élaborées et implémentées dans différents logiciels. Leurs particularités théoriques sont présentées, accompagnées d'un exemple illustratif de la méthode retenue dans la suite de ce travail.

- ***Cas général – algorithme de Rubin***

Si les données sont désignées par une matrice *Y*, comprenant une partie observée Y^{obs} et une partie manquante Y^{miss}, Y est distribuée selon la fonction $f(Y|\phi)$, où ϕ désigne l'ensemble des paramètres du modèle.

Sous l'hypothèse que les données sont manquantes au hasard, la distribution prédictive $f\left(Y^{miss}\middle|Y^{obs}\right)$ s'exprime comme :

$$f\left(Y^{miss}\middle|Y^{obs}\right)=\int f\left(Y^{miss},\phi\middle|Y^{obs}\right)d\phi$$
$$=\int f\left(Y^{miss}\middle|Y^{obs},\phi\right)f\left(\phi\middle|Y^{obs}\right)d\phi$$

Les données manquantes sont imputées en deux étapes. (i) Une valeur est estimée pour les paramètres d'après leur distribution a posteriori observée sur les données $f\left(\phi\middle|Y^{obs}\right)$. Les tirages dans les distributions de probabilité a posteriori sont réalisés à partir d'un algorithme de type Monte Carlo par chaîne de Markov (MCMC). Cette première étape permet d'obtenir une variabilité dans les paramètres. (ii) Les données manquantes sont estimées d'après leur distribution conditionnelle a posteriori $f\left(Y^{miss}\middle|Y^{obs},\phi\right)$ en utilisant la valeur de ϕ générée à la première étape.

- ***Modèle multivarié normal et modèle par équations chainées***

A partir de l'algorithme originel de Rubin, deux adaptations principales ont été élaborées [20]. Il s'agit de l'imputation multiple basée sur une distribution multivariée normale (Multivariate Normal Imputation, MVNI) originellement implémentée par Schafer [17] et de la méthode d'imputation multiple par équations chaînées (Multiple Imputation by Chained Equations, MICE), décrite par Van Buuren sous le terme de Regression *Switching* et plus récemment dénommée Fully Conditional Specification (FCS), et implémentée de façon indépendante par Van Buuren et al. [2], Raghunatan et al. [10], et Royston [40].

Le modèle multivarié normal est directement dérivé de l'algorithme de Rubin. Il est donc basé sur l'hypothèse que toutes les variables incluses dans le modèle d'imputation suivent une distribution multivariée jointe. Cette hypothèse n'est pas toujours vérifiée, en particulier lorsque le modèle inclut des variables binaires et catégorielles. Cependant, Schafer [29] suggère que les estimations obtenues à partir d'un modèle multivarié normal peuvent souvent être considérées comme valides, même si l'hypothèse de distribution multivariée normale n'est pas plausible. De ce fait, cette méthode a été largement utilisée, et est implémentée dans un logiciel gratuit (NORM), ainsi que dans une procédure SAS (MI et MIANALYZE) et STATA (version 11, MI IMPUTE) [39].

L'imputation par équations chaînées est une méthode plus flexible puisqu'elle ne fait pas l'hypothèse d'une distribution multivariée normale. En effet, une distribution est spécifiée pour chaque variable incomplète conditionnellement à toutes les autres variables incluses dans le modèle d'imputation. Concrètement, cette méthode permet de ramener un problème multivarié de dimension *k* en *k* problèmes univariés successifs conditionnant à chaque pas une variable imputée sur les valeurs observées et sur les valeurs les plus récentes générées des autres variables [41].

Le tirage dans les distributions conditionnelles s'effectue avec un algorithme spécifique dérivé des chaînes de Markov (l'échantillonneur de Gibbs), dont le principe est détaillé dans le paragraphe suivant. Le nombre d'itérations nécessaire à la convergence est faible, entre 5 et 20 dans la plupart des applications selon la taille du jeu de données et la proportion de cas incomplets, puisque la convergence se fait vers une distribution estimée et non une distribution exacte [41].

Cette approche est particulièrement flexible puisqu'une fonction de lien ainsi qu'un ensemble de variables prédictrices peuvent être spécifiés pour chaque variable incomplète, par exemple une régression logistique pour une variable binaire ou une régression multinomiale pour une variable catégorielle. L'hypothèse de normalité n'est ainsi plus requise que pour les variables continues.

Une limite théorique de cette méthode est due au fait que les distributions conditionnelles pourraient ne pas être compatibles avec la distribution jointe, ce qui causerait des problèmes de convergence du modèle d'imputation. Les répercussions pratiques ne sont cependant pas documentées dans la littérature [2;42], et quelques études de simulation de cas théoriques, montrant des distributions conditionnelles non compatibles avec la distribution jointe, tendent à montrer que les estimations sont non-biaisées [41;43].

Cette approche a d'abord été implémentée en tant que programme additionnel (ado) sous STATA (ICE), mais elle est à présent disponible comme fonction de base dans la version 12 de STATA. Elle est également disponible comme un ensemble de routines pouvant être appelé à partir de SAS (IVEware) et sous R (librairie MICE). Son utilisation est en progression constante [39].

Les résultats obtenus en appliquant ces deux méthodes sont variables selon les études. Ainsi, Faris et al. [35], Yu et al. [44] et Van Buuren et al. [43] concluent à partir d'études par

simulation que la méthode par équations chaînées donne des résultats plus fiables que l'approche multivariée normale en termes de biais et de couverture de l'intervalle de confiance. Dans une étude récente, Lee et al. [45] comparent les performances des deux méthodes à partir de données simulées de différents types, et montrent qu'elles constituent toutes les deux une approche valide. Ils concluent que, même si l'approche par équations chaînées se distingue par sa flexibilité et sa capacité à gérer les données discrètes, la méthode multivariée normale produit également des estimations valides quel que soit le type de données.

Un avantage majeur de l'imputation par équations chaînées réside cependant dans la gestion adéquate des variables discrètes, ainsi que dans la possibilité de spécifier un jeu de prédicteurs pour chaque variable, permettant d'inclure des liens complexes entre les variables. En relation avec le type de données que nous avons été amenés à traiter, à savoir des données essentiellement discrètes, nous avons fait le choix d'appliquer la méthode d'imputation multiple par équations chaînées. Un exemple d'imputation complexe présenté dans ce travail et réalisé à partir de la base de données de surveillance du VIH illustre bien les avantages liés à la flexibilité de cette approche (chapitre 4).

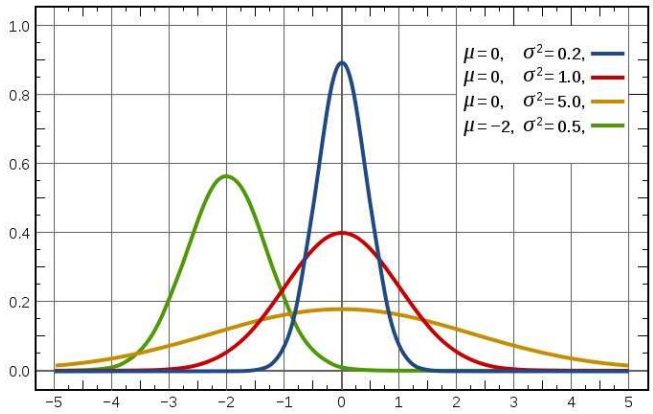
- ***Echantillonneur de Gibbs (d'après [3])***

- **Rappel sur les distributions de probabilité**

On note $P(X)$ la densité de probabilité de X qui est égale à la probabilité que X prenne la valeur x : $P(X) = P(X = x)$.

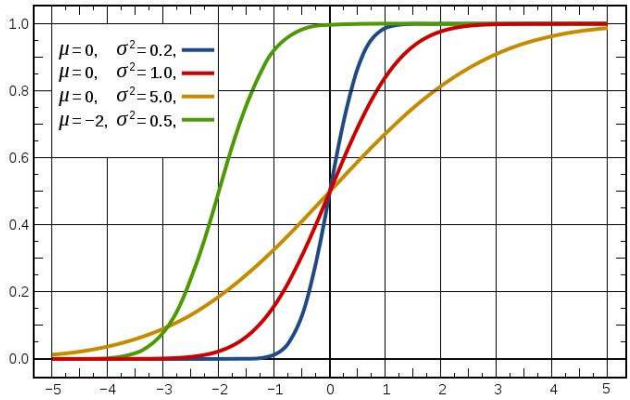
Si X suit une loi normale de moyenne μ et de variance σ^2 alors $P(X) = P(X = x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$.

Figure 1.3 – Densités de probabilité de distributions normales



On note $F(X)$ la fonction de répartition de X égale à la densité de probabilité cumulée : $F(X) = P(X \leq x)$. Si $X \sim N(\mu;\sigma^2)$, alors la fonction de répartition peut être représentée comme suit.

Figure 1.4 – Fonctions de répartition de distributions normales



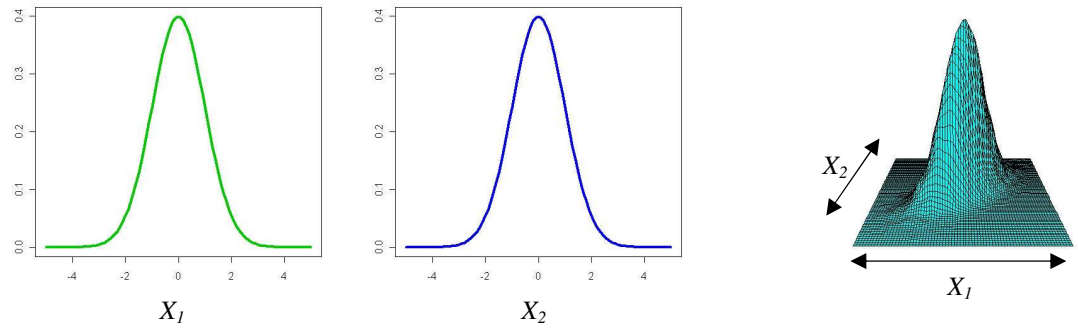
On tire une valeur dans la distribution de X en utilisant sa fonction de répartition. On tire un nombre aléatoire u entre 0 et 1 et on obtient la valeur x qui correspond à ce nombre aléatoire : $x = F^{-1}(u)$.

On note $P(X_1, X_2)$ la densité de probabilité du couple (X_1, X_2) égale à la probabilité que X_1 prenne la valeur x_1 et que X_2 prenne la valeur x_2 : $P(X_1, X_2) = P(X_1 = x_1, X_2 = x_2)$.

Exemple : $X_1 \sim N(\mu_1;\sigma_1^2)$ et $X_2 \sim N(\mu_2;\sigma_2^2)$

Figure 1.5 – Distributions marginales et jointes de X_1 et X_2

Distributions marginales		**Distribution jointe**
$P(X_1)$	$P(X_2)$	$P(X_1, X_2)$



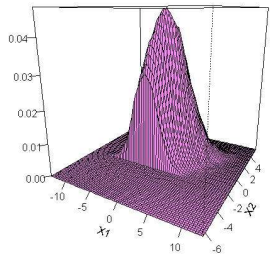
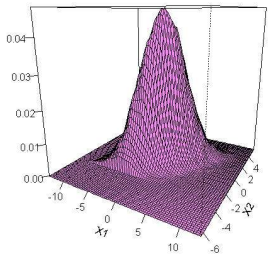
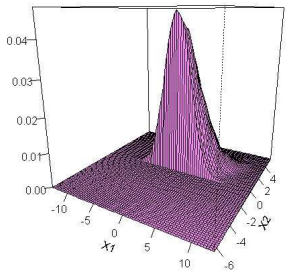
La distribution de X_1 conditionnellement à X_2 s'écrit $P(X_1|X_2)$ et la distribution de X_2 conditionnellement à X_1 s'écrit $P(X_2|X_1)$.

Ces deux distributions conditionnelles s'expriment en fonction de la distribution jointe et des distributions marginales par la relation de Bayes :

$$P(X_1|X_2) = \frac{P(X_1, X_2)}{P(X_2)} \text{ et } P(X_2|X_1) = \frac{P(X_1, X_2)}{P(X_1)}.$$

La figure 1.6 illustre des distributions conditionnelles $P(X_1|X_2)$ pour différentes valeurs de X_2.

Figure 1.6 – Exemples de distributions conditionnelles

$P(X_1 \mid X_2 = -3.02)$	$P(X_1 \mid X_2 = -1.08)$	$P(X_1 \mid X_2 = 0.21)$
		

- **Tirages avec l'échantillonneur de Gibbs**

Soit un ensemble de p variables aléatoires $X_1,...,X_p$. On souhaite tirer aléatoirement un ensemble de valeurs dans la distribution jointe de ces p variables aléatoires, notée $P(X_1,...,X_p)$. Il est cependant très difficile de tirer ces valeurs directement à partir de la distribution jointe. Une solution est d'utiliser l'échantillonneur de Gibbs. Celui-ci génère des valeurs issues de la distribution jointe, mais uniquement à partir des distributions conditionnelles :

$P(x_j \mid x_1,..., x_{j-1}, x_{j+1},..., x_p)$ pour tout $j = 1,..., p$.

L'algorithme de l'échantillonneur de Gibbs se décompose en trois étapes.

Etape 1 : Des valeurs initiales $x_1^{(0)},...,x_p^{(0)}$ sont choisies d'une certaine manière et on initialise $t=0$.

Etape 2 : On tire les valeurs selon les distributions conditionnelles de la manière suivante :

$$x_1^{(t+1)} \sim p\left(x_1 \middle| x_2^{(t)}, x_3^{(t)},...., x_p^{(t)}\right)$$

$$x_2^{(t+1)} \sim p\left(x_2 \middle| x_1^{(t+1)}, x_3^{(t)},...., x_p^{(t)}\right)$$

$$x_3^{(t+1)} \sim p\left(x_3 \middle| x_1^{(t+1)}, x_2^{(t+1)}, x_4^{(t)},...., x_p^{(t)}\right)$$

....

$$x_p^{(t+1)} \sim p\left(x_p \middle| x_1^{(t+1)}, x_2^{(t+1)},...., x_{p-1}^{(t+1)}\right)$$

Etape 3 : On incrémente *t* et on retourne à l'étape 2.

Etape 1		Cycle 1		Cycle 2			Cycle k
$x_1^{(0)}$	$\longrightarrow$	$x_1^{(1)}$	$\longrightarrow$	$x_1^{(2)}$	$\longrightarrow$	...	$x_1^{(k)}$
$x_2^{(0)}$	$\longrightarrow$	$x_2^{(1)}$	$\longrightarrow$	$x_2^{(2)}$	$\longrightarrow$	...	$x_2^{(k)}$
.		.		.			.
.		.		.			.
.		.		.			.
$x_p^{(0)}$	$\longrightarrow$	$x_p^{(1)}$	$\longrightarrow$	$x_p^{(2)}$	$\longrightarrow$	...	$x_p^{(k)}$

Le vecteur des valeurs tirées au sort $\left(x_1^{(k)}, x_2^{(k)},..., x_p^{(k)}\right)$, qui est issu de distributions conditionnelles successives, s'apparente, sous des conditions très générales et pour un nombre *k* de cycles suffisamment grand, à un vecteur de valeurs qui aurait été tiré de la distribution jointe de $\left(X_1, X_2,...X_p\right)$ que l'on n'a pas eu à déterminer [46].

Illustration 1 :

On se place dans le cas où l'on ne dispose que de deux variables X_1 et X_2. La variable X_1 a des données manquantes et la variable X_2 est complète.

X1	X2
1	18
0	24
1	33
0	67
0	49
1	43
.	26
.	61
.	45

Notons qu'il n'y a pas ici de notion de variable explicative et de variable à expliquer. Toutes les variables du modèle d'imputation sont alternativement des variables explicatives et des variables à expliquer, au sens statistique.

Etape 1 : On remplace les données manquantes de X_1 par des valeurs initiales $\left(x_1^{(0)}\right)$. Celles-ci peuvent être tirées de la distribution des valeurs observées.

X1	X2
1	18
0	24
1	33
0	67
0	49
1	43
.	26
.	61
.	45

X1	X2
1	18
0	24
1	33
0	67
0	49
1	43
0	26
1	61
0	45

Etape 2 : On spécifie la distribution de X_1 conditionnellement à X_2. Dans notre cas, X_1 est une variable binaire et on suppose que :

$$P\left(x_{i1}\middle|x_{i2,\beta}\right)=\left(\frac{\exp\left(x_{i2}\beta\right)}{1+\exp\left(x_{i2}\beta\right)}\right)^{x_{i1}}\times\left(\frac{1-\exp\left(x_{i2}\beta\right)}{1+\exp\left(x_{i2}\beta\right)}\right)^{1-x_{i1}}.$$

Cette spécification entraîne le choix d'un modèle d'imputation, qui est un modèle de régression incluant des variables explicatives, ainsi que de la fonction de lien entre la variable qui présente des données manquantes et ces variables explicatives. Puisque la variable est binaire, le lien est une fonction logit.

Le paramètre β et sa variance sont estimés en maximisant une fonction de vraisemblance.

```
logit X1 X2

Iteration 0:   log likelihood = -6.1826542
Iteration 1:   log likelihood = -6.1298599
Iteration 2:   log likelihood = -6.1298463

Logistic regression                                     Number of obs   =          9
                                                        LR chi2(1)      =       0.11
                                                        Prob > chi2     =     0.7452
Log likelihood = -6.1298463                             Pseudo R2       =     0.0085

------------------------------------------------------------------------------
          X1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          X2 |  -.0138385   .0428534    -0.32   0.747    -.0978295    .0701526
       _cons |   .3367752   1.851899     0.18   0.856    -3.292881    3.966431
------------------------------------------------------------------------------
```

On obtient une estimation du coefficient $\hat{\beta}=-0.0138385$ et de sa variance $\hat{V}\left(\hat{\beta}\right)=0.0428534^2$.

On tire une nouvelle valeur β^* à partir de la distribution supposée normale de β, $N\left(\hat{\beta}=-0.0138385,\hat{V}\left(\hat{\beta}\right)=0.0428534^2\right)$.

On impute les valeurs manquantes. Pour chaque individu *i* pour lequel la variable X_1 est manquante :

- On calcule sa probabilité d'être $X_1 = 1$: $w_i = \dfrac{\exp(x_i\beta^*)}{1+\exp(x_i\beta^*)}$

- On tire un nombre aléatoire entre 0 et 1 : $u_i \sim unif(0.1)$. Si $u_i > w_i$ on impute $X_{i1} = 0$, sinon on impute $X_{i1} = 1$.

X1	X2	w	u
1	18		
0	24		
1	33		
0	67		
0	49		
1	43		
0	26	**0.667**	**0.876**
1	61	**0.836**	**0.658**
0	45	**0.769**	**0.036**

X1	X2
1	18
0	24
1	33
0	67
0	49
1	43
0	26
1	61
1	45

Variante (1) – Bootstrap (non paramétrique)

Au lieu de tirer une valeur β^* dans la distribution de β en faisant l'hypothèse que cette distribution est normale, on peut introduire de l'incertitude, non pas au niveau du paramètre mais au niveau des données elles-mêmes. Pour cela, on tire au sort avec remise un échantillon d'individus (échantillon bootstrapé) et on estime β à partir de ce jeu de données. Le principal intérêt est de ne pas avoir à faire l'hypothèse d'une distribution normale pour β.

Variante (2) – Prédiction matching

Au lieu de calculer les quantités w_i et u_i pour les individus ayant des données manquantes et de les comparer ensuite pour imputer les valeurs, on calcule w_i pour tous les individus. Pour chaque individu ayant une donnée manquante, on cherche l'individu n'ayant pas de donnée manquante avec un w_i le plus proche du sien. On lui attribue alors la valeur correspondante.

Illustration 2 :

On se place dans le cas où l'on ne dispose que de deux variables X_1 et X_2, toutes deux incomplètes.

Etape 1 : On remplace les données manquantes de X_1 et X_2 par des valeurs initiales $\left(x_1^{(0)}, x_2^{(0)}\right)$. Celles-ci peuvent être tirées dans chacune des distributions des données observées.

X1	X2
1	18
0	24
1	33
0	.
0	.
1	43
.	.
.	61
.	45

X1	X2
1	18
0	24
1	33
0	**45**
0	**38**
1	43
0	**29**
1	61
0	45

Etape 2 : On entre dans le cycle 1 de l'échantillonneur de Gibbs. On impute les données manquantes de X_1 comme précisé précédemment : $x_1^{(1)} \sim p\left(x_1 \middle| x_2^{(0)}\right)$. Puis on impute les données manquantes de X_2 en utilisant les valeurs que l'on vient d'imputer à X_1 : $x_2^{(1)} \sim p\left(x_2 \middle| x_1^{(1)}\right)$. Le cycle 1 est achevé.

Cycle 1

X1	X2
1	18
0	24
1	33
0	**45**
0	**38**
1	43
0	**29**
1	61
1	45

X1	X2
1	18
0	24
1	33
0	**43**
0	**36**
1	43
0	**25**
1	61
0	45

On rentre ensuite dans le cycle 2 de l'échantillonneur de Gibbs. On impute les données manquantes de X_1 : $x_1^{(2)} \sim p\left(x_1 \middle| x_2^{(1)}\right)$. Puis on impute les données manquantes de X_2 en utilisant les valeurs que l'on vient d'imputer à X_2 : $x_2^{(1)} \sim p\left(x_2 \middle| x_1^{(2)}\right)$. Le cycle 2 est achevé.

		Cycle 1				Cycle2			
X1	**X2**	**X1**	**X2**	**X1**	**X2**	**X1**	**X2**	**X1**	**X2**
1	18	1	18	1	18	1	18	1	18
0	24	0	24	0	24	0	24	0	24
1	33	1	33	1	33	1	33	1	33
0	67	0	45	0	43	0	43	0	42
0	49	0	38	0	36	0	36	0	35
1	43	1	43	1	43	1	43	1	43
.	26	0	29	0	25	0	25	0	26
.	61	1	61	1	61	0	61	0	61
.	45	1	45	0	45	1	45	1	45

On recommence ce processus jusqu'à ce que les valeurs imputées ne varient plus d'un cycle à l'autre. On a alors atteint la convergence de l'échantillonneur.

On retient les valeurs obtenues au dernier cycle et l'ensemble de ces valeurs vont former une base de données complète. En pratique, le nombre de cycles est fixé par l'utilisateur, en général 10 cycles. Le processus complet est répété *M* fois pour obtenir *M* bases de données complètes.

5.2.3. Phase d'analyse

La séparation de la phase d'imputation et d'analyse représente un des avantages majeurs de la méthode d'imputation multiple. En effet, une équipe peut ainsi réaliser l'imputation et livrer une base de données complète qui pourra ensuite être analysée selon les procédures classiques par des utilisateurs variés, portant sur l'ensemble de la base de données ou sur des sous-parties. C'est dans cette optique que Rubin avait envisagé à l'origine le processus d'imputation multiple.

Les résultats des analyses individuelles portant sur chacune des *M* bases de données sont ensuite combinés selon des règles édictées par Rubin [3] et présentées ici pour un seul paramètre d'intérêt.

Soit Q le paramètre à estimer (proportion, moyenne, coefficient de régression) et U sa variance. L'imputation de M bases de données conduit à M estimations de Q et U, notées $\hat{Q}_m$ et $\hat{U}_m$ pour $m = 1,...,M$.

L'estimateur combiné $\hat{Q}^*$ est la moyenne des $\hat{Q}_m$ pour les M imputations :

$$\hat{Q}^* = \frac{1}{M}\sum_{m=1}^{M}\hat{Q}_m$$

La variance combinée $\hat{U}^*$ se décompose en deux parties.

La variance intra-imputation $\overline{U}$ rend compte des variances de chacun des $\hat{Q}_m$. Elle est estimée par la moyenne des M variances :

$$\overline{U} = \frac{1}{M}\sum_{m=1}^{M}\hat{U}_m$$

La variance inter-imputation B rend compte de la variance de chacun des $\hat{Q}_m$ par rapport à l'estimateur combiné. Elle correspond à la variance des moyennes a posteriori des $\hat{Q}_m$:

$$B = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{Q}_m - \hat{Q}^*\right)^2$$

La variance totale ou combinée $\hat{U}^*$ est la somme pondérée de la variance intra-imputation et de la variance inter-imputation :

$$\hat{U}^* = \overline{U} + \left(1 + \frac{1}{M}\right)B$$

Une approximation de Student est utilisée pour les tests et le calcul des intervalles de confiance : $\left(\hat{Q}^* - Q\right)\Big/\sqrt{\hat{U}^*} \sim t_\nu$, avec le nombre de degrés de liberté égal à $\nu = (M-1)\left[1 + \frac{\overline{U}}{(1+1/M)B}\right]^2$.

Les phases d'imputation et d'analyse sont réalisées par le biais de procédures dédiées, et s'effectuent de façon automatisée. Lors de la phase d'imputation, il est seulement nécessaire de choisir certains paramètres comme le nombre de cycles de l'échantillonneur et le nombre de bases à imputer, ainsi que certaines options spécifiques de la méthode. La phase d'analyse est encore plus aisée, puisque les résultats combinés sont obtenus directement avec les commandes classiques précédées de commandes propres à l'imputation multiple, sachant que les résultats des analyses intermédiaires n'apparaissent pas.

Du fait de la simplification des procédures implémentées sous les différents logiciels, l'imputation multiple apparaît d'une application abordable. Cependant, même si certains cas simples ne requièrent pas d'expertise particulière, l'imputation multiple nécessite une mise en œuvre méthodique et rigoureuse, aussi bien pour la construction du modèle d'imputation que pour l'interprétation des résultats et la vérification des hypothèses. C'est ce processus que nous allons maintenant détailler.

5.3. Etapes pratiques de la mise en application

En généralisant le cas présenté précédemment (paragraphe 2.3.3), on considère le cas d'une imputation réalisée en vue d'une analyse étiologique expliquant une variable complète représentant la maladie et notée M à partir de plusieurs variables d'exposition incomplètes E_i.

5.3.1. Etape 1 : Examen et analyse de la base de données incomplète

- ***Examen préliminaire de la base de données***

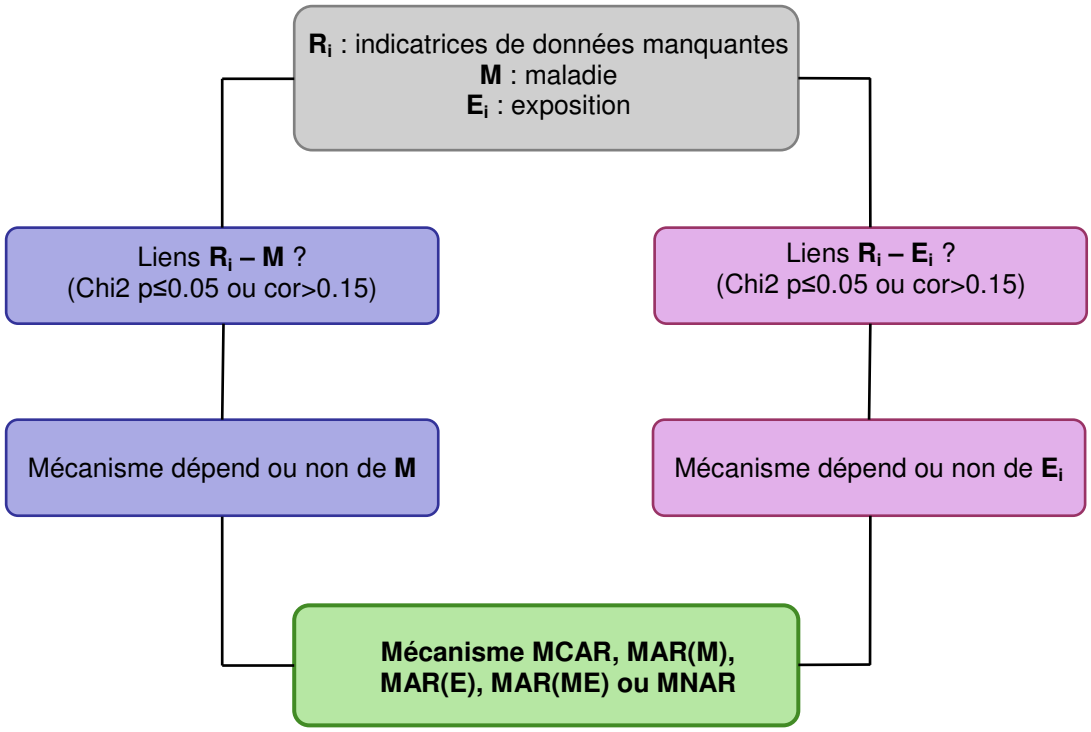
Dans un premier temps, cet examen consiste à quantifier le nombre de variables incomplètes ainsi que la proportion de données manquantes de chaque variable. Cette étape simple permet de sélectionner les variables qui peuvent être retenues pour un processus d'imputation, soit parce qu'elles seront incluses dans les analyses ultérieures, soit parce qu'elles ont des capacités prédictrices pour les variables incomplètes.

Selon la taille de la base de données d'origine, il est parfois intéressant de constituer une base restreinte limitée à ces variables, en restreignant la proportion de données manquantes maximale à environ 30%. Ce chiffre est cité à titre d'exemple car il n'existe pas de limite théorique à la proportion de données manquantes que l'on peut traiter par imputation multiple [47;48]. Cette limite peut être fixée de façon empirique en fonction du nombre de variables incomplètes mais aussi selon la proportion d'information annexe disponible dans la base de données. Des outils diagnostiques appliqués après imputation permettent de valider ce choix.

Il est informatif à ce stade d'identifier la typologie des données manquantes afin de décrire le motif et le mécanisme des données manquantes. Des commandes spécifiques permettent de visualiser le motif de répartition des données manquantes de toutes les variables ainsi que d'évaluer la perte d'effectifs attendue lors d'une analyse cas complet. Afin d'identifier le mécanisme de données manquantes, on associe à chaque variable d'exposition incomplète E_i une indicatrice de données manquantes R_i binaire (codée 1 si E_i n'est pas renseignée et 0 sinon). On croise chaque indicatrice R_i avec, d'une part la variable à expliquer M , d'autre part chacune des variables d'exposition E_i (Figure 1.7). Le lien statistique recherché est un test du Chi2 significatif ($p \leq 0.05$) pour des variables binaires ou catégorielles, ou bien un coefficient de corrélation supérieur à 0.15 pour des variables continues [42]. Cet examen univarié simple permet de proposer une première synthèse des mécanismes de données manquantes observés. Un examen multivarié peut également être réalisé.

Cette analyse permet généralement d'exclure un mécanisme MCAR pour une variable donnée si au moins un des croisements avec la variable à expliquer (maladie) ou une des variables explicatives (expositions) est significatif. Le mécanisme de données manquantes n'est alors pas complètement aléatoire car il dépend au moins d'une autre variable de la base de données. Par ailleurs, cet examen permet d'identifier pour chaque variable un mécanisme de type MAR(M), MAR(E) ou MAR(ME), c'est-à-dire dépendant seulement de la maladie, seulement d'une ou plusieurs variables d'exposition, ou conjointement des deux.

Figure 1.7 – Représentation graphique du processus d'identification du mécanisme de données manquantes



A partir des résultats de cet examen présentés dans le Tableau 1.1, il est possible de dégager un mécanisme de données manquantes global, c'est-à-dire d'anticiper le risque de biais en analyse cas-complet si un mécanisme de type MAR(ME) est identifié. Il faut noter que l'on ne peut exclure un mécanisme de type MNAR si des liens existent entre des indicatrices de données manquantes et des variables incomplètes, car le mécanisme de données manquantes peut aussi dépendre des valeurs non-observées de ces variables.

A la suite de cet examen statistique des mécanismes de données manquantes, il est important de proposer des hypothèses épidémiologiques pour ces mécanismes, tout particulièrement lorsqu'un mécanisme de type MNAR est envisagé.

Tableau 1.1 – Exemples de mécanismes de données manquantes avec une variable à expliquer (M) et 4 variables d'expositions (Ei)

	R_1	R_2	R_3	R_4
M			**X**	**X**
E_1		X		
E_2				X
E_3		X		
E_4				
Mécanisme	MCAR	MAR(E)	MAR(M)	MAR(ME)

R_1 est la variable indicatrice de données manquantes de E_1, R_2 de E_2 etc.
La zone en gris clair représente les croisements entre les indicatrices de données manquantes et la variable à expliquer, et les zones en gris foncé les intersections entre chaque variable d'exposition et son indicatrice de données manquantes.
X représente une liaison significative entre les variables définie comme (i) un coefficient de corrélation >0.15 ou un test du Chi2 significatif ($p \leq 0.05$) en analyse univariée, ou (ii) un test de Wald significatif ($p \leq 0.05$) en analyse multivariée.

- ***Analyse Cas Complet***

Une analyse cas complet approfondie doit être réalisée lors de cette phase préliminaire. On effectue tout d'abord une analyse univariée afin de sélectionner les variables qui seront retenues pour les analyses multivariées cas-complet et imputation multiple, et d'identifier les termes d'interaction potentiels entre les variables. Puis, l'analyse multivariée permet d'une part de mesurer l'impact des données manquantes sur le processus de sélection des variables, et d'autre part d'obtenir des estimations qui pourront être comparées avec les résultats obtenus après imputation multiple. L'analyse cas-complet est donc une étape indispensable car elle permet d'appréhender les relations entre les variables, relations qui seront prises en compte lors du processus d'imputation.

5.3.2. Etape 2 : Construction du modèle d'imputation

- ***Sélection des variables prédictrices***

La spécificité de la méthode d'imputation par équations chaînées est de permettre l'élaboration d'un modèle d'imputation conditionnelle pour chaque variable incomplète. Ainsi, une fonction de lien et un ensemble de variables prédictrices sont spécifiés pour chaque variable. La sélection des variables prédictrices, automatisée pour certaines implémentations (IVEware, MICE), peut également être réalisée manuellement en suivant des critères de sélection simples.

Rubin [3] propose d'inclure le plus de variables explicatives possible, tout en construisant un modèle d'imputation de façon raisonnée, c'est-à-dire en respectant les relations entre les variables. Selon lui, il est en principe plus dangereux d'omettre des variables explicatives que d'en inclure trop dans le modèle d'imputation. Rubin fait alors référence à un type particulier d'imputation qui consiste à estimer les données manquantes dans des bases de données importantes, et ceci sans tenir compte des analyses qui seront réalisées par la suite sur des sous-parties des bases complétées.

Meng [49] précise la méthode de sélection des variables en introduisant la notion d'homogénéité entre la procédure d'imputation et les analyses ultérieures. Ainsi, il montre qu'il est nécessaire d'inclure au minimum dans le modèle d'imputation toutes les variables retenues pour les analyses ultérieures, sous peine d'introduire des biais. Concrètement, si on réalise une analyse étiologique à partir d'une base de données incomplète, on doit inclure dans le modèle d'imputation les covariables retenues au terme de l'analyse univariée cas-complet ainsi que la variable à expliquer.

Bien que cette démarche ne paraisse pas intuitive, il est nécessaire d'inclure la variable à expliquer dans le modèle d'imputation afin de reproduire le niveau de lien entre les valeurs observées de la variable à expliquer et les valeurs imputées des covariables incomplètes, tel qu'il existe pour les autres données observées de la base [50;51;52]. Si la variable à expliquer n'est pas retenue, le niveau de lien entre la variable à expliquer et les covariables sera sous-estimé lors de l'imputation, ce qui faussera l'analyse étiologique ultérieure [20]. Ces variables, retenues systématiquement pour l'imputation, sont dénommées variables principales. Elle composent le modèle d'imputation le plus réduit.

Cependant, suivant les indications princeps de Rubin [30], Van Buuren [42] propose une stratégie de construction du modèle d'imputation consistant à sélectionner le plus de variables prédictrices possible, dans la mesure où elles contiennent des informations sur le mécanisme de données manquantes des variables incomplètes. Il faut donc retenir d'une part les variables liées aux variables à imputer, et d'autre part les variables liées au mécanisme de données manquantes des variables à imputer. Ces variables additionnelles, dites variables auxiliaires [14], permettent d'améliorer les capacités prédictrices du modèle d'imputation [53], même lorsqu'elles ne sont pas retenues dans les analyses ultérieures des données imputées. Le modèle d'imputation est alors plus général que le modèle d'analyse. Il est conseillé de retenir

un total de 15 à 25 variables prédictrices [42], mais nous verrons qu'en pratique il est peu courant de pouvoir sélectionner plus de 15 variables.

L'ajout de variables auxiliaires permet de capturer au mieux le mécanisme de données manquantes et ainsi de rendre l'hypothèse MAR plus plausible, puisque la modélisation des variables incomplètes dépend davantage de valeurs observées et donc dans une moindre mesure de valeurs non-observées [10;14]. Van Buuren [42] a ainsi démontré dans une étude par simulation que l'imputation de données manquantes selon un mécanisme MNAR était nettement améliorée par l'ajout de variables auxiliaires, c'est-à-dire que les estimateurs étaient moins biaisés par rapport à la vraie valeur que lors d'une imputation réalisée à partir d'un modèle plus réduit.

En pratique, la sélection des variables s'effectue en deux temps. Tout d'abord, on retient les variables sélectionnées pour les analyses ultérieures, les variables principales, puis on identifie les variables auxiliaires en croisant les indicatrices de données manquantes des variables incomplètes avec les variables disponibles dans la base de données.

La sélection des variables prédictrices, principales et auxiliaires, est présentée par Van Buuren [42] comme une recherche de corrélations entre les variables deux à deux, aussi bien pour les liens entre variables incomplètes, qu'entre indicatrices de données manquantes et variables incomplètes ou/et complètes. Il est cependant informatif de tenir compte dans cette sélection de l'ajustement sur les autres variables, c'est-à-dire de se placer dans un cas multivarié. On sélectionne ainsi les variables prédictrices par régression de chaque indicatrice de données manquantes et de chaque variable à imputer sur l'ensemble des variables de la base de données [41]. Cependant, si cette stratégie permet de prendre en compte les relations entre les données utilisées lors du processus d'imputation, ces analyses multivariées sont réalisées en cas-complet et peuvent donc être affectées par une perte de puissance ainsi que par un biais de sélection du fait des données manquantes.

Il a donc été proposé par Wood et al.[54] et plus récemment par White et al. [28] d'effectuer cette sélection à partir des données imputées. Cette approche, qui peut paraître circulaire, a été retenue dans notre travail essentiellement comme un procédé de validation du jeu de variables prédictrices retenues. En effet, sachant qu'une imputation multiple valide doit préserver les relations entre les variables, il parait cohérent d'examiner les liens entre les variables à partir des données imputées puisqu'ils doivent refléter les liens initiaux entre les variables incomplètes.

Notre approche a consisté à construire un modèle multivarié expliquant chaque indicatrice de données manquantes à partir des covariables (variables explicatives et variable à expliquer), puis à appliquer pour chaque modèle une stratégie de sélection pas à pas descendante portant sur l'ensemble des bases imputées, par le biais des commandes propres à l'imputation. Les résultats obtenus sont utilisés pour valider la sélection réalisée en analyse cas-complet univariée. Les applications pratiques présentées dans ce rapport montrent que le processus multivarié est plus sélectif que le processus univarié, et nous avons fait le choix de privilégier le modèle d'imputation le plus complet, tout en testant des modèles plus parcimonieux.

- ***Type des variables incomplètes***

Une fonction de lien est spécifiée pour chaque variable selon son type. Les variables sont imputées par régression logistique pour les variables binaires, régression multivariée pour les variables nominales ou ordinales et régression linéaire pour les variables continues.

Les variables catégorielles sont bien prises en compte par la méthode d'imputation par équations chaînées. Il a été démontré que la méthode d'imputation multivariée normale, en traitant les variables catégorielles comme des variables continues, peut induire des biais dans les estimations [55]. A l'inverse, la méthode par équations chaînées permet de spécifier une fonction de lien multinomiale pour les variables catégorielles. Ainsi, chaque variable catégorielle est décomposée en un jeu de variables indicatrices binaires qui sont utilisées pour l'estimation des autres variables incomplètes [56].

L'imputation des variables continues repose sur l'hypothèse d'une distribution normale. Or les résultats de l'imputation multiple sont sensibles au non-respect de la normalité aussi bien pour l'approche multivariée normale que pour l'approche par équations chaînées [45]. Il est donc recommandé de transformer chaque variable dont la distribution s'écarte de la normalité pour assurer la validité de l'imputation de cette variable, mais aussi celle des autres variables incomplètes du modèle. Une transformation inverse est dans ce cas appliquée après imputation.

- ***Choix du nombre de bases à imputer***

Les indications de la littérature, essentiellement basées sur les règles édictées par Rubin, stipulent qu'un nombre restreint d'imputations, c'est-à-dire de 3 à 5, est suffisant. Cet argument standard est basé sur la notion d'efficacité statistique relative des estimateurs, qui

évalue l'efficacité de l'utilisation d'un nombre fini M d'imputations par rapport à un nombre infini d'imputations.

Si la fraction d'information manquante, que l'on dénotera *FMI* (Fraction of Missing Information), est définie comme une fonction des variances intra-imputation (W) et inter-imputation (B) avec $FMI = \frac{B}{W+B}$, alors l'efficacité statistique relative (Relative Efficiency, *RE*) est donnée par $RE = 1 + \frac{FMI}{m}$.

La valeur de la FMI d'une variable dépend forcément de la proportion de données manquantes initiale de cette variable, et Bodner [57] a proposé d'approximer la FMI par cette proportion. Cependant, la FMI dépend également du processus d'imputation, et représente un indicateur de la qualité de l'imputation pour chaque variable, selon qu'elle est égale, inférieure ou supérieure à la proportion de données manquantes.

Le Tableau 1.2, d'après [3], illustre les variations d'efficacité statistique relative (RE) selon la *FMI* et *M*. Si l'on tolère une perte d'efficacité de 5% $(RE = 95\%)$, alors $\frac{FMI}{M} \leq 0.05$ et donc $M = 5$ est adéquat si $FMI \leq 0.25$. On se place ici dans un cas univarié, ce qui signifie qu'il suffit d'imputer 5 bases pour obtenir une efficacité statistique relative de 95% si la variable incomplète a une FMI de 25%. Pour une même efficacité statistique, il suffirait d'imputer 10 bases pour prendre en compte une FMI de 50%.

Tableau 1.2 – Efficacité statistique relative en % selon la FMI et le nombre de bases imputées M

M	FMI					
	0,1	0,2	0,3	0,5	0,7	0,9
2	95	91	87	80	74	69
3	97	94	91	86	81	77
5	98	96	94	91	88	85
10	99	98	97	95	93	92
20	100	99	99	98	97	96

Formulé autrement, une perte d'efficacité statistique relative correspond à une inflation de la variance de l'estimation. Ainsi, tolérer une efficacité statistique de 95% implique une augmentation de la variance d'un facteur de 1.02 pour une proportion de données manquantes de 25% et $M = 5$. Ce facteur atteint 1.08 pour une FMI=50% et $M = 10$.

Cependant, de nouveaux arguments en faveur d'un nombre d'imputations plus important sont décrits dans la littérature. Ainsi, Graham et al. [58] reprennent les arguments d'efficacité statistique proposés par Rubin en les précisant. Ils observent par simulation une baisse notable de puissance statistique, avec un impact sur la largeur de l'intervalle de confiance et sur la p-valeur, baisse supérieure à 5% pour une $FMI \geq 0.25$ et pour $M = 10$, par rapport à 100 imputations. Selon cet exemple, le nombre de bases qu'il faudrait imputer pour limiter la perte de puissance statistique à 5% serait le double de celui retenu selon les arguments précédents, pour une *FMI* similaire.

White et al. soulignent cependant que les arguments d'efficacité statistique relative et de puissance statistique ne sont pas suffisants pour déterminer le nombre de bases à imputer. Comme suggéré par Horton et al. [38], il faudrait tenir compte dans le choix du nombre de bases de la stabilité des résultats obtenus en analysant des jeux de données imputées à partir de la même base de données initiale (et le même modèle d'imputation). On se base ainsi sur l'estimation de l'erreur de Monte Carlo, donnée par l'expression $MC_{error} = \sqrt{B/M}$ qui tend vers 0 lorsque M augmente. A partir d'une étude par simulation, Bodner [57] propose d'estimer l'erreur de Monte Carlo d'un paramètre β. Afin de valider le choix du nombre de bases imputées, on peut retenir comme critère que l'erreur de Monte Carlo de $\hat{\beta}$ doit être égale à environ 10% de l'écart type de $\hat{\beta}$.

De façon à s'assurer que tous ces critères sont remplis, White et al. suggèrent de choisir, quand cela est possible, un nombre de bases imputées M tel que $M \geq 100 \times FMI$. En approximant la FMI par la proportion de données manquantes, il faudrait alors imputer, si une seule variable est incomplète, environ 100 fois la proportion de données manquantes de cette variable.

En pratique, il est rare d'avoir à traiter une base de données contenant une seule variable incomplète et la FMI peut différer de la proportion de données manquantes. Nous verrons dans les applications présentées qu'un diagnostic simple du nombre de bases imputées peut être obtenu en combinant ces différents critères, et qu'il est souvent difficile en pratique d'imputer un nombre de bases adéquat. Il reste cependant intéressant d'explorer ces critères diagnostiques, tout particulièrement lorsqu'un processus d'imputation peut être amélioré au cours du temps, comme c'est le cas pour les données de systèmes de surveillance.

- ***Autres options***

Nombre de cycles

Plusieurs cycles sont nécessaires seulement si plusieurs variables sont incomplètes. L'échantillonneur de Gibbs converge rapidement (5 à 10 itérations) et le programme additionnel ICE spécifie 10 cycles par défaut. Dans la littérature, les auteurs observent peu de variations au-delà de 20 cycles [35;42;59]. En pratique, le temps d'imputation varie peu entre 10 et 30 cycles, ce qui peut autoriser un choix conservateur.

Racine

La racine ("seed") est un nombre que l'on peut spécifier dans les options du modèle d'imputation. C'est un nombre qui initialise le générateur de nombres aléatoires. Si ce nombre est spécifié, il est alors possible d'obtenir un jeu de données identique lorsque l'on répète le processus d'imputation à partir du même modèle. Dans le cas contraire, une racine est générée par ICE et chaque nouvelle imputation donnera des résultats potentiellement différents.

5.3.3. Etape 3 : Analyse des données imputées et présentation des résultats

- ***Diagnostic de l'imputation multiple***

Deux conditions sont requises afin d'obtenir des inférences valides après imputation multiple : (i) le modèle d'imputation doit être correctement spécifié et (ii) l'hypothèse MAR doit être plausible, ou du moins l'impact d'un processus MNAR doit être limité.

Il est recommandé de tester plusieurs bases de données imputées issues de modèles d'imputation différents. Ceux-ci peuvent être plus ou moins généraux, c'est-à-dire inclure plus ou moins de variables prédictrices, et plusieurs transformations de variables peuvent être évaluées (par exemple lors de recodage de variables catégorielles ou de transformation de variables continues).

Les implémentations classiques de l'imputation multiple requièrent que les données soient manquantes aléatoirement, c'est-à-dire qu'elles ne dépendent que des données observées, et non de données non-observées. Une approche théorique pour tester cette hypothèse a été proposée par Pothoff et al. [60] mais elle n'est pas applicable en pratique courante (sans implémentation dans un logiciel).

Lorsque des informations issues de la littérature ou spécifiques au mode de recueil de données laissent supposer que certaines variables peuvent contenir des données MNAR, les estimations issues de l'imputation multiple peuvent être biaisées. Deux approches sont envisageables : (i) modéliser la non-réponse pour ces variables en appliquant des modèles par sélection ou par mélange à partir d'hypothèses sur le mécanisme de données manquantes, (ii) appliquer une méthode d'analyse de sensibilité afin d'évaluer l'impact d'un mécanisme MNAR sur les résultats de l'imputation multiple.

En pratique, même si la procédure d'imputation multiple ne peut être validée à partir des valeurs observées, les résultats obtenus peuvent être vérifiés en tenant compte de standards raisonnables. Ainsi, les différences entre données observées et données imputées peuvent être testées, afin d'évaluer si ces variations ont un sens dans le contexte du recueil de données. En effet, avant la réalisation de l'imputation, une analyse des données observées permet de définir des mécanismes MAR dépendant de certaines variables, et ainsi d'anticiper des variations entre données observées et imputées.

Abayomi et al. [61] proposent des comparaisons numériques et graphiques afin de dépister des anomalies dans le processus d'imputation, appliquées consécutivement à plusieurs bases de données imputées. Comme souligné par Raghunatan et al. [62], les comparaisons numériques ne sont pas toujours interprétables en raison de la différence d'effectifs entre valeurs observées et imputées, car elles peuvent être de ce fait artificiellement significatives. Les diagnostics graphiques permettent de tester aisément les données imputées, sur une seule base ou sur l'ensemble des bases imputées. Dans le cas des variables continues, l'imputation se fait sous l'hypothèse d'une distribution normale, ce qui implique souvent une transformation préalable de ces variables. Une analyse graphique permet ainsi de dépister des défauts de superposition des distributions observées et imputées.

Cette première étape de validation du processus d'imputation doit prendre place avant la phase d'analyse. L'expérience montre que, bien qu'elle repose sur des examens qualitatifs, cette étape est cruciale pour s'assurer (i) de la validité du modèle d'imputation retenu, (ii) de la plausibilité de l'hypothèse MAR. Dans le cas où les tests diagnostiques montrent des variations inexpliquées, et si le modèle d'imputation parait valide, il est important d'envisager une analyse de sensibilité sur la base d'hypothèses épidémiologiques. Cette approche est illustrée dans le chapitre 3. Une autre approche consiste à appliquer une procédure de validation croisée [63] à partir d'échantillons simulés selon plusieurs mécanismes de données manquantes. Cette procédure permet de tester l'étendue des biais sur les estimateurs, ainsi que la couverture des intervalles de confiance selon les mécanismes de données manquantes générés. Elle est illustrée dans le chapitre 4.

- ***Analyse jointe des bases de données imputées***

Puisque les variables sont imputées sous leur forme originelle, une transformation inverse n'est nécessaire que pour les variables continues qui ont dû être normalisées avant imputation. Selon les analyses prévues, il peut être nécessaire de générer des variables à partir des variables imputées. Elles sont alors créées automatiquement dans l'ensemble des M bases de données.

Pour les analyses descriptives, les estimations sont assorties d'un intervalle de confiance rendant compte de la variabilité liée au processus d'estimation. Dans le cas des analyses étiologiques, le calcul de la variance intègre la variabilité entre les bases imputées, et les principales fonctions de lien peuvent être spécifiées. La prise en compte d'un plan de sondage est également prévue pour les implémentations classiques et pour certaines régressions [64].

La réalisation d'une analyse multivariée implique une stratégie de sélection de variables basée sur le test d'hypothèses. Après imputation multiple, le test de Wald est approximé par un test de Student pour tester les coefficients de régression. Un test de Fisher permet également de tester conjointement une série de coefficients de régression sur l'ensemble des bases imputées [65].

Cependant, certaines statistiques ne peuvent être obtenues directement à partir de données imputées, car elles ne peuvent être combinées sur les *M* bases imputées selon les règles de Rubin. White et al. [28] ont récemment synthétisé cette problématique pour les statistiques les plus fréquemment utilisées (Tableau 1.3).

Tableau 1.3 – Statistiques courantes pouvant être combinées ou non selon les règles de Rubin (d'après [28]).

Statistiques pouvant être combinées sans transformation	Moyenne, proportion, coefficient de régression, C-index, aire sous la courbe ROC
Statistiques nécessitant une transformation adéquate pour être combinées	Odds ratio, risque relatif, probabilité de survie, écart-type, corrélation, proportion de variance expliquée, skewness, kurtosis
Statistiques ne pouvant pas être combinées	p-valeur, test de rapport de vraisemblance, test du Chi2 du modèle, test d'adéquation du modèle

Les critères statistiques permettant la sélection du meilleur modèle tels que le test de rapport de vraisemblances, le test de la déviance, les critères d'adéquation AIC (Akaïke Information Criterion) et BIC (Bayesian Information Criterion) ainsi que le test du Chi2 du modèle (test d'adéquation) ne peuvent pas être obtenus directement. Une approximation de la statistique du rapport de vraisemblances a été proposée par Meng et Rubin [66]. La problématique de la sélection d'un modèle selon des critères d'adéquation sera abordée dans le chapitre 2.

- ***Règles de publication des résultats de l'imputation multiple***

En relation avec le nombre croissant d'études dans lesquelles une méthode d'imputation multiple est appliquée, des règles de publication se dégagent de la littérature [26;52] et sont synthétisées ci-après en suivant la structure de l'article.

Matériel et méthodes

- Présenter les variables incomplètes, la proportion de données manquantes par variable et leur motif global de répartition. Donner un ordre de grandeur de la perte d'effectifs attendue en analyse cas-complet.

- Proposer un mécanisme de données manquantes par variable à partir des variables indicatrices de données manquantes. Si des analyses étiologiques sont prévues, préciser le risque de biais attendu en analyse cas-complet, c'est-à-dire si un mécanisme MAR(ME) global est attendu.

- Identifier selon le type de données recueillies et le mode de collecte les principaux mécanismes générant des données manquantes. Formuler des hypothèses MCAR, MAR, MNAR en fonction de ces informations.

- Dégager l'intérêt de l'estimation des données manquantes par imputation multiple par rapport à l'analyse cas-complet. Préciser la méthode retenue (modèle multivarié normal ou modèle par équations chaînées), ainsi que le logiciel utilisé.

- Détailler le processus d'élaboration du modèle d'imputation. Préciser les variables incluses dans le modèle, leur type ainsi que les transformations éventuelles ainsi que les termes d'interaction retenus. Spécifier le nombre de bases de données imputées en le rapportant à la proportion de données manquantes.

Résultats

- Proposer un diagnostic de l'imputation en comparant les données imputées et observées pour les variables contenant une proportion non-négligeable de données manquantes.

- Présenter si possible les résultats des deux analyses, cas-complet et imputation multiple. Préciser, si c'est le cas, les variables qui ne sont pas communes aux modèles finaux retenus au terme des deux analyses. Présenter les estimations obtenues sur la base d'un modèle incluant les même variables pour les bases de données incomplète et imputée.

Discussion

- Discuter les différences entre les deux analyses en termes de (i) sélection des variables (c'est-à-dire les variables retenues dans le modèle final), (ii) comparaison des estimations tenant compte de la variabilité des résultats à l'aide du coefficient de variation $\left(CV = \beta / SE \right)$, et (iii) biais attendus pour les deux analyses selon les hypothèses sur le mécanisme de données manquantes (MAR(ME) en cas-complet, MNAR en imputation multiple).

- Discuter la validité de l'hypothèse MAR selon (i) la richesse du modèle d'imputation en variables auxiliaires (rendant l'hypothèse MAR plus plausible), (ii) les hypothèses épidémiologiques sur le mécanisme de données manquantes proposé pour les variables "sensibles".

- Discuter l'intérêt d'une analyse de sensibilité permettant de tester la robustesse des résultats au non-respect de l'hypothèse MAR, sachant que ce type d'analyse reste très majoritairement l'apanage d'articles à visée statistique.

6. Mécanismes de données manquantes et biais : Etude de simulation

6.1. Contexte

Il est établi qu'une analyse cas-complet peut donner des résultats biaisés lorsque le mécanisme de données manquantes dépend conjointement de la variable à expliquer et d'une autre variable. En revanche, si le mécanisme de données manquantes ne dépend pas de la variable à expliquer, une analyse cas-complet peut s'avérer plus robuste, c'est-à-dire donner des résultats non-biaisés, en cas de données de type MNAR [4;12;67].

Les méthodes d'imputation multiple, telles qu'elles sont implémentées dans les logiciels standards, sont basées sur l'hypothèse que les données sont manquantes selon un mécanisme MAR, que le mécanisme de données manquantes dépende ou non de la variable à expliquer. Ainsi, une analyse par imputation multiple permet de redresser les biais observés en analyse cas-complet lorsque le mécanisme de données manquantes dépend de la variable à expliquer.

Cependant, si les données sont MNAR, c'est-à-dire que le mécanisme de données manquantes dépend de valeurs non-observées d'une variable, alors l'imputation multiple peut donner des résultats biaisés. Selon Carpenter et al., ce biais peut être plus important si le mécanisme de données manquantes dépend de la variable à expliquer [68]. Ce dernier concept n'est cependant pas explicité clairement dans la littérature, et l'article de Wang et al. [69] tend à montrer que la relation entre le type de mécanisme de données manquantes et les biais observés en imputation multiple est probablement plus complexe.

6.2. Etude de simulation

6.2.1. Matériel et méthodes

En reprenant l'exemple proposé par Vach et al. [4] et illustré par Chavance et al. [70], nous avons choisi de simuler une enquête cas-témoins, avec un témoin par cas. Comme présenté dans le Tableau 1.4, la base de données contient 792 individus, et trois variables binaires M (maladie), E (exposition) et C (confusion).

Tableau 1.4 – Répartition des effectifs de l'enquête cas-complet simulée et odds ratios associés, en global et selon les deux strates du facteur de confusion

	M=1	M=0	Total
E=1	180	162	342
E=0	216	234	450
Total	396	396	792

OR=1.204

IC95% = [0.908;1.595]

C=1	M=1	M=0	Total
E=1	90	135	225
E=0	36	117	153
Total	126	252	378

OR=2.167

IC95% = [1.349;3.429]

C=0	M=1	M=0	Total
E=1	90	27	117
E=0	180	117	297
Total	270	144	414

OR=2.167

IC95% = [1.329;3.533]

Les effectifs ont été calculés de façon à respecter certaines contraintes : obtenir pour l'odds ratio stratifié une valeur plausible d'un point de vue épidémiologique (environ 2) et similaire pour les deux strates du facteur de confusion.

Des données manquantes sont ensuite simulées pour le facteur de confusion selon les mécanismes présentés en paragraphe 2.3.3. Puis des analyses cas-complet et par imputation multiple sont réalisées.

6.2.2. Résultats

- ***Analyse cas-complet***

Une proportion variable de données manquantes est appliquée pour chaque strate selon le mécanisme recherché. Par exemple, pour un mécanisme MCAR, une proportion de 11% de données manquantes est appliquée à chaque cellule des tableaux de contingence des deux strates du facteur de confusion. Pour simuler un mécanisme MAR(M) dépendant de la maladie, une proportion de 22% de données manquantes est appliquée pour les malades et 33% pour les non-malades. Il nous a paru cohérent de simuler une proportion plus élevée de données manquantes parmi les non-malades que parmi les malades, ainsi que parmi les non-exposés que parmi les exposés. Ce principe est décliné pour tous les autres mécanismes. Par exemple, pour le mécanisme MAR(ME), 2% de données manquantes sont simulées pour les non-malades/non-exposés, 4% pour les malades/non-exposés, 3% pour les non-malades/exposés et enfin 50% chez les non-malades/non-exposés.

Le Tableau 1.5 montre, pour chaque mécanisme simulé, la répartition des effectifs dans les tableaux de contingence, en global et pour les deux strates du facteur de confusion, et donne les odds ratios, brut et pour chaque strate du facteur de confusion.

Tableau 1.5 – Effectifs et odds-ratios ajustés selon les différents mécanismes de données manquantes

MCAR : 11% dm

	M=1	M=0	Total
E=1	160	144	304
E=0	192	208	400
Total	352	352	704

OR=1.204

IC95% = [0.908;1.595]

C=1	M=1	M=0	Total
E=1	80	120	200
E=0	32	104	136
Total	112	224	336

OR=2.167

IC95% = [1.332;3.526]

C=0	M=1	M=0	Total
E=1	80	24	104
E=0	160	104	264
Total	240	128	368

OR=2.167

IC95% = [1.290;3.640]

Maladie - MAR(M) : %M1 = 22% - %M0 = 33%

	M=1	M=0	Total
E=1	140	108	249
E=0	168	156	324
Total	308	264	572

OR=1.204

IC95% = [0.908;1.595]

C=1	M=1	M=0	Total
E=1	70	90	160
E=0	28	78	106
Total	98	168	266

OR=2.167

IC95% = [1.204;3.898]

C=0	M=1	M=0	Total
E=1	70	18	88
E=0	140	78	218
Total	210	96	306

OR=2.167

IC95%=[1.290;3.640]

Exposition - MAR(E) : %E1 = 22% - %E0 = 33%

	M=1	M=0	Total
E=1	160	144	304
E=0	144	156	300
Total	304	300	604

OR=1.204

IC95% = [0.908;1.595]

C=1	M=1	M=0	Total
E=1	80	120	200
E=0	24	78	102
Total	104	198	302

OR=2.167

IC95% = [1.265;3.710]

C=0	M=1	M=0	Total
E=1	80	24	104
E=0	120	78	198
Total	200	102	302

OR=2.167

IC95% = [1.265;3.710]

Maladie - Exposition

MAR(ME) : %E1M1 = 2% - %E1M0 = 4% - %E0M1 = 3% - %E0M0 = 50%

	M=1	M=0	Total
E=1	176	156	332
E=0	210	118	328
Total	386	274	660

OR=0.634

IC95% = [0.464;0.866]

C=1	M=1	M=0	Total
E=1	88	130	218
E=0	35	59	94
Total	123	189	312

OR=1.141

IC95% = [0.693;1.878]

C=0	M=1	M=0	Total
E=1	88	26	114
E=0	175	59	234
Total	263	85	348

OR=1.141

IC95% = [0.673;1.934]

Confusion - MNAR(C) : %C1 = 11% - %C0 = 33%

	M=1	M=0	Total
E=1	140	138	278
E=0	152	182	334
Total	292	320	612

OR=1.215

IC95% = [0.883;1.671]

C=1	M=1	M=0	Total
E=1	80	120	200
E=0	32	104	136
Total	112	224	336

OR=2.167

IC95% = [1.332;3.526]

C=0	M=1	M=0	Total
E=1	60	18	78
E=0	120	78	198
Total	180	96	276

OR=2.167

IC95% = [1.190;3.944]

Maladie - Confusion

MNAR(MC) : %C1M1 = 6% - %C1M0 = 11% - %C0M1 = 22% - %C0M0 = 33%

	M=1	M=0	Total
E=1	155	138	293
E=0	174	182	356
Total	329	320	649

OR=1.175

IC95% = [0.862;1.601]

C=1	M=1	M=0	Total
E=1	85	120	205
E=0	34	104	138
Total	119	224	343

OR=2.167

IC95% = [1.345;3.490]

C=0	M=1	M=0	Total
E=1	70	18	88
E=0	140	78	218
Total	210	96	306

OR=2.167

IC95% = [1.204;3.898]

Exposition - Confusion

MNAR(EC) : %C1E1 = 2% - %C1E0 = 11% - %C0E1 = 22% - %C0E0 = 44%

	M=1	M=0	Total
E=1	158	153	311
E=0	132	169	301
Total	290	322	612

OR=1.175

IC95% = [0.962;1.818]

C=1	M=1	M=0	Total
E=1	88	132	220
E=0	32	104	136
Total	120	236	356

OR=2.167

IC95% = [1.342;3.499]

C=0	M=1	M=0	Total
E=1	70	21	91
E=0	100	65	165
Total	170	86	256

OR=2.167

IC95% = [1.214;3.866]

Maladie - Exposition - Confusion

MNAR(MEC) : %C1E1M1 = 2% - %C1E1M0 = 4% - %C1E0M1 = 3% - %C1E0M0 = 50%
%C0E1M1 = 11% - %C0E1M0 = 26% - %C0E0M1 = 17% - %C0E0M0 = 43%

	M=1	M=0	Total
E=1	168	150	318
E=0	185	126	311
Total	353	276	629

OR=0.763

IC95% = [0.447;1.079]

C=1	M=1	M=0	Total
E=1	88	130	218
E=0	35	29	94
Total	123	189	312

OR=1.141

IC95% = [0.643;1.639]

C=0	M=1	M=0	Total
E=1	80	20	100
E=0	150	67	217
Total	230	87	317

OR=1.787

IC95% = [1.218;2.355]

Les résultats du Tableau 1.5 montrent que les estimations des OR stratifiés sont biaisées pour le mécanisme MAR(ME) avec un odds ratio de 1.141 [0.373 ;1.934] pour C=0. L'intervalle de confiance de l'odds ratio pour ce mécanisme ne contient donc pas la vraie valeur de l'OR qui est égale à 2.167 [1.349 ;3.429]. Pour le mécanisme MEC, on observe un biais de la même amplitude, même si l'intervalle de confiance de l'OR stratifié contient la vraie valeur de l'OR pour C=0.

- ***Analyse après imputation multiple***

Pour chacune des bases de données incomplètes créées, on estime les données manquantes par imputation multiple et 50 bases sont générées. Puis, une analyse bivariée est effectuée sur l'ensemble des bases imputées. Les résultats de ces analyses réalisées en cas-complet et après imputation multiple sont synthétisés dans le Tableau 1.6.

Tableau 1.6 – Résultats des analyses bivariées en analyse cas-complet et après imputation multiple

Mécanisme de données manquantes		Cas Complet	Imputation Multiple
MCAR		2.167	2.162
		[1.33;3.53]	[1.54;3.06]
MAR(M)	Maladie	2.167	2.185
		[1.20;3.90]	[1.53;3.11]
MAR(E)	Exposition	2.167	2.171
		[1.26;3.71]	[1.53;3.08]
MAR(ME)	Maladie + Exposition	**1.141**	2.157
		[0.69;1.88]	[1.54;3.02]
MNAR(C)	Confusion	2.167	**2.132**
		[1.33;3.53]	**[1.51;3.02]**
MNAR(EC)	Exposition + Confusion	2.167	**1.964**
		[1.34;3.50]	**[1.40;2.75]**
MNAR(MC)	Maladie + Confusion	2.167	**2.219**
		[1.34;3.49]	**[1.57;3.14]**
MNAR(MEC)	Maladie + Exposition + Confusion	**1.395**	**2.184**
		[0.64;1.64]	**[1.55;3.07]**

Sont présentés les odds ratios associés à la variable d'exposition après ajustement sur le facteur de confusion. Les zones en vert représentent les OR biaisés en analyse cas-complet, et les zones en jaune les OR biaisés en analyse par imputation multiple.

6.2.3. Discussion

Les résultats montrent que l'analyse cas-complet est effectivement biaisée pour les mécanismes impliquant la variable à expliquer et l'une des variables d'exposition, c'est-à-dire les mécanismes MAR(ME) et MNAR(MEC). L'imputation multiple donne des résultats non-

biaisés lorsque le mécanisme de données manquantes est de type MCAR ou MAR, et permet donc de redresser le biais observé en analyse cas-complet pour le mécanisme MAR(ME). Pour les mécanismes MNAR, les résultats apparaissent effectivement biaisés après imputation multiple.

Cependant, selon les connaissances sur ce sujet, des biais plus importants sont attendus pour les mécanismes MNAR dépendant de la variable à expliquer, c'est-à-dire les mécanismes MNAR(MC) et MNAR(MEC). Or, c'est pour le mécanisme MAR(EC) que le biais est le plus important. Cela montre tout d'abord les limites d'une analyse simple de simulation. Il aurait en effet été intéressant de faire varier les effectifs de départ, ainsi que les proportions de données manquantes selon les mécanismes, mais les contraintes imposées ne permettaient pas d'envisager d'autres scénarios. Par ailleurs, et cela sera confirmé dans une application plus complexe sur des données réelles (chapitre 3), l'amplitude de la dépendance entre le mécanisme de données manquantes et la variable à expliquer peut varier selon les variables, avec un impact différentiel sur les biais attendus en cas de mécanisme MNAR.

Conclusion

Nous avons abordé dans ce chapitre le traitement des données manquantes dans des bases de données transversales ou de systèmes de surveillance. Nous n'avons pas présenté le mode de traitement de données longitudinales, car l'imputation est alors basée sur des modèles différents de ceux décrits, tenant compte de la répétition des observations pour les mêmes individus. Ainsi, la prise en compte de données manquantes de type monotone et selon un processus MAR liées à des sorties d'études dans le cadre de données de cohorte est plus facilement gérable en pratique par une estimation par maximum de vraisemblance des paramètres d'un modèle mixte plutôt que par imputation multiple. Par ailleurs, certaines données transversales ont une structure hiérarchique du fait de la présence de sous-groupes définis au sein de l'échantillon analysé. Il est alors possible d'imputer ce type de données avec les méthodes d'imputation présentées ici, en incluant des variables indicatrices identifiant les sous-groupes. Cette adaptation a cependant ses limites si le nombre de sous-groupes est élevé, et il est alors préférable d'appliquer des procédures dédiées [31].

CHAPITRE 2

PROCESSUS D'IMPUTATION MULTIPLE DEDIE A DES ANALYSES SPECIFIQUES : APPLICATION A DES DONNEES D'ENQUETES ET DE SURVEILLANCE

1. Etude du risque de transmission du VIH par les dons de sang

En France, les hommes ayant des rapports sexuels avec des hommes (HSH) sont exclus de façon permanente du don de sang parce qu'ils ont un risque plus élevé d'être infectés par le VIH [71;72], mais aussi par l'hépatite B ou la syphilis, et donc un risque de transmettre ces infections. Depuis 1985, des progrès très importants ont été réalisés en matière de tests de dépistage pour le VIH, incluant depuis 2001 le Dépistage Génomique Viral (DGV) qui permet la détection d'infections très récentes.

Cependant, un risque de transmission du VIH persiste, essentiellement dû à l'existence d'une fenêtre silencieuse, c'est-à-dire d'une période qui survient après que le donneur ait été infecté et avant que les marqueurs de l'infection soient détectables. Du fait de l'existence de cette fenêtre silencieuse, les autorités de santé ont maintenu l'exclusion permanente des HSH des dons de sang. Cette mesure, considérée comme discriminatoire par certains activistes, n'est pas totalement respectée car certains HSH ne déclarent pas leur comportement sexuel avant le don.

L'objectif de cette étude était d'estimer la part de risque résiduel de transmission du VIH par transfusion attribuable aux HSH.

1.1. Méthodes

1.1.1. Recueil de données

Depuis 1992 en France, tous les centres de transfusion collectent chaque trimestre le nombre total de dons et de donneurs selon le statut du donneur (nouveau ou connu) et les caractéristiques épidémiologiques (sexe, âge, sous-type du VIH 1, origine géographique) des donneurs confirmés VIH positifs au système national de surveillance des donneurs de sang [73]. Parmi les donneurs identifiés comme positifs pour le VIH, plus des trois quarts reviennent pour une consultation médicale après le don durant laquelle l'information sur le mode probable de contamination est recueillie.

1.2.2. Examen des données manquantes

La base de données constituée à partir de la surveillance de l'infection par le VIH parmi les donneurs de sang pour la période 1994-2008 contient 1105 observations et 47 variables. Les variables sélectionnées pour les analyses sont l'année de diagnostic (1994-2008), le sexe, l'âge (variable continue), le département de domicile, le pays de naissance (France métropolitaine, Antilles/Guyane, Afrique sub-Saharienne, autre), le mode de contamination (homosexuel, hétérosexuel, autres), le type de donneur (nouveau donneur, donneur connu) et la présence d'anticorps VIH (variable binaire).

La variable mode de contamination est incomplète avec 29.7% de données manquantes. On génère une indicatrice de données manquantes égale à 1 si le mode de contamination est manquant et 0 sinon. L'examen univarié du mécanisme de données manquantes de la variable mode de contamination montre des liens significatifs (test du Chi2, $p \leq 0.05$) entre l'indicatrice de données manquantes et les variables année de diagnostic, sexe, département de domicile, pays de naissance et type de donneur. Toutes ces variables sont complètes et le mécanisme de données manquantes est donc a priori MAR.

Cependant, le mode probable de contamination est recueilli lors d'un entretien après le don de sang pour les personnes positives pour le VIH. Puisque les donneurs HSH sont exclus de façon permanente du don de sang, certains d'entre eux pourraient considérer cette mesure comme discriminatoire et être tentés de cacher leur orientation sexuelle afin de pouvoir donner leur sang. Par ailleurs, le don de sang pourrait être un moyen simple pour les HSH d'accéder à un dépistage du VIH réalisé au moyen de tests performants. Ainsi, un mécanisme

de données manquantes de type MNAR ne peut être exclu pour la variable mode de contamination.

1.2. Construction du modèle d'imputation

1.2.1. Construction des équations de prédiction

La variable mode de contamination est catégorielle et une fonction de lien multinomiale est spécifiée. Les variables retenues pour les analyses de risque résiduel sont l'année de diagnostic, le sexe, le type de donneur et la présence d'anticorps VIH. Ces 4 variables constituent les variables principales, systématiquement incluses dans le modèle d'imputation. Peu de variables supplémentaires ont pu être retenues dans la base de données de départ, et 3 variables auxiliaires ont été identifiées : l'âge, le département de domicile et le pays de naissance.

1.2.2. Choix du nombre de bases et de cycles

En tenant compte des indications de la littérature récente, nous avons sélectionné $M = 30$ bases puisque la variable incomplète contient 30% de données manquantes. Une seule variable est incomplète et donc un seul cycle de l'échantillonneur de Gibbs est nécessaire.

1.3. Diagnostic de l'imputation

Afin d'estimer si le modèle d'imputation est correct et si l'hypothèse MAR est plausible, nous avons comparé les proportions observées et estimées de la variable mode de contamination. En effet, si le processus d'imputation est valide, on s'attend à ce que les variations entre valeurs observées et imputées restent dans des limites raisonnables et qu'elles soient explicables par un mécanisme identifié.

Les résultats de cet examen sont présentés dans le Tableau 2.1. Les proportions observées de la variable mode de contamination ne différent pas significativement des proportions estimées, puisqu'elles appartiennent aux intervalles de confiance à 95% (IC 95%) des proportions estimées, aussi bien pour chacune des périodes de 3 ans que sur l'ensemble de la période d'étude.

Tableau 2.1 – Comparaison des proportions observées et estimées de la variable mode de contamination selon la période

Mode de transmission	1994-1996	1997-1999	2000-2002	2003-2005	2006-2008	1994-2008
Homosexuel			**Proportions (%)**			
Observé	29.30	24.59	32.58	26.25	34.15	31.4
Imputé	32.96	26.25	34.02	28.38	34.90	33.72
IC 95%	26.2,39.8	18.6,33.9	24.5,43.5	18.6,38.2	24.7,45.1	30.5,37.0
Hérérosexuel						
Observé	65.61	72.13	65.17	72.5	64.63	64.35
Imputé	61.50	70.39	63.55	70.13	63.85	61.43
IC 95%	54.5,68.5	62.4,78.3	54.0,73.1	60.2,80.1	53.6,74.0	58.1,64.8
Autres modes de transmission						
Observé	5.1	3.28	2.25	1.25	1.22	4.25
Imputé	5.54	3.36	2.43	1.48	1.25	4.86
IC 95%	2.1,9.0	0.2,6.6	0.8,5.7	0.9,5.7	0.8,3.6	3.2,6.5

Nous en concluons que le risque d'avoir des données MNAR est limité, ou que le modèle est suffisamment riche en variables prédictrices pour rendre l'hypothèse MAR plausible.

1.4. Calcul du risque résiduel

1.4.1. Risque résiduel VIH pour la période 2006-2008

Le risque résiduel de transmettre le VIH par transfusion a été calculé par périodes de 3 ans en faisant le produit de l'incidence du VIH chez les donneurs de sang ayant donné au moins deux fois au cours de la période d'étude par la durée de la fenêtre silencieuse (en années) [74]. L'incidence du VIH correspond au nombre de donneurs réguliers ayant séroconverti pour le VIH pendant la période de 3 ans (S) divisé par le nombre de donneurs-années (DA) calculé en sommant les délais entre le premier et le dernier don de chaque donneur au cours de la période d'étude. La fenêtre silencieuse pour le VIH correspond à la période pendant laquelle le virus est indétectable, et a été estimée à 12 jours [75].

1.4.2. Evaluation de la part du risque résiduel attribuable aux HSH

Malgré la mesure d'exclusion permanente, des HSH sont régulièrement testés positifs pour le VIH à l'occasion d'un don de sang. Afin d'estimer l'impact des dons de ces donneurs HSH sur le risque résiduel de contamination par le VIH, l'incidence du VIH chez les donneurs de sang a été calculée en deux parties : l'incidence estimée chez les HSH et l'incidence estimée chez les autres donneurs (femmes et hommes non HSH).

Si l'on note I_{HSH} et I_{AUTRES} l'incidence du VIH chez les donneurs HSH et les autres donneurs, et DA_{HSH} et DA_{AUTRES} le nombre de Donneurs-Années chez les donneurs HSH et les autres donneurs, alors l'incidence du VIH pour l'ensemble des donneurs est estimée comme suit :

$$I = \left[\frac{DA_{HSH}}{\left(DA_{HSH} + DA_{AUTRES}\right)} \times I_{HSH} \right] + \left[\frac{DA_{AUTRES}}{\left(DA_{HSH} + DA_{AUTRES}\right)} \times I_{AUTRES} \right]$$

avec $I_{HSH} = \frac{S_{HSH}}{DA_{HSH}} \times 10^5$ et $I_{AUTRES} = \frac{S_{AUTRES}}{DA_{AUTRES}} \times 10^5$ où S_{HSH} et S_{AUTRES} sont respectivement le nombre de séroconversions VIH observées chez les donneurs HSH et chez les autres donneurs pendant la période d'étude.

Cependant, contrairement aux nombres de séroconversions, les nombres de Donneurs-Années n'ont pu être obtenus séparément pour les HSH et les autres donneurs à partir du système de surveillance des donneurs de sang.

Les deux quantités DA_{HSH} et DA_{AUTRES} ont pu être estimées à partir des données suivantes :

(1) Pendant la période 2006-2008, **50.2%** des donneurs étaient des hommes

(2) Un intervalle a été proposé pour estimer la proportion d'HSH parmi les donneurs masculins, déduit des résultats d'une étude en population générale sur le comportement sexuel des Français (CSF-2006) [76]. L'intervalle proposé est [1.5% ; 4.1%] et les bornes sont utilisées pour proposer des valeurs des incidences et des risques résiduels pour les donneurs de sang HSH et les autres donneurs.

Ainsi on aurait, selon le scénario, une proportion de **1.5%** ou de **4.1%** de HSH parmi les donneurs de sang masculin, ce qui donne :

$$I_{HSH-1} = \frac{S_{HSH}}{DA \times 0.502 \times 0.015} \times 10^5 \text{ et } I_{AUTRES-1} = \frac{S_{AUTRES}}{DA \times (1 - (0.502 \times 0.015))} \times 10^5$$

$$I_{HSH-2} = \frac{S_{HSH}}{DA \times 0.502 \times 0.041} \times 10^5 \text{ et } I_{AUTRES-1} = \frac{S_{AUTRES}}{DA \times (1 - (0.502 \times 0.041))} \times 10^5$$

où DA est le nombre total de Donneurs-Années.

1.5. Résultats

1.5.1. Risque résiduel VIH pour la période 2006-2008

Les résultats sont présentés dans le Tableau 2.2. Pour la période 2006-2008, 31 séroconversions VIH ont été observées parmi les donneurs de sang ayant fait au moins deux dons, ce qui donne une incidence de 1.3 pour 100 000 Donneurs-Années. A partir de cette incidence, le risque résiduel a été estimé à 0.41 par million de dons soit 1 sur 2 440 000 de dons.

Tableau 2.2 – Risque résiduel de transmission de l'infection par le VIH par transfusion selon la période fenêtre, 2006-2008

Nombre de cas VIH incidents	Nombre de DA	Incidence pour 100 000 DA (IC 95%)	Durée de la période fenêtre en jours (étendue)	Risque résiduel par million de dons (IC 95%)
31	2 467 560	1.26 (0.87-1.81)	12 (0 – 28)	0.41 (0.0 – 1.39)

1.5.2. Part de risque résiduel attribuable aux HSH

Parmi les 31 cas incidents observés entre 2006 et 2008, 23 étaient des hommes et 8 des femmes. Après imputation multiple de la variable mode de contamination, l'analyse jointe des 30 bases permet d'estimer que, parmi les 23 hommes, 15 se sont contaminés par voie homosexuelle et 8 par voie hétérosexuelle. Le nombre de cas incidents non-HSH a donc été estimé à 16. Les estimations d'incidence et de risque résiduel de transmission du VIH par transfusion sont présentées dans le Tableau 2.3 pour les deux scénarios proposés.

Tableau 2.3 – Incidence du VIH et risque résiduel de transmission de l'infection par le VIH par transfusion parmi les donneurs HSH et les autres donneurs

	Donneurs HSH	**Autres donneurs**
Nombre de cas VIH incidents (séroconversions)	$S_{HSH} = 15$	$S_{AUTRES} = 16$
Nombre de Donneurs-Années		
Scénario 1: 1.5% donneurs masculins HSH	18 590	2 448 970
Scénario 2: 4.1% donneurs masculins HSH	50 813	2 416 747
Incidence du VIH pour 100 000 Donneurs-Années		
Scénario 1: 1.5% donneurs masculins HSH	$I_{HSH-1} = 80.7$	$I_{AUTRES-1} = 0.65$
Scénario 2: 4.1% donneurs masculins HSH	$I_{HSH-2} = 29.5$	$I_{AUTRES-2} = 0.66$
Risque résiduel pour 1 million de dons		
Scénario 1: 1.5% donneurs masculins HSH	$RR_{HSH-1} = 26.5$	$RR_{AUTRES-1} = 0.21$
Scénario 2: 4.1% donneurs masculins HSH	$RR_{HSH-2} = 9.7$	$RR_{AUTRES-2} = 0.22$

Pour le scénario 1 (1,5% des donneurs masculins seraient des HSH), le risque résiduel d'infection par le VIH lié aux HSH a été estimé à 26.5 par million de dons ou 1 sur 38 000 dons, soit 125 fois plus que le risque imputable aux autres donneurs (1 sur 4 700 000 dons).

Pour le scénario 2 (4,5% des donneurs hommes seraient des HSH), le risque résiduel d'infection par le VIH lié aux HSH a été estimé à 9,7 par million de dons ou 1 sur 100 000 dons, soit 45 fois plus que le risque imputable aux autres donneurs (1 sur 4 600 000 dons).

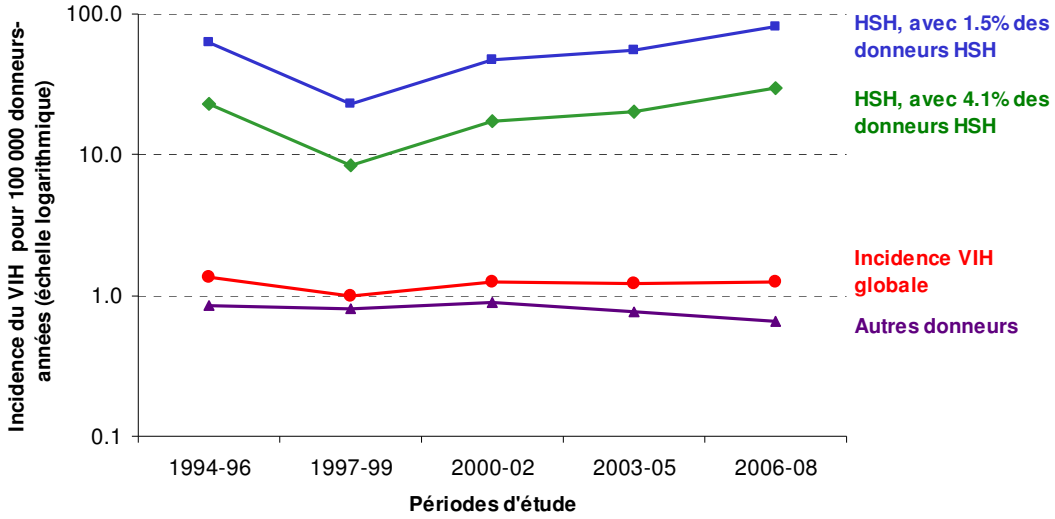
Ces résultats montrent que, si tous les HSH avaient renoncé à donner leur sang au cours de la période 2006-2008, le risque résiduel de transmission du VIH se serait situé dans l'intervalle [1 sur 4 700 000 – 1 sur 4 600 000 dons] soit la moitié du risque actuel (1 sur 2 440 000 dons).

Notons que l'incidence, et donc le risque résiduel, sont plus élevés pour le scénario 1 (1.5% des donneurs masculins sont des HSH) que pour le scénario 2 (4.1% des donneurs masculins sont des HSH) car, pour un nombre de séroconversions identique, le nombre de Donneurs-Années est plus faible pour le scénario 1 que pour le scénario 2.

1.5.3. Evolution de l'incidence du VIH sur l'ensemble de la période de surveillance

En appliquant les proportions estimées par imputation multiple aux modalités de la variable mode de contamination, l'incidence a été estimée pour chaque période de 3 ans pour les donneurs HSH selon les deux scénarios ainsi que pour les autres donneurs.

Figure 2.1 – Tendances d'incidence du VIH parmi les donneurs HSH et les autres donneurs pour la période 1994-2008



Les résultats illustrés par la Figure 2.1 montrent que l'incidence a augmenté de façon significative parmi les donneurs HSH entre la période 1994-1996 et la période 2006-2008 (p=0.02) alors que l'incidence globale est restée stable.

1.6. Discussion

L'estimation du nombre d'hommes se contaminant par voie homosexuelle était précédemment réalisée par réaffectation proportionnelle. Sachant que les données manquantes pour cette variable sont à fort risque d'être MNAR, il convient de choisir une méthode d'estimation prenant en compte toute l'information disponible, afin de rendre l'hypothèse MAR plausible. Le modèle d'imputation inclut 7 variables prédictrices et les résultats de l'étape de diagnostic montrent que les données imputées et observées sont proches. La variable mode de contamination contient 30% de données manquantes et la stabilité des résultats permet de conclure que l'imputation est valide.

Les résultats de cette étude montrent que la part du risque de contamination par le VIH lors de transfusion attribuable aux HSH est importante, puisque ce risque serait réduit de moitié si les HSH s'étaient abstenus de donner leur sang. La politique actuelle d'exclusion du don de sang peut être ressentie comme discriminatoire pour certains HSH, ce qui en amène certains à donner leur sang tout en dissimulant leur orientation sexuelle. Cependant, l'incidence du VIH chez les donneurs de sang HSH (0.08%) reste très inférieure à celle estimée chez les HSH de la population générale (1%) [71].

Pour compléter cette étude, un modèle a été construit afin d'évaluer l'impact d'une modification de cette politique d'exclusion actuelle en la remplaçant par une mesure consistant à n'exclure que les HSH multipartenaires au cours des 12 mois précédant le don. Ce modèle est basé sur des données obtenues à partir d'enquêtes comportementales et épidémiologiques. Les résultats montrent qu'une modification de la politique d'exclusion des HSH, au profit d'une politique plus proche de celle pratiquée pour les autres donneurs, peut augmenter le risque de transmission du VIH par transfusion [77]. Toutefois, le modèle ne prend pas en compte l'amélioration possible de la compliance des HSH avec une mesure d'exclusion moins stricte et qui serait donc perçue comme plus équitable. A l'inverse, l'assouplissement de cette mesure pourrait encourager certains HSH à se faire dépister dans des centres de transfusion sanguine. Des études qualitatives doivent donc être mises en œuvre pour évaluer les changements possibles de compliance des donneurs de sang liés à une nouvelle stratégie.

Conclusion

Cette illustration à partir de données de surveillance du VIH chez les donneurs de sang présente une application simple de la méthode d'imputation multiple. En effet, une seule variable est manquante et le schéma de données manquantes est donc univarié, cas de figure rarement rencontré en pratique. Cependant, on doit faire l'hypothèse que les données sont MNAR pour cette variable, puisqu'il est établi à partir des entretiens postérieurs au don de sang que certains HSH ont volontairement dissimulé leur préférence sexuelle. Parmi les 35% d'individus positifs pour le VIH qui ne se présentent pas à l'entretien, certains sont probablement en partie des HSH avec un comportement à risque qui recherchent un test VIH rapide et fiable. Des données récentes de l'InVS (non publiées) indiquent d'ailleurs que les cas d'infection dépistés par DGV à un stade très précoce sont très majoritairement des homosexuels.

2. Enquête cas-témoins sur l'infection à campylobacter

Dans la plupart des pays industrialisés, les infections à Campylobacter représentent, avec les infections par les Salmonelles, la cause la plus fréquente de gastroentérite bactérienne d'origine alimentaire. Des études antérieures ont montré que les principaux facteurs de risque d'infection à Campylobacter actuellement identifiés sont : la consommation de poulet peu cuit ainsi que d'autres produits carnés, la consommation d'eau non-traitée, les contacts avec des animaux de compagnie ou de ferme ainsi que le fait de voyager à l'étranger [78-80].

L'objectif de cette étude était d'identifier les facteurs de risque des cas sporadiques d'infection à Campylobacter survenant en France. Cette étude consistait en une enquête cas-témoins incidents exploratoire nationale avec appariement, et elle a été menée du 15/08/02 au 30/06/04.

2.1. Population d'étude et critères d'inclusion

La population cible était composée de personnes de tout âge résidant en France métropolitaine.

La population source était composée des personnes résidant dans la zone géographique du praticien prescripteur de la coproculture et du laboratoire d'analyse et de biologie médicale qui avait déclaré le cas.

Un cas était défini comme toute personne résidant en France métropolitaine pour lequel un diagnostic clinique de gastroentérite ou d'infection systémique avait été posé puis confirmé par l'isolement d'un Campylobacter dans les selles ou un liquide biologique normalement stérile.

Les cas étaient sélectionnés à partir du système de surveillance des infections à Campylobacter [81]. Chaque cas était apparié selon l'âge, le sexe (pour les plus de 15 ans) et le lieu de résidence, avec un témoin désigné par le médecin traitant du cas.

Au final, 269 cas et témoins ont été retenus pour les analyses.

2.2. Recueil de données

2.2.1. Données collectées

La période d'exposition explorée correspondait aux 8 jours précédant la date d'apparition des premiers symptômes du cas. Le questionnaire a porté sur des expositions alimentaires (types d'aliments, mode de cuisson et/ou lieu de consommation) et sur des contacts avec des animaux (vivants ou morts) ainsi qu'avec des malades diarrhéiques dans l'entourage du cas. L'hygiène en cuisine a été explorée comme une habitude comportementale et non sur une période donnée.

A partir de la base de données d'origine, nous avons constitué une base de données restreinte pour les analyses (cas-complet et imputation multiple). Pour cela, nous avons d'abord sélectionné et/ou reconstruit 28 variables binaires. Des variables ayant un pourcentage élevé de données manquantes (30%) ont été retenues, mais les variables ayant de faibles effectifs dans la catégorie des exposés ont été exclues. La base de données contient au final 21 variables d'exposition.

2.2.2. Examen des données manquantes

- ***Examen quantitatif***

Le questionnaire contenait 200 questions portant sur la consommation détaillée de nombreuses catégories d'aliments pendant une période de 8 jours. Malgré la qualité du recueil des informations, le mode de recueil rétrospectif a induit un problème de mémorisation et les réponses "ne sait pas" ont généré des données manquantes pour la majorité des variables d'exposition. Sur les 21 variables d'exposition, 3 variables sont complètes, 10 variables ont moins de 8% de données manquantes, 5 variables de 8 à 15% et 3 variables de 15 à 30% (Tableau 2.4).

Afin de déterminer la typologie des données manquantes, on examine la répartition des données manquantes parmi les différentes combinaisons des 21 variables incomplètes. Même si des variables appartenant à un même sous-groupe (par exemple les variables de consommation de poulet) ont des données manquantes communes, la répartition paraît suivre un motif arbitraire, c'est-à-dire que la proportion de données manquantes est du même ordre de grandeur pour la plupart des combinaisons des 21 variables.

De ce fait, les 21 variables sont entièrement renseignées pour 202 individus seulement, sur un effectif total de 538.

- ***Examen qualitatif***

Afin d'identifier le mécanisme de données manquantes, on associe à chaque variable d'exposition incomplète E_i une indicatrice de données manquantes R_i binaire, qui vaut 1 si la variable E_i est manquante pour l'individu *i* et 0 sinon. On croise chaque indicatrice R_i avec d'une part la variable à expliquer M (infection à Campylobacter), et d'autre part chacune des variables d'exposition E_i. Le lien statistique recherché est un test du Chi2 significatif ($p \leq 0.05$). Cet examen univarié permet de proposer une première synthèse des mécanismes de données manquantes.

Le mécanisme de données manquantes dépend de la variable à expliquer pour une seule variable, avoir mangé au restaurant. Cela signifie que, pour les autres variables d'exposition, la proportion de données manquantes ne diffère pas significativement entre les cas et les témoins. Le détail des relations entre les variables indicatrices de réponse et les variables d'exposition est donné dans le Tableau 2.4. Il montre que 18 des 21 variables d'exposition sont liées significativement à au moins une variable indicatrice de données manquantes.

Le mécanisme de données manquantes serait donc a priori de type MAR(ME) pour la variable avoir mangé au restaurant, MAR(E) pour 17 autres variables, et MCAR pour les 3 variables de contact (avec des malades diarrhéiques ou des animaux). On peut en déduire que le risque de biais des estimateurs en analyse cas-complet est réduit. Notons qu'un mécanisme MCAR pour les 3 variables de contact est peu réaliste, même si aucune relation statistique n'a été mise en évidence. Par ailleurs, un mécanisme de type MNAR doit être envisagé pour les 18 autres variables, puisqu'un lien existe entre leurs indicatrices de données manquantes et des variables d'exposition incomplètes, donc potentiellement avec les valeurs non-observées de ces variables.

Tableau 2.4 – Examen de la base de données incomplète

Variables d'exposition		Données manquantes (%)		Variables de non réponse *	
Libellé	Signification	Témoins (N=269)	Cas (N=269)	Libellé	Lien avec les variables d'exposition †
Consommation de volaille					
E1	Avoir mangé du poulet	32.7	27.1	R1	R2 R8 R16
E2	Avoir mangé du poulet acheté au détail	23.8	28.6	R2	R1
Consommation de bœuf					
E3	Avoir mangé du bœuf	11.1	13.4	R3	R1 R2 R6 R7
E4	Avoir mangé du bœuf hors du domicile	4.5	2.2	R4	R2 R6 R7
E5	Avoir mangé du bœuf acheté en boucherie,à la ferme ou au marché	0.0	0.4	R5	R3 R6 R7
E6	Avoir mangé du bœuf acheté au détail	15.6	18.6	R6	R2 R18
E7	Avoir mangé du bœuf peu cuit	7.8	11.1	R7	R3
Autres consommations alimentaires					
E8	Avoir mangé au restaurant	6.0	1.1	R8	R1 R2 R6 R10 R16 R17 R18
E9	Avoir mangé de la viande cuite au barbecue	0.0	0.0	R9	R16 R17 R18
E10	Avoir mangé du poisson ou des fruits de mer	11.9	9.7	R10	R1 R6
E11	Avoir mangé des légumes crus ou des salades	7.4	4.5	R11	R1 R3 R6
E12	Avoir mangé des fruits ou des baies	7.4	11.5	R12	R8
E13	Avoir consommé des produits laitiers	3.0	3.0	R13	R11
E14	Avoir mangé du fromage	0.0	0.0	R14	R1 R6 R8 R21
E15	Avoir bu de l'eau du robinet	0.7	1.1	R15	R6
Comportement en cuisine pour la préparation des repas					
E16	Hygiène insuffisante des mains	4.1	6.3	R16	R21
E17	Hygiène insuffisante des ustensiles	3.3	5.6	R17	R8 R18
E18	Hygiène insuffisante des plans de travail	5.2	7.1	R18	R21
Contacts avec des animaux ou des personnes malades					
E19	Avoir eu des contacts avec des animaux de compagnie ou des animaux de ferme	6.0	3.0	R19	_
E20	Expositions professionnelles (animaux morts ou vivants)	0.0	0.0	R20	_
E21	Avoir été en contact avec une personne diarrhéique	3.3	5.2	R21	_

* Variables indicatrices de données manquantes indiquant si la variable est entièrement renseignée ou non
† Variables indicatrices de données manquantes liées significativement à chacune des variables d'exposition (Chi2, $p \leq 0,05$)

Il est possible d'évaluer la pertinence de l'hypothèse MAR. D'un point de vue épidémiologique, les mécanismes de données manquantes explorés concernent des variables de consommation alimentaire, d'hygiène et de contact avec des animaux ou des personnes. Parmi ces expositions, un mécanisme de type MAR parait plausible pour les variables de consommation alimentaire puisque les données manquantes proviennent sans doute d'un défaut de mémorisation. Le mécanisme de données manquantes dépendrait alors seulement des valeurs observées des autres variables (un consommateur régulier de produits carnés pourrait consommer moins souvent du poisson).

En revanche, les expositions liées aux comportements d'hygiène en cuisine pourraient être ressenties comme sensibles et il parait cohérent que certaines personnes, plus particulièrement parmi celles qui ont été malades, préfèrent ne pas fournir ce type d'information. Le mécanisme de données manquantes dépendrait alors de la valeur non-observée des variables, et les données seraient MNAR. Cependant, la proportion de données manquantes pour ces variables est faible (5%) et n'est pas plus élevée chez les cas que chez les témoins.

2.3. Construction et validation du modèle d'imputation

2.3.1. Analyse cas-complet

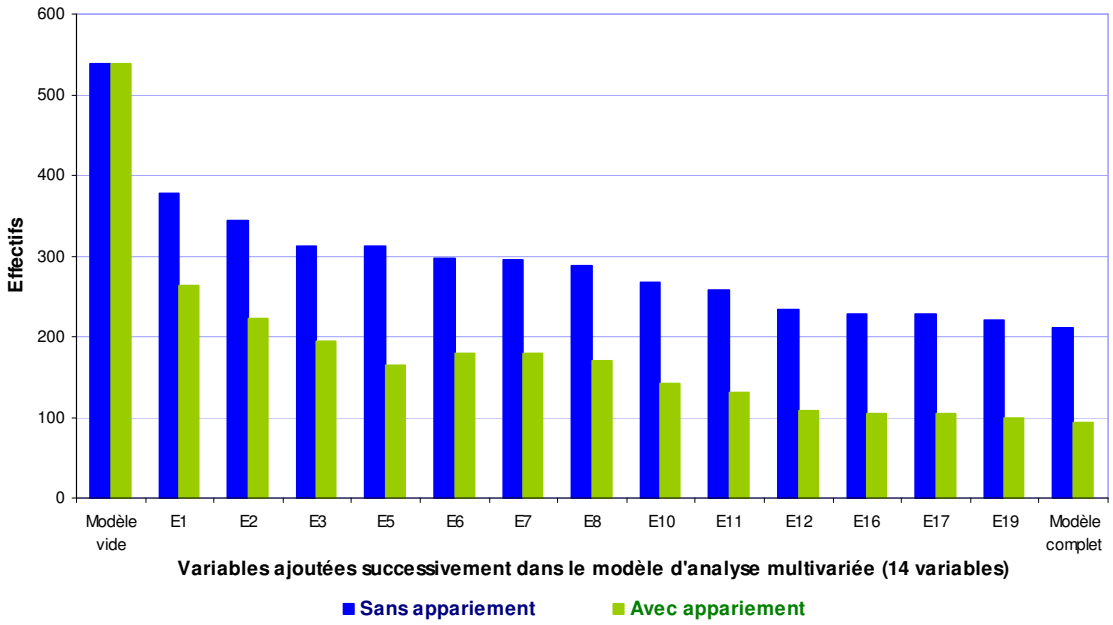
Une analyse cas-complet est réalisée afin (i) d'identifier les variables retenues pour être incluses dans le modèle d'analyse multivariée en cas-complet et après imputation multiple, et (ii) d'obtenir des estimations qui pourront être comparées à celles obtenues après imputation multiple. Au terme de l'analyse univariée, 14 variables ont été retenues pour l'analyse multivariée car elles sont liées à la variable à expliquer avec un seuil de significativité $p \leq 0.2$ (Tableau 2.5). Des interactions d'ordre 1 ont été recherchées mais aucune n'était significative.

Tableau 2.5 – Sélection des variables incluses dans les modèles d'analyse et d'imputation

Variables d'exposition		Odds Ratio apparié	
Libellé	Signification	(IC 95%)	p
Variables d'exposition liées à la variable à expliquer et à la non réponse : variables principales			
E1	Avoir mangé du poulet	0,7 (0,4;1,2)	0.2
E2	Avoir mangé du poulet acheté au détail	0,5 (0,3;1,0)	0.05
E3	Avoir mangé du bœuf	0,8 (0,5;1,2)	0.2
E5	Avoir mangé du bœuf acheté en boucherie,à la ferme ou au marché	0,6 (0,4;1,0)	0.03
E6	Avoir mangé du bœuf acheté au détail	0,6 (0,4;0,9)	0.02
E7	Avoir mangé du bœuf peu cuit	2,0 (1,2;3,4)	0.009
E8	Avoir mangé au restaurant	1,6 (1,0;2,7)	0.06
E10	Avoir mangé du poisson ou des fruits de mer	0,5 (0,3;0,9)	0.01
E11	Avoir mangé des légumes crus ou des salades	0,4 (0,2;0,7)	0.002
E12	Avoir mangé des fruits ou des baies	0,5 (0,4;0,8)	0.004
E16	Hygiène insuffisante des mains	1,5 (1,0;2,2)	0.04
E17	Hygiène insuffisante des ustensiles	1,7 (1,1;2,6)	0.009
E19	Avoir eu des contacts avec des animaux de compagnie ou des animaux de ferme	1,5 (1,0;2,3)	0.06
E21	Avoir été en contact avec une personne diarrhéique	2,3 (1,3;3,9)	0.003
Variables d'exposition non liées à la variable à expliquer mais liées à la non réponse : variables auxiliaires			
E4	Avoir mangé du bœuf hors du domicile	0,9 (0,4;2,1)	0.8
E14	Avoir mangé du fromage	0,8 (0,5;1,2)	0.3
E9	Avoir mangé de la viande cuite au barbecue (toute viande)	1,5 (0,8;2,3)	0.3
E13	Avoir consommé des produits laitiers	0,8 (0,4;1,5)	0.4
E15	Avoir bu de l'eau du robinet	1,0 (0,7;1,4)	0.9
E18	Hygiène insuffisante des plans de travail	1,3 (0,8;2,0)	0.3
Variable d'exposition ni liée à la variable à expliquer ni liée à la non réponse : variable accessoire			
E20	Expositions professionnelles (animaux morts ou vivants)	1,3 (0,8;2,0)	0.3

Nous avons réalisé une analyse multivariée en appliquant une stratégie de sélection des variables pas à pas descendante. Cependant, les 14 variables incluses dans le modèle sont incomplètes, et la perte d'effectifs en analyse multivariée est importante et aggravée par l'appariement. En effet, comme illustré sur la Figure 2.2, la présence d'une donnée manquante pour un cas ou un témoin entraine la perte de la paire complète lors de l'analyse.

Figure 2.2 – Evolution des effectifs en fonction des variables incluses successivement dans le modèle d'analyse multivariée cas-complet



Ainsi, après inclusion des 14 variables dans le modèle multivarié tenant compte de l'appariement, les effectifs sont ramenés à 94 individus sur les 538 initiaux. De ce fait, l'algorithme de maximisation de la vraisemblance du modèle ne converge pas et les paramètres ne peuvent pas être estimés. Nous avons donc appliqué une stratégie d'analyse par sous-modèles qui consiste à effectuer une première sélection des variables en les incluant par groupes de variables corrélées dans 4 modèles multivariés indépendants. Les 9 variables retenues au terme des 4 analyses ($p \leq 0.2$) sont incluses dans un modèle unique et le modèle final obtenu porte sur 340 individus.

2.3.2. Construction du modèle d'imputation

- ***Construction des équations de prédiction***

La spécification des équations de prédiction consiste à préciser une fonction de lien et un ensemble de variables prédictrices pour chaque variable à imputer. Les variables de la base de données sont toutes binaires et une fonction de lien logit est spécifiée. La sélection des variables à inclure dans les équations de prédiction est réalisée comme suit (Tableau 2.5).

La règle de base est d'inclure systématiquement toutes les variables significativement liées à la variable à expliquer en analyse univariée (test du Chi2, $p \leq 0.2$). Elles constituent les variables dites principales et sont incluses dans le modèle d'analyse multivariée. La variable à expliquer doit également être retenue pour l'imputation. Le modèle d'imputation le plus réduit se limite à ces 15 variables.

Le modèle d'imputation gagne à être plus riche que le modèle d'analyse par l'inclusion de variables dites auxiliaires, non liées à la variable à expliquer mais liées aux variables indicatrices de données manquantes des variables principales (test du Chi2, $p \leq 0.05$). Six variables auxiliaires potentielles sont identifiées.

L'inclusion de variables qui ne sont liées ni à la variable à expliquer, ni aux variables indicatrices de données manquantes, peut réduire la précision des estimations. Elles sont dites variables accessoires. Dans cette étude, une seule variable est de type accessoire.

Le modèle d'imputation retenu se compose donc de 21 variables.

- ***Nombre de bases et de cycles***

En se basant sur les critères d'efficacité statistique, il est recommandé d'imputer 5 bases pour une fraction d'information manquante (FMI) d'environ 25%, tolérant ainsi une perte d'efficacité statistique de moins de 5% par rapport à un nombre infini d'imputations [3]. Afin de limiter la perte de puissance lors des analyses, il faudrait imputer 10 bases pour avoir une perte de puissance de moins de 5% par rapport à 100 imputations [31]. Nous avons fait le choix conservateur d'imputer 30 bases de données en raison de la taille réduite de la base de départ ainsi que de la proportion et de la répartition des données manquantes.

Le programme additionnel utilisé (ICE) spécifie 10 cycles par défaut car la convergence de l'échantillonneur de Gibbs est rapide si le modèle est correctement spécifié, et nous avons retenu cette valeur.

2.3.3. Analyse et diagnostic des données imputées

- ***Analyse***

Les 30 bases de données imputées sont analysées de façon séparée puis combinée selon les règles de Rubin. Le modèle d'analyse multivariée inclut la variable à expliquer ainsi que les 14 variables identifiées lors de l'analyse univariée cas-complet. Les commandes spécifiques de l'imputation permettent d'effectuer une analyse par régression logistique tenant compte de l'appariement (régression logistique conditionnelle).

Notons que les variables associées à une diminution du risque de Campylobactériose ont été conservées dans le modèle d'analyse (avoir mangé du poisson ou des fruits de mer, avoir mangé des légumes crus ou des salades et avoir mangé des fruits ou des baies). En effet, on peut faire l'hypothèse que l'effet protecteur de ces variables est probablement indirect, lié à une préférence alimentaire et non à un mécanisme biologique.

- ***Diagnostic***

Cette étape, qui en pratique est réalisée en parallèle de la phase d'analyse, a pour objectif de valider le modèle d'imputation, et d'estimer si l'hypothèse que les données soient MAR est plausible.

- **Comparaison des données observées et imputées**

Même si la procédure d'imputation multiple ne peut être validée à partir des données observées, la comparaison des données observées et des données imputées est informative. Ainsi, il est logique d'observer des différences modérées entre données observées et imputées, puisque l'examen des mécanismes de données manquantes montre de nombreuses relations entre les indicatrices de données manquantes et les variables. Nous avons choisi de comparer les OR obtenus en analyse univariée cas-complet et après imputation. Le Tableau 2.6 présente ces résultats, ainsi qu'un critère de variation relative entre ces OR.

Tableau 2.6 – Résultats comparés de l'analyse univariée en analyse cas-complet et après imputation multiple

Variables d'exposition		Cas Complet		Imputation Multiple		Diagnostic (%)	
		OR_{CC} (IC 95%)	SE	OR_{IM} (IC 95%)	SE	$\frac{\lvert OR_{IM}-OR_{CC}\rvert^*}{OR_{CC}}$	Données manquantes (%)
E1	Avoir mangé du poulet	0.69 (0.42-1.16)	0.18	0.91 (0.60-1.38)	0.19	31.9	30.0
E2	Avoir mangé du poulet acheté au détail	0.55 (0.30-0.99)	0.17	0.73 (0.47-1.13)	0.16	32.7	26.2
E3	Avoir mangé du bœuf	0.80 (0.54-1.19)	0.16	0.85 (0.60-1.21)	0.15	6.3	12.3
E5	Avoir mangé du bœuf acheté en boucherie, à la ferme ou au marché	0.64 (0.43-0.97)	0.13	0.69 (0.46-1.03)	0.14	7.8	3.3
E6	Avoir mangé du bœuf acheté au détail	0.62 (0.41-0.94)	0.13	0.78 (0.54-1.12)	0.14	25.8	17.1
E7	Avoir mangé du bœuf peu cuit	2.00 (1.18-3.38)	0.53	2.26 (1.39-3.69)	0.56	13.0	9.5
E8	Avoir mangé au restaurant	1.64 (1.00-2.70)	0.42	1.67 (1.02-2.72)	0.41	1.8	3.5
E10	Avoir mangé du poisson ou des fruits de mer	0.55 (0.34-0.88)	0.13	0.55 (0.36-0.86)	0.12	0.0	10.8
E11	Avoir mangé des légumes crus ou des salades	0.42 (0.25-0.72)	0.12	0.45 (027-0.74)	0.12	7.1	6.0
E12	Avoir mangé des fruits ou des baies	0.55 (0.37-0.83)	0.11	0.64 (0.44-0.94)	0.12	16.4	9.5
E16	Hygiène insuffisante des mains	1.50 (1.01-2.22)	0.3	1.51 (1.03-2.21)	0.29	0.7	5.2
E17	Hygiène insuffisante des ustensiles	1.71 (1.15-2.55)	0.35	1.68 (1.14-2.49)	0.34	1.8	4.5
E19	Avoir eu des contacts avec des animaux de compagnie ou de ferme	1.52 (0.98-2.35)	0.34	1.46 (0.95-2.23)	0.32	3.9	4.5
E21	Avoir été en contact avec une personne diarrhéique	2.26 (1.32-3.88)	0.62	2.03 (1.22-3.39)	0.53	5.8	4.3

* Variation relative des OR obtenus en analyse cas-complet (OR_{CC}) et après imputation multiple (OR_{IM})

Des variations relatives de 20 à 30% sont observées pour les 3 variables ayant une proportion de données manquantes élevée, de 17 à 30%. Pour ces variables, les OR_{CC} et les OR_{IM} ne diffèrent pas significativement puisque les OR_{CC} appartiennent aux intervalles de confiance à 95% des OR_{IM}. Pour les autres variables, le critère de variation relative des OR est inférieur à 10% pour 10 variables sur 12.

En ce qui concerne les variables d'hygiène pour lesquelles les données sont à risque d'être MNAR, la variation relative est inférieure à 2%. Cela peut être lié au faible pourcentage de données manquantes pour ces variables (<5%), ainsi que par l'inclusion des principaux prédicteurs de ces variables dans le modèle d'imputation.

- **Choix du nombre de bases**

Lors de la réalisation de cette étude, les critères diagnostiques tels que l'efficacité statistique et l'erreur de Monte Carlo n'étaient pas directement accessibles lors des analyses. De ce fait, nous avons tenu compte, lors du choix du nombre de bases imputées, des variations de la valeur de l'odds ratio ajusté (ORa) de 4 variables clés en analyse multivariée en fonction du nombre de bases imputées.

Tableau 2.7 – Evolution des ORa en analyse multivariée en fonction du nombre de bases imputées

Variables	Données manquantes (%)	Nombre de bases imputées				
		5	10	20	30	50
E1	29.9	1.05 (0.70 - 1.58)	0.96 (0.64 - 1.45)	0.97 (0.65 - 1.45)	0.99 (0.66 - 1.49)	0.99 (0.66 - 1.49)
E6	17.1	0.96 (0.64 - 1.43)	0.90 (0.61 - 1.32)	0.91 (0.62 - 1.34)	0.91 (0.61 - 1.34)	0.90 (0.61 - 1.32)
E10	16.8	0.64 (0.40 - 1.06)	0.66 (0.41 - 1.08)	0.67 (0.41 - 1.10)	0.67 (0.41 - 1.09)	0.67 (0.41 - 1.09)
E17	4.5	1.80 (1.15 - 2.82)	1.83 (1.19 - 2.82)	1.82 (1.18 - 2.81)	1.82 (1.18 - 2.81)	1.82 (1.18 - 2.82)

E1 : Avoir mangé du poulet ; E6 : Avoir mangé du bœuf acheté au détail ; E10 : Avoir mangé du poisson ou des fruits de mer ; E17 : Hygiène insuffisante des ustensiles

Les résultats présentés dans le Tableau 2.7 montrent peu de variations au-delà de 20 bases, aussi bien pour la valeur de l'ORa que pour l'intervalle de confiance à 95%. On peut en déduire qu'il est suffisant d'imputer 30 bases.

2.4. Résultats

Les résultats des analyses cas-complet et par imputation multiple sont présentés dans le Tableau 2.8. Pour le modèle d'analyse final, les effectifs sont respectivement de 340 et 538 individus.

Tableau 2.8 – Analyse multivariée des facteurs associés à une augmentation ou une diminution du risque d'infection à Campylobacter, analyse cas-complet et imputation multiple

Variables	Cas complet† (N* = 340)			Imputation Multiple (N* = 538)		
	OR	IC 95%	p	OR	IC 95%	p
E5 - Avoir mangé du bœuf acheté en boucherie, à la ferme ou au marché	0.51	0.28 - 0.93	0.03	0.59	0.37 - 0.94	0.03
E7 - Avoir mangé du bœuf peu cuit	**2.71**	1.37 - 5.39	0.004	**2.76**	1.62 - 4.73	< 0.001
E8 - Avoir mangé au restaurant			NS‡	**1.75**	1.02 - 3.03	0.04
E11 - Avoir mangé des légumes crus ou des salades	0.45	0.23 - 0.88	0.002	0.40	0.22 - 0.70	0.002
E17 - Hygiène insuffisante des ustensiles			NS‡	**2.10**	1.32 - 3.30	0.002
E21 - Avoir été en contact avec une personne diarrhéique	**3.19**	1.57 - 6.48	0.001	**2.01**	1.13 - 3.58	0.02

Sont notés en caractères gras les ORa des variables associées à une augmentation du risque d'infection à Campylobacter.
* Nombre d'individus pris en compte dans le calcul des ORa dans le modèle final
† Analyse cas-complet réalisée par l'intermédiaire de sous-modèles
‡ Odds ratio non significatif

Pour les deux analyses, la consommation de bœuf insuffisamment cuit et le contact avec une personne diarrhéique sont des facteurs indépendamment associés au risque de survenue d'une Campylobactériose. Par ailleurs, la consommation de bœuf acheté en boucherie, à la ferme ou au marché, ainsi que la consommation de légumes crus ou de salades sont des facteurs associés à une diminution du risque de Campylobactériose.

L'analyse réalisée par imputation multiple met en évidence deux facteurs de risque supplémentaires de Campylobactériose : le fait d'avoir mangé au restaurant et une hygiène insuffisante des ustensiles de cuisine lors de la préparation des repas.

Les odds ratios ajustés (ORa) présentés dans le Tableau 2.8 sont issus de modèles multivariés différents en analyse cas-complet et en imputation multiple. Nous avons restreint l'analyse

par imputation multiple aux 4 variables sélectionnées par l'analyse cas-complet afin de pouvoir comparer les estimations en fonction du type d'analyse. Les résultats sont présentés dans le Tableau 2.9.

Tableau 2.9 – Résultats comparés de l'analyse multivariée (modèle final à 4 variables), analyse cas-complet et imputation multiple

Variables	Cas Complet (N* = 330)			Imputation Multiple (N* = 538)		
	OR	SE	(SE/OR)x100†	OR	SE	(SE/OR)x100†
E5 - Avoir mangé du bœuf acheté en boucherie, à la ferme ou au marché	0.51 (0.28-0.93)	0.15	29.4	0.61 (0.39-0.97)	0.14	23.1
E7 - Avoir mangé du bœuf peu cuit	2.71 (1.37-5.38)	0.95	35.1	2.62 (1.57-4.38)	0.69	26.3
E11 - Avoir mangé des légumes crus ou des salades	0.45 (0.23-0.88)	0.15	33.3	0.47 (0.27-0.80)	0.13	27.7
E21 - Avoir été en contact avec une personne diarrhéique	3.19 (1.57-6.48)	1.15	36.1	2.03 (1.17-3.53)	0.57	28.1

* Nombre d'individus pris en compte dans le calcul des OR dans le modèle final
† Coefficient de variation associé à l'OR, exprimé en %

Les ORa sont proches pour les deux analyses pour les variables E5, E7 et E11. On observe un gain de précision après imputation multiple avec une baisse du coefficient de variation. Pour la variable contact avec une personne diarrhéique, la différence entre les ORa obtenus dans les deux analyses est plus marquée avec une variation relative de 36%. Notons que, pour cette variable, l'écart type est divisé par 2 après imputation et le coefficient de variation est réduit de presque 10%.

2.5. Analyse de sensibilité selon le modèle d'imputation

Il est recommandé de tester plusieurs bases de données imputées issues de modèles d'imputation différents, car cela peut constituer une forme simple d'analyse de sensibilité [35]. Dans cette étude, le type des variables ne peut être modifié, mais il est possible de faire varier le nombre de variables incluses dans le modèle d'imputation.

Le Tableau 2.10 présente les valeurs des ORa et de leur écart-type pour les 6 variables retenues au terme de l'analyse multivariée après imputation, et pour des modèles incluant de 15 à 24 variables d'exposition.

Les modèles 1 à 4 sont composés des variables principales et d'un nombre croissant de variables auxiliaires, le modèle 4 étant le modèle retenu pour l'imputation. On note peu de variations de la valeur de chacun des ORa et de son écart-type selon les modèles.

Le modèle 5 inclut en plus une variable accessoire, c'est-à-dire sans capacité prédictrice, et le modèle 6 contient les 3 variables d'appariement (sexe, âge, région de résidence). Notons qu'en toute rigueur il aurait été justifié de tester les capacités prédictrices des 3 variables d'appariement afin de déterminer si elles pouvaient être des variables auxiliaires. On observe un impact faible de l'ajout de ces variables, avec au maximum une variation relative par rapport aux résultats du modèle 4 de 6% pour la variable E8 (avoir mangé au restaurant).

Les modèles 8 à 11 correspondent au modèle 1, c'est-à-dire le modèle minimal, dont on a exclu certaines variables principales. On se trouve donc dans des conditions où le modèle d'imputation spécifié est incorrect, ce qui en théorie peut provoquer des biais dans les estimations. Or l'impact sur les estimations est limité, avec une variation relative maximale de 8% par rapport au modèle 4 pour la variable E8.

Cette analyse montre que les estimations sont stables selon le modèle d'imputation retenu, même lors d'incompatibilité des modèles d'imputation et d'analyse. Cela peut être lié à la robustesse de la méthode d'imputation, ainsi qu'à la richesse du modèle d'imputation en variables prédictrices bien corrélées entre elles par sous-groupes.

Tableau 2.10 – Résultats comparés du modèle d'analyse multivariée final selon le modèle d'imputation

Modèles d'imputation	ORa (Ecart type)					
	E5	E7	E8	E11	E17	E21
Modèle 1 : 15 variables (variable à expliquer + 14 variables principales)	0.60 (0.14)	2.79 (0.76)	1.69 (0.47)	0.41 (0.12)	2.05 (0.47)	2.02 (0.60)
Modèle 2 : 18 variables (variable à expliquer + 14 variables principales + 3 variables auxiliaires)	0.59 (0.14)	2.74 (0.76)	1.74 (0.49)	0.39 (0.11)	2.10 (0.49)	2.04 (0.60)
Modèle 3 : 19 variables (variable à expliquer + 14 variables principales + 4 variables auxiliaires)	0.59 (0.14)	2.84 (0.78)	1.73 (0.48)	0.39 (0.11)	2.11 (0.49)	2.03 (0.60)
Modèle 4 : 21 variables (variable à expliquer + 14 variables principales + 6 variables auxiliaires)	0.58 (0.14)	2.77 (0.76)	1.75 (0.49)	0.40 (0.12)	2.10 (0.49)	2.01 (0.59)
Modèle 5 : 22 variables (variable à expliquer + 14 variables principales + 6 variables auxiliaires + 1 variable accessoire)	0.60 (0.14)	2.75 (0.76)	1.75 (0.49)	0.39 (0.11)	2.10 (0.49)	1.98 (0.58)
Modèle 6 : 24 variables (variable à expliquer + 14 variables principales + 6 variables auxiliaires + 3 variables d'appariement)	0.60 (0.14)	2.71 (0.75)	1.64 (0.45)	0.44 (0.13)	2.07 (0.48)	1.96 (0.57)
Modèle 8 : 14 variables (Modèle 1 sans la variable E1)	0.61 (0.14)	2.85 (0.77)	1.73 (0.48)	0.40 (0.11)	2.10 (0.47)	1.99 (0.58)
Modèle 9 : 14 variables (Modèle 1 sans la variable E3)	0.60 (0.14)	2.78 (0.80)	1.63 (0.45)	0.40 (0.11)	2.00 (0.47)	1.93 (0,57)
Modèle 10 : 14 variables (Modèle 1 sans la variable E10)	0.59 (0.14)	2.86 (0.83)	1.61 (0.44)	0.39 (0.11)	2.07 (0.48)	1.95 (0.57)
Modèle 11 : 13 variables (Modèle 1 sans les variables E1 et E3)	0.60 (0.14)	2.69 (0.73)	1.66 (0.46)	0.41 (0.12)	2.09 (0.48)	2.03 (0.60)

E5 : Avoir mangé du bœuf acheté en boucherie, à la ferme ou au marché ; E7 : Avoir mangé du bœuf peu cuit ; E8 : Avoir mangé au restaurant ; E11 : Avoir mangé des légumes crus ou des salades ; E17 : Hygiène insuffisante des ustensiles ; E21 : Avoir été en contact avec une personne diarrhéique.
E1 : Avoir mangé du poulet ; E3 : Avoir mangé du bœuf ; E10 : Avoir mangé du poisson ou des fruits de mer.
En gris foncé : modèle retenu pour l'imputation ; en gris clair : modèle incluant toutes les variables disponibles.

2.6. Discussion

2.6.1. Interprétation des résultats

- ***Processus de sélection des variables***

L'analyse cas-complet, bien que potentiellement non-biaisée, n'est pas satisfaisante puisque les variables sélectionnées en analyse univariée n'ont pu être incluses simultanément dans le modèle d'analyse multivariée. Le processus de sélection des variables en analyse multivariée est donc faussé puisque toutes les relations entre les variables n'ont pu être explorées. Un éventuel effet de confusion pourrait donc être omis en relation avec les données manquantes. De plus l'analyse multivariée cas-complet ne porte que sur 340 individus pour le modèle final sur les 538 individus de départ.

De ce fait, l'analyse par imputation multiple est plus efficace car elle permet de tenir compte de l'ensemble des informations contenues dans la base de données. La sélection des variables n'est alors pas faussée par les données manquantes et les variables retenues au terme de l'analyse multivariée diffèrent de celles de l'analyse cas-complet. Deux facteurs de risque supplémentaires de Campylobactériose sont mis en évidence après imputation multiple, le fait d'avoir mangé au restaurant et une hygiène en cuisine insuffisante.

Par ailleurs, certaines variables qui auraient pu être occultées lors de l'analyse cas-complet par manque de puissance, car contenant une proportion élevée de données manquantes (25 à 30%), ne sont pas retenues au terme de l'analyse après imputation multiple. Ainsi, l'imputation multiple permet de confirmer que la consommation de poulet, identifiée par d'autres études comme un facteur de risque de Campylobactériose, n'est pas mise en évidence dans cette enquête. Un des objectifs de cette étude était de préciser ce point important.

- ***Puissance et précision***

Outre une amélioration du processus de sélection des variables, l'imputation multiple permet d'obtenir un gain de puissance et donc de précision, ainsi que de redresser les biais potentiels en analyse cas-complet, sous l'hypothèse MAR.

L'analyse cas-complet est réalisée sur 38.6% des effectifs totaux pour les 4 variables communes aux deux analyses. Les coefficients de variation rendent compte de l'impact de

cette perte d'effectifs en analyse cas-complet. Ils ont des valeurs proches pour les 4 variables considérées. Notons que la perte d'effectifs est alors indépendante de la proportion de données manquantes initiale de ces variables, mais dépend de l'effet conjugué des données manquantes des 4 variables. Lors de l'analyse après imputation multiple, portant sur des effectifs restaurés, la variabilité liée à la présence de données estimées est prise en compte dans le calcul de la variance (en intégrant une variance inter-bases). L'imputation d'un nombre suffisant de bases de données, tenant compte de la proportion de données manquantes dans la base de données de départ, doit permettre de limiter cette variabilité. Dans cette étude, cet objectif est atteint puisque le coefficient de variation est nettement réduit après imputation multiple pour les 4 variables considérées. Par ailleurs, un examen simple des ORa de variables clés selon le nombre de bases imputées montre que le choix d'imputer 30 bases est valide. Une limite de cette étude est de n'avoir pas précisé cette notion par le calcul pour chacune des variables de la FMI, de l'efficacité statistique et de l'erreur de Monte-Carlo.

- ***Biais – hypothèse MAR***

Lors de la comparaison des résultats obtenus par analyse cas-complet et par imputation multiple, la question d'un biais potentiel des estimations doit être abordée. Notons cependant que, lors de l'analyse d'une base de données incomplète, le biais lié aux données manquantes n'est quantifiable que lorsque les données complètes sont disponibles, soit uniquement par des études de simulation. L'examen initial de la base de données incomplète montre que les mécanismes de données manquantes des variables d'intérêt sont très majoritairement de type MAR(E) ou MAR(EC) si l'on tient compte d'un effet de confusion. Sous ces scénarios, une analyse de type cas-complet est non-biaisée.

L'imputation multiple peut donner des résultats biaisés (i) si le modèle d'imputation n'est pas correctement spécifié, ou (ii) si des données sont manquantes selon un mécanisme MNAR. Dans cette étude, la base de données contient des variables qui ont pu être ajoutées au modèle d'imputation afin d'améliorer ses capacités prédictrices et 6 variables auxiliaires ont été sélectionnées. Une limite est de ne pas avoir recherché les capacités prédictrices des variables d'appariement, même si l'inclusion de la variable à expliquer dans le modèle d'imputation permet de tenir compte du statut cas/témoin. De plus, les estimations obtenues avec ce modèle plus complet (modèle 6 à 24 variables) sont proches de celles présentées (modèle 4 à 21 variables). Par ailleurs, nous avons testé l'effet de l'exclusion d'une ou plusieurs variables

principales du modèle d'imputation et les estimations se sont révélées robustes à cette incompatibilité entre les modèles d'imputation et d'analyse.

Dans cette étude, nous avons fait l'hypothèse que les données manquantes étaient essentiellement dues à un défaut de mémorisation, en relation avec le type d'information collectée sur une période précise et avec un mode de recueil rétrospectif. Cependant, les réponses aux questions d'hygiène pourraient être volontairement omises, menant à un mécanisme MNAR. Lorsque l'hypothèse MAR est posée à tort, les estimations obtenues après imputation peuvent être biaisées, c'est-à-dire les estimations des coefficients de régression des variables du modèle final. Cependant, la proportion de données manquantes des variables d'hygiène est de l'ordre de 5%, et ne dépend pas du statut cas/témoin.

Les estimations obtenues avec les deux types d'analyse sont proches pour les 4 variables communes. La différence relative observée entre les ORa associés à la variable cuisson du bœuf et plus nettement pour la variable contact avec une personne diarrhéique peut être liée au gain de précision après imputation multiple, objectivé par la baisse du coefficient de variation. On peut donc conclure que le processus d'imputation sous l'hypothèse MAR, basé sur un modèle riche en variables prédictrices, permet de redresser les biais liés à un mécanisme MNAR potentiel.

2.6.2. Intérêt méthodologique de l'étude

La principale difficulté rencontrée a consisté à effectuer la sélection des variables à inclure dans le modèle d'imputation. Au moment de l'étude, le mode de sélection des variables était encore peu documenté dans la littérature, essentiellement détaillé par Van Buuren et Schafer [17;42]. Nous avons exploré en analyse univariée les liens entre les variables indicatrices de données manquantes des variables incomplètes et toutes les autres covariables retenues. Cela nous a permis de synthétiser les mécanismes de données manquantes, et de sélectionner des variables auxiliaires. Cependant, la notion de variable prédictrice est imprécise puisque le nombre de liens entre les variables auxiliaires et les indicatrices de données manquantes des variables auxiliaires varie de 1 à 3 liens significatifs sur 14 variables à imputer. De même, les variables principales ont des capacités prédictrices très variables puisque de 1 à 8 relations significatives sont observées entre ces variables et leurs indicatrices de données manquantes. Nous avons fait le choix conservateur d'inclure chaque variable liée à au moins une

indicatrice de données manquantes dans l'équation de prédiction de toutes les variables à imputer.

Cependant, les liens identifiés ne tiennent pas compte d'un ajustement sur les autres covariables. Une analyse multivariée des indicatrices de données manquantes est recommandée dans la littérature récente, mais n'aurait pas pu être mise en œuvre à partir de la base de données incomplète, en raison de la proportion et de la répartition des données manquantes. Ce type d'analyse multivariée aurait pu être effectuée à partir de bases de données complétées selon un modèle d'imputation très général, afin de valider le choix des variables prédictrices.

Par ailleurs, dans cette étude, la stratégie d'analyse à partir des données imputées est réalisée de façon linéaire, en se basant sur la significativité des coefficients, ce qui permet de sélectionner un modèle final sans faire appel à des procédures de post-estimation. Notons que l'estimation des données manquantes par imputation multiple permet de minimiser le risque de ne pas dépister un effet de confusion lié aux données manquantes, si le processus d'imputation est valide. Dans cette étude, l'imputation multiple par équations chaînées est particulièrement adaptée puisque les variables incomplètes sont discrètes. Par ailleurs, la richesse de la base de données en variables prédictrices très corrélées entre elles permet de conclure que les résultats issus de l'analyse multivariée après imputation sont fiables.

Pour le choix du modèle d'imputation, aucun critère d'adéquation ne peut être calculé. Il est cependant possible de tester l'adéquation aux valeurs observées de chacun des modèles construits à partir des équations de prédiction. Il aurait ainsi fallu élaborer 14 modèles, un par variable à imputer, en testant pour chaque modèle l'effet de l'inclusion de chacune des 6 variables auxiliaires.

Conclusion

Cette application pratique a été réalisée à partir d'une base de données contenant de nombreuses variables incomplètes, avec des données qui sont manquantes selon un schéma arbitraire et un mécanisme majoritairement MAR. L'intérêt de cette étude est donc d'illustrer le processus de sélection des variables lorsque de nombreuses variables sont disponibles. Notons que pour certaines bases de données de santé publique, pouvant contenir plus de 100 variables, des procédures automatiques de construction des équations de prédiction sont disponibles pour certaines implémentations. Par exemple, le programme IVEware retient par

une procédure d'analyse pas à pas descendante automatique les variables considérées comme prédictrices. La sélection est automatisée en spécifiant au préalable la valeur de la part de la variance expliquée du modèle de régression [82]. Cependant, le nombre de variables retenues dans l'équation de prédiction de chaque variable varie sensiblement selon la valeur de ce critère. Ce type de sélection automatique contient donc une part d'arbitraire qu'il convient de contrôler au mieux.

3. Estimation par une méthode capture-recapture à trois sources du nombre de nouveaux diagnostics VIH chez les enfants : imputation d'une variable de stratification

L'objectif méthodologique de ce travail était d'estimer par imputation multiple les données manquantes d'une variable de stratification non-renseignée pour une des trois sources de données, puis d'appliquer la méthode capture-recapture à des bases de données imputées.

3.1. Contexte

Au niveau mondial, l'immense majorité des infections par le VIH chez les enfants est due à une transmission mère-enfant (TME) au cours de la grossesse, de l'accouchement ou de l'allaitement. Environ 430 000 nouvelles infections pédiatriques ont été diagnostiquées dans le monde en 2008 [83]. La plupart de ces infections auraient pu être évitées par des programmes de prévention de la TME.

En France, le risque de transmission du VIH à l'enfant a été réduit de façon marquée depuis la fin des années 80 grâce à l'utilisation prophylactique et thérapeutique des thérapies antirétrovirales (ART). Ces traitements administrés à la mère pendant la grossesse, mais aussi au bébé pendant les premières semaines de vie, ont permis d'obtenir un taux de transmission très faible, de l'ordre de 1% pour les femmes ayant accouché entre 1997 et 2004 [84]. Ainsi, un diagnostic précoce durant la grossesse et une mise sous traitement rapide de la mère permet une prévention efficace de la TME. En France, une nouvelle politique nationale datant de 1993 stipule qu'un test de dépistage de l'infection par le VIH doit être systématiquement proposé aux femmes enceintes lors du premier examen prénatal (1er trimestre de la grossesse).

Il est communément admis que le nombre d'enfants vivant avec le VIH en France est d'environ 1 500 sur un total estimé de 150 000 personnes séropositives. De plus, 10 à 15 nouvelles infections sont diagnostiquées chez des nouveau-nés chaque année [85], auxquelles il faudrait ajouter les diagnostics chez les enfants nés en pays de forte endémie, arrivant en France après leur naissance et pour lesquels aucune estimation n'est disponible.

3.2. Objectif

Notre objectif était d'estimer par la méthode capture-recapture le nombre total de nouveaux diagnostics d'infection à VIH chez les enfants de moins de 13 ans en France métropolitaine pour la période 2003-2006, à partir de trois sources de données : la déclaration obligatoire du VIH (DOVIH), l'Enquête Périnatale Française (EPF) et la surveillance de l'activité de dépistage du VIH auprès des laboratoires (LaboVIH). L'objectif secondaire de cette étude était d'évaluer l'exhaustivité de la DOVIH chez l'enfant et celle d'EPF.

La méthode capture-recapture permet, en croisant au moins deux sources d'informations d'une même maladie, d'estimer le nombre total de cas de la maladie et ainsi l'exhaustivité de chaque source [86].

Définition de cas

Un cas est un enfant de moins de 13 ans pour lequel un nouveau diagnostic d'infection à VIH a été effectué en France métropolitaine entre le 01 janvier 2003 et le 31 décembre 2006. Le diagnostic est basé sur un examen virologique (culture virale ou PCR) chez l'enfant de moins de 18 mois ou sur une sérologie confirmée positive (Western Blot) chez l'enfant de plus de 18 mois [85].

3.3. Description des trois sources de données

3.3.1. La déclaration obligatoire du VIH (DOVIH)

La déclaration obligatoire des nouveaux diagnostics de l'infection à VIH a été mise en place en 2003 par l'Institut de Veille Sanitaire (InVS) afin de suivre les tendances temporelles et spatiales de l'épidémie et de décrire les caractéristiques des nouveaux cas [87]. Chaque cas avec une sérologie positive confirmée pour la première fois doit être notifié par les biologistes. La création d'un code d'identification unique pour chaque individu, construit à partir de la date de naissance, du prénom, de l'initiale du nom et du sexe, permet de détecter les éventuels doublons. Les cliniciens complètent le formulaire de notification avec des informations épidémiologiques et cliniques. Pour les enfants de moins de 13 ans, la notification n'était basée jusqu'en 2007 que sur les pédiatres et non sur la base d'une double déclaration par les biologistes et les cliniciens comme chez les adultes. Les formulaires de notification sont transmis aux autorités sanitaires, puis centralisés à l'InVS où les données

sont saisies dans une base de données nationale. Afin de tenir compte des délais de déclaration, plus longs pour les notifications des enfants que pour celles des adultes, la base DOVIH utilisée a inclus des notifications jusqu'au 31 mars 2010.

3.3.2. L'enquête périnatale française (EPF)

L'enquête périnatale française, financée par l'ANRS (ANRS-EPF CO1/CO10/CO11), est une cohorte de mères séropositives, incluses au plus tard à l'accouchement, ayant donné leur accord pour leur inclusion et le suivi de leurs enfants, qu'ils soient ou non infectés par le VIH. Initiée en 1984, cette enquête collecte prospectivement des données grâce à la participation volontaire d'une centaine de services de maternité et de pédiatrie sur l'ensemble du territoire français [84]. Les critères d'inclusion des enfants ont été étendus depuis 2005 à tous les enfants de moins de 13 ans récemment diagnostiqués pour le VIH dans des sites pédiatriques et nés de mères non incluses dans EPF, avec un consentement parental. Pour ces enfants, le recueil de données a été rétrospectif entre 2003 et 2004 puis prospectif à partir d'octobre 2004. Des examens cliniques et biologiques sont effectués tous les 6 mois entre la naissance et 18 ans pour les enfants infectés par le VIH et jusqu'à 24 mois pour les enfants non-infectés. L'objectif de cette enquête était d'étudier les facteurs associés à la TME du VIH, la tolérance aux antirétroviraux administrés en périnatal chez les mères et leurs enfants, ainsi que le pronostic de l'infection pédiatrique. La base de données constituée pour notre étude inclut des cas sélectionnés à partir du fichier disponible en avril 2008, pour lequel les doublons avaient été identifiés et supprimés.

3.3.3. L'enquête LaboVIH

Afin d'évaluer l'activité de dépistage du VIH en France, l'InVS a mis en place en 2001 l'enquête LaboVIH auprès de l'ensemble des laboratoires d'analyses biologiques et médicales (4200 laboratoires). Chaque laboratoire envoie à l'InVS des données agrégées par semestre concernant le nombre de tests VIH effectués et le nombre de tests confirmés positifs [88]. Il précise également le nombre de diagnostics d'infection par le VIH confirmés chez des enfants âgés de moins de 15 ans.

Une troisième source de données a donc pu être constituée à partir de ce réseau de surveillance afin d'améliorer les conditions d'application de la méthode capture-recapture. Parmi les laboratoires participant chaque année (taux de participation de 85%), 137

laboratoires métropolitains ont signalé au moins un diagnostic d'infection à VIH confirmé pour la période 2003-2006. Un questionnaire leur a été envoyé afin de recueillir des données individuelles pour chacun de ces diagnostics. Parmi ces 137 laboratoires, 113 ont finalement rapportés au moins un cas pédiatrique d'infection au VIH. Seuls les diagnostics pour les enfants de moins de 13 ans ont été retenus pour l'étude. Les doublons de notifications ont été identifiés et supprimés.

3.4. Méthodes

3.4.1. Identification des cas communs

En l'absence d'un identifiant unique pour chaque nouveau diagnostic, des algorithmes d'identification des cas communs aux sources ont été construits en fonction de critères communs disponibles dans chacune des bases de données. L'année de naissance, le sexe, l'hôpital de suivi (ou son département) et la date de diagnostic (ou de prise en charge) étaient renseignés dans les trois sources. L'algorithme d'identification des cas communs à la DOVIH et EPF a également pris en compte la maternité de naissance pour les enfants nés en France ou le pays de naissance, l'origine géographique de la mère et le statut vital de l'enfant.

L'identification des cas communs a été réalisée par la procédure SQL de SAS (version 9.1) et a été complétée par une vérification manuelle.

3.4.2. Identification de variables d'hétérogénéité de capture

Des variables dites d'hétérogénéité de capture permettent de prendre en compte des variations de la probabilité de notification au sein d'une même source [86]. Trois variables ont été identifiées comme variables d'hétérogénéité potentielles.

- ***La région de diagnostic (Ile de France, province)***

Pour EPF, l'exhaustivité pourrait varier selon la région de diagnostic puisque la région Ile de France regroupe les sites les plus importants. Pour la DOVIH, il existe une disparité selon les régions, avec une sous-déclaration allant jusqu'à 50% pour certaines [89]. Enfin le taux de participation des laboratoires à l'enquête LaboVIH varie également selon la région (de 78 à 98% en 2005) [89].

- ***Le pays de naissance de l'enfant (France, étranger)***

Pour EPF, il est probable que l'exhaustivité soit meilleure pour les enfants nés en France que pour les enfants nés à l'étranger. En effet, l'inclusion d'enfants après leur découverte de séropositivité date de 2004, et le recueil de données a été rétrospectif pour la période 2003-2004. De plus, un consentement parental est nécessaire pour inclure ces enfants, et il est probable que le risque de refus est plus important que dans le cas de mères séropositives incluses d'emblée dans la cohorte. Pour la DOVIH, il est possible que les enfants nés à l'étranger et porteurs du VIH soient moins bien dépistés que les enfants nés en France du fait d'un accès aux soins plus difficile pour les enfants nés à l'étranger.

- ***L'année de diagnostic (2003-2006)***

Pour la source DOVIH, on peut considérer que les délais de notification ont été pris en compte, du fait de la mise à jour de la base de données jusqu'en 2010. En revanche pour la source EPF, même si le problème de délais ne se pose pas avec la même acuité que pour la DO, le recueil de données a été clôturé en avril 2008, et il est possible que la base de données ne soit pas complète.

3.4.3. Imputation de la variable pays de naissance

Nous avons fait le choix d'estimer le nombre total de nouveaux diagnostics VIH en prenant en compte le pays de naissance (né en France, né à l'étranger). Cette variable binaire n'a jamais été recueillie dans la source LaboVIH. Cependant, cette information était partiellement disponible pour les cas de LaboVIH communs à au moins l'une des deux autres sources (DOVIH et EPF) pour lesquelles cette information avait été collectée. La proportion globale de données manquantes pour la variable pays de naissance était donc d'environ 30% (66/216) et variait selon les sources : 59.9% (64/129) pour LaboVIH, 1.8% (2/110) pour DOVIH et 0.0% pour EPF.

Les données manquantes de la variable pays de naissance ont été estimées par une méthode d'imputation multiple. Dans une étude de type capture-recapture, les variables communes aux trois sources sont peu nombreuses. Dans cette étude, la base regroupant les données des trois sources contenait les 3 variables indicatrices de sources (codées 1 lorsque l'individu est présent dans la source et 0 sinon), le sexe, l'âge et les 3 variables d'hétérogénéité de capture. Les 4 variables sexe (variable binaire), âge (variable continue), région de diagnostic (variable

catégorielle) et année de diagnostic (variable catégorielle) étaient entièrement observées et ont été incluses comme variables prédictrices dans le modèle d'imputation. Le modèle d'imputation était donc réduit aux 3 variables dites principales car utilisées dans les analyses après imputation, et à 1 variable auxiliaire, le sexe. Une fonction de lien multinomiale a été spécifiée pour la variable pays de naissance. Les variables catégorielles ont été décomposées en variables indicatrices binaires avant d'être incluses dans le modèle d'imputation.

Du fait de la proportion importante de données manquantes de la variable pays de naissance (30%) nous avons choisi de générer 100 bases de données complètes. Nous avons tenu compte dans ce choix d'indications de la littérature récente qui suggèrent de générer un nombre important de bases de données afin de limiter la perte de puissance lors des analyses [28;58].

3.5. Estimations issues de la méthode capture-recapture

3.5.1. Conditions d'application

Pour la méthode capture-recapture, la fiabilité des résultats dépend du respect de différentes conditions d'application : (1) l'absence de faux-positifs parmi les cas, (2) l'identification de tous les vrais cas communs, (4) l'homogénéité de capture, (5) la même période et la même zone géographique, (6) l'indépendance entre les sources et (7) une population d'étude fermée [86].

L'homogénéité de capture est vérifiée quand la probabilité de notification d'un cas par une source donnée est la même pour tous les cas, c'est-à-dire que cette probabilité ne dépend pas des caractéristiques individuelles des cas (par exemple l'âge, le sexe ou le lieu de naissance). Cette probabilité peut varier d'une source à une autre ou être globalement constante. L'inclusion dans l'analyse de variables dites d'hétérogénéité de capture permet cependant de prendre en compte des variations de la probabilité de notification au sein d'une même source [86].

Deux sources sont dites indépendantes si la probabilité de notification d'un cas par une source ne dépend pas de la probabilité de notification de ce cas par une autre source. Pour des analyses réalisées à partir de trois sources ou plus, l'hypothèse d'indépendance n'est plus requise puisque des termes d'interaction peuvent être inclus dans les modèles de régression de façon à ajuster sur les dépendances potentielles entre les sources.

3.5.2. Analyses préliminaires

La dépendance entre les sources peut être prise en compte dans le calcul du nombre total de cas. Ainsi, Wittes propose de comparer les estimations obtenues pour chaque combinaison des sources deux à deux et d'évaluer la dépendance entre les sources en comparant ces estimations [86]. Si un couple de sources donne une estimation très différente des autres couples, une dépendance entre ces sources peut être suspectée (positive ou négative). La dépendance entre deux sources peut être plus directement approchée par le calcul du rapport de cotes (OR) de Wittes et son intervalle de confiance à 95%.

Une analyse préliminaire réalisée à partir de la base de départ (non-imputée) a été appliquée en ajustant 8 modèles log-linéaires aux données réparties dans un tableau de contingence à 2^3 cellules selon la présence ou non des cas dans chacune des 3 sources. La variable dépendante pour chaque modèle est le logarithme népérien du nombre de cas dans chacune des 7 cellules non-vides du tableau de contingence. Ces analyses permettent d'obtenir des estimations pour la population d'étude tenant compte des dépendances entre les sources. Les 8 modèles log-linéaires construits vont du modèle le plus simple sans interaction au modèle saturé incluant les trois interactions d'ordre 1. Cette analyse préliminaire est basée sur l'hypothèse d'une homogénéité de capture au sein de chaque source et a été réalisée en utilisant le programme additionnel RECAP de STATA [90]. Les intervalles de confiance des estimations ont été calculés selon une méthode basée sur la vraisemblance proposée par Regal et Hook [91].

3.5.3. Analyses incluant les variables d'hétérogénéité

Nous avons choisi de prendre en compte trois variables potentielles d'hétérogénéité de capture : le pays de naissance (né en France, né à l'étranger), la région de diagnostic (Ile de France, province) et l'année de diagnostic (2003-2006). Les données sont donc réparties dans un tableau de contingence formé de 2^3x2x2x4 cellules, afin d'obtenir tous les croisements avec les variables d'hétérogénéité. Par ailleurs, les données se présentent sous la forme d'un fichier contenant les 100 bases de données imputées et la base de départ (dite base 0), chacune des bases étant identifiée par une variable indicatrice.

La commande permettant de générer ce tableau de contingence a donc été appliquée successivement à chacune des 100 bases de données imputées. Puis la base 0 et les bases imputées ont été regroupées dans un fichier global et une variable indicatrice permettant

d'identifier les bases a été reconstruite. Le fichier obtenu a été ensuite mis en forme de façon à permettre l'analyse jointe des 100 bases de données.

Des modèles log-linéaires ont été ajustés aux données imputées avec la commande GLM de STATA dédiée aux modèles linéaires généralisés, en spécifiant un lien logarithmique et une distribution de Poisson. Une procédure manuelle de sélection des variables pas à pas descendante a été effectuée en partant d'un modèle incluant toutes les interactions d'ordre 2 entre sources, entre sources et variables d'hétérogénéité, et entre variables d'hétérogénéité. Pour chaque modèle log-linéaire construit lors de l'analyse pas à pas descendante, les analyses ont été effectuées séparément puis de façon jointe selon les règles de Rubin, en utilisant le module MI ESTIMATE de STATA 11 [65]. La sélection des variables d'interaction a été basée sur un test de la significativité des coefficients, par un test de Student pour les variables binaires et par un test de Fisher pour les interactions avec la variable catégorielle année de diagnostic.

Les estimations de la taille de la population, calculées comme une somme d'exponentielles de coefficients de régression, ont été obtenues par des commandes spécifiques à l'imputation multiple. Leurs variances respectives ont été estimées en appliquant la méthode delta. Les intervalles de confiance des estimations ont été calculés en utilisant une approximation de Student, avec un nombre de degrés de liberté dépendant pour chaque variable des variances intra et inter-imputation et du nombre de bases imputées.

Classiquement, dans les études appliquant la méthode capture-recapture, le choix du modèle final est basé sur l'adéquation du modèle aux données par la statistique du rapport de vraisemblances (entre le modèle testé et le modèle saturé) dit test de la déviance (G2), mais aussi sur l'AIC (Akaïke Information Criterion) et le DIC correspondant au BIC (Bayesian Information Criterion) adapté par Draper [92;93]. Les critères AIC et DIC ont été calculés pour chaque base de données imputées selon les expressions suivantes : $AIC = G^2 - 2(ddl)$ et $DIC = G^2 - \left(\ln\left(N_{obs}/2\pi\right)\right)\cdot\left(ddl\right)$ où ddl est le nombre de degrés de liberté associé à chaque modèle et N_{obs} est le nombre de cellules renseignées du tableau de contingence.

L'approche naïve consistant à faire la moyenne des tests de la déviance sur l'ensemble des bases de données imputées ne donne pas une p-valeur valide [28]. Une approche proposée par Meng et Rubin [66], et récemment illustrée par Marshall et al. [94], a été appliquée pour calculer un test de la déviance global sur les données imputées et la p-valeur correspondante.

Les coefficients de régression de chacun des 100 modèles log-linéaires ont été contraints aux valeurs des coefficients de régression obtenus par l'analyse jointe, c'est-à-dire à la moyenne des coefficients de régression sur les 100 bases de données. Les estimations des AIC et DIC ont alors été calculées comme la moyenne des 100 AIC et DIC obtenues avec ce modèle contraint.

La stratégie d'analyse retenue consiste à sélectionner un seul modèle final, et fait ainsi abstraction de l'incertitude liée au choix du modèle. Nous avons donc calculé un estimateur composite dérivé de tous les *K* modèles impliqués dans la procédure pas-à-pas descendante comme suggéré par Draper [92]. Nous avons choisi de retenir l'estimateur pondéré par le DIC exprimé comme $\hat{N}_{WDIC} = \sum_{i=1}^{K} \hat{N}_i \cdot e^{-(DICi/2)} \Big/ \sum_{i=1}^{K} e^{-(DICi/2)}$ où $\hat{N}_i$ est l'estimateur associé au modèle *i*, et DIC_i est le *DIC* associé au modèle *i*, $i(i = 1,...,K)$.

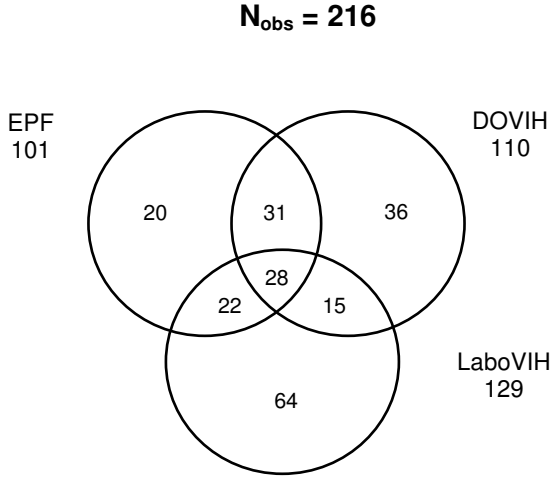
L'exhaustivité des sources a été estimée en divisant le nombre de nouveaux diagnostics rapporté par chaque source par le nombre total estimé par le modèle log-linéaire final. L'exhaustivité a également été calculée pour chaque strate des variables pays de naissance, région de diagnostic et année de diagnostic. Le taux annuel de nouveaux diagnostics VIH a été obtenu en faisant le rapport du nombre estimé de nouveaux diagnostics VIH sur la taille de la population des enfants de moins de 13 ans recensés en France métropolitaine en décembre 2007 [95]. Ce taux a également été calculé selon le pays de naissance, en se basant sur le nombre d'enfants de moins de 13 ans nés en France ou à l'étranger.

3.6. Résultats

3.6.1. Cas communs aux sources

Entre le 1er janvier 2003 et le 31 décembre 2006, 216 nouveaux diagnostics d'infection à VIH chez des enfants de moins de 13 ans en France métropolitaine ont été recensés par au moins l'une des trois sources (Figure 2.3).

Figure 2.3 – Distribution du nombre de diagnostics d'infection à VIH selon la source



DOVIH : Systéme de déclaration obligatoire du VIH
EPF : Enquête périnatale française
LaboVIH : Surveillance de l'activité de dépistage par les laboratoires

3.6.2. Imputation de la variable pays de naissance

Après imputation de la variable pays de naissance, 60% des enfants nouvellement diagnostiqués étaient nés à l'étranger (Tableau 2.11). La distribution des nouveaux diagnostics selon le pays de naissance et la classe d'âge (< 1 an, ≥ 1 an) variait selon la source. Ainsi, la proportion d'enfants nés à l'étranger et âgés de plus de 1 an était plus importante dans les sources EPF et surtout DOVIH que dans la source LaboVIH.

Tableau 2.11 – Distribution des nouveaux diagnostics d'infection à VIH identifiés par chacune des trois sources selon l'âge et le pays de naissance, après imputation multiple

	Total	DOVIH		EPF		Labo VIH	
		< 1 an	≥ 1 an	< 1 an	≥ 1 an	< 1 an	≥ 1 an
Né à l'étranger	130 (60.2%)	1 (12.5%)	24 (85.8%)	1 (12.5%)	9 (75%)	4 (33.3%)	37 (71.2%)
Né en France	86 (39.8%)	7 (87.5%)	4 (14.3%)	7 (87.5%)	3 (25%)	8 (66.7%)	18 (28.8%)
	216	8	28	8	12	12	52

DOVIH : Systéme de déclaration obligatoire du VIH ; EPF : Enquête périnatale française ; LaboVIH : Surveillance de l'activité de dépistage par les laboratoires.

3.6.3. Estimation du nombre total de diagnostics et évaluation de la dépendance

- ***Analyses préliminaires***

L'analyse capture-recapture, appliquée à chaque combinaison de sources deux à deux, a donné les résultats suivants : l'estimation du nombre de nouveaux diagnostics VIH produite par l'analyse des sources DOVIH et EPF (N_{est}=188 ; IC 95% [171;206]) était plus petite que celle issue de l'analyse des sources LaboVIH et EPF (N_{est}=261 ; IC 95% [224;297]) et des sources LaboVIH et DOVIH (N_{est}=330 ; IC 95% [272;389]), suggérant une dépendance positive entre les sources DOVIH et EPF. Le calcul des odds ratios de Wittes a confirmé la dépendance entre les sources DOVIH et EPF (OR = 5.4 ; IC 95% [2.5;12.1]) et a suggéré l'existence d'une dépendance positive entre LaboVIH et EPF (OR = 2.2 ; IC 95% [1.0;4.8]).

L'analyse préliminaire incluant les 3 sources et les dépendances entre sources dans les modèles log-linéaires a donné une estimation de 369 (IC 95% [294 ;532]) nouveaux diagnostics pour la période 2003-2006 (Tableau 2.12). Ce modèle a pris en compte deux interactions entre sources (DOVIH*EPF et EPF*LaboVIH).

Tableau 2.12 – Estimations du nombre de nouveaux diagnostics VIH par les modèles log-linéaires, analyses préliminaires

Modèles log-linéaires	$\hat{n}$	$\hat{N}$	IC 95%	ddl	G^2	p	AIC	BIC
Dépendances entre sources								
LaboVIH*DOVIH, LaboVIH*EPF, DOVIH*EPF	126	342	259,573	0	0	1	0	0
LaboVIH*DOVIH, LaboVIH*EPF	23	239	225,263	1	18.83	$<10^{-4}$	16.83	16.89
LaboVIH*DOVIH, DOVIH*EPF	58	274	243,331	1	3.78	0.05	1.78	1.84
LaboVIH*EPF, DOVIH*EPF	153	369	294,521	1	0.24	0.63	-1.76	-1.71
LaboVIH*DOVIH, EPF	29	249	234.272	2	18.49	$<10^{-4}$	14.49	14.6
LaboVIH*EPF, DOVIH	51	267	245,300	2	30.12	$<10^{-4}$	26.12	26.23
DOVIH*EPF, LaboVIH	85	301	268,349	2	5.96	0.05	1.96	2.07
LaboVIH, DOVIH, EPF	49	265	246,292	3	30.2	$<10^{-4}$	24.2	24.36

$\hat{n}$: estimation du nombre de diagnostics recensés dans aucune des trois sources ; $\hat{N}$: estimation du nombre de diagnostics ; IC 95% : intervalle de confiance à 95% ; ddl : nombre de degrés de liberté ; G^2 : déviance ; p : p-valeur ; AIC : Akaïke Information Criterion ; BIC : Bayesian Information Criterion.

- ***Modèles log-linéaires incluant les variables d'hétérogénéité***

En considérant les interactions avec les variables d'hétérogénéité, le modèle ayant le plus petit AIC et un test de la déviance non-significatif (p>0.05) a donné une estimation de 387 (IC 95% [271;503]) nouveaux diagnostics VIH durant la période 2003-2006 (Tableau 2.13). Ce modèle (modèle 8) incluait deux interactions entre sources (DOVIH*EPF et EPF*LaboVIH) et des interactions entre sources et variables d'hétérogénéité (DOVIH*pays de naissance, EPF*pays de naissance, DOVIH*région de diagnostic, EPF*région de diagnostic, LaboVIH*région de diagnostic et EPF*année de diagnostic). L'estimation pondérée correspondante était de 384 nouveaux diagnostics VIH. L'estimation annuelle du nombre de nouveaux diagnostics diminuait au cours du temps, variant de 108 en 2003 à 89 en 2006 (Tableau 2.14).

- ***Exhaustivité***

L'exhaustivité estimée des trois sources combinées était de 55.8% (IC 95% [42.9 ;79.7]), mais variait selon la source (Tableau 2.4). L'exhaustivité de la source DOVIH (28.4%) et d'EPF (26.1%) était plus basse que celle de LaboVIH (33.3%). L'exhaustivité a diminué légèrement depuis 2004 aussi bien pour DOVIH que pour EPF, puis plus nettement pendant la dernière année d'étude (2006). L'exhaustivité était plus élevée en région parisienne qu'en province pour les trois sources, et était meilleure pour les enfants nés en France que pour les enfants nés à l'étranger pour les sources EPF et LaboVIH.

- ***Taux annuel de nouveaux diagnostics***

En se basant sur le nombre estimé de nouveaux diagnostics présentés dans le Tableau 2.14, le taux annuel de nouveaux diagnostics de VIH chez les enfants de moins de 13 ans en France métropolitaine était de 9.1 cas par million (IC 95% [5.7 ;12.5]) en 2006. Ce taux annuel était 38 fois plus élevé pour les enfants nés à l'étranger (161.1 cas par million) que pour les enfants nés en France (4.2 cas par million).

Tableau 2.13 – Estimation du nombre de nouveau diagnostics par les modèles log-linéaires incluant les variables d'hétérogénéité, analyse pas à pas descendante

Modèles log-linéaires	$\hat{n}$	$\hat{N}$	IC 95%	df	G^2	p	AIC	DIC
Modèle 1: DO*LABO, DO*EPF, LABO*EPF,DO*pays, Labo*pays, EPF*pays, DO*région, Labo*région, EPF*région, DO*année, Labo*année, EPF*année, pays*région, pays*année, région*année.	108.3	324.3	210.9,437.6	78	93.84	0.07	-62.16	-130.85
Modèle 2: DO*LABO, DO*EPF, LABO*EPF,DO*pays, EPF*pays, DO*région, Labo*région, EPF*région, DO*année, Labo*année, EPF*année, pays*région, pays*année, région*année.	106.9	322.9	211.2,434.5	79	93.95	0.08	-64.05	-133.62
Modèle 3: DO*LABO, DO*EPF, LABO*EPF,DO*pays, EPF*pays, DO*région, Labo*région, EPF*région, DO*année, Labo*année, EPF*année, pays*region, pays*année.	108.6	324.6	211.7,437.5	82	96.17	0.10	-67.83	-140.05
Modèle 4: DO*LABO, LABO*EPF,DO*pays, EPF*pays, DO*région, Labo*région, EPF*région, DO*année, Labo*année, EPF*année, pays*région, pays*année.	148.0	364.0	257.9,470.1	83	96.82	0.10	-69.18	-142.27
Modèle 5: DO*LABO, LABO*EPF,DO*pays, EPF*pays, DO*région, Labo*région, EPF*région, DO*année, Labo*année, EPF*année, pays*année.	152.3	368.3	259.9,476.7	84	98.31	0.10	-69.69	-143.66
Modèle 6: DO*LABO, LABO*EPF,DO*pays, EPF*pays, DO*région, Labo*région, EPF*région, Labo*année, EPF*année, pays*année.	143.3	359.3	259.1,459.6	87	102.28	0.09	-71.72	-148.34
Modèle 7: DO*LABO, LABO*EPF,DO*pays, EPF*pays, DO*région, Labo*région, EPF*région, EPF*année, pays*année.	140.8	356.8	258.0,455.6	90	105.54	0.09	-72.46	-153.71
Modèle 8: DO*LABO, LABO*EPF,DO*pays, EPF*pays, DO*région, Labo*région, EPF*région, EPF*année.	171.0	387.0	271.0,502.9	93	112.72	0.07	-73.28	-155.17
Modèle 9: DO*LABO, LABO*EPF,DO*pays, EPF*pays, DO*région, Labo*région, EPF*région.	171.0	387.0	271.0,502.9	96	118.48	0.05	-73.52	-158.06
Estimation pondérée par le DIC		384.3						

$\hat{n}$: estimation du nombre de diagnostics recensés dans aucune des trois sources ; $\hat{N}$: estimation du nombre de diagnostics ; IC 95% : intervalle de confiance à 95% ; ddl : nombre de degrés de liberté ; G^2 : déviance ; p : p-valeur ; AIC : Akaïke Information Criterion ; DIC : Draper Information Criterion ; estimation pondérée : estimation de $\hat{N}$ pondérée par le DIC de tous les modèles (p>0.05).

Tableau 2.14 – Estimation de l'exhaustivité pour chaque source selon l'année de diagnostic, le pays de naissance et la région de diagnostic

Strate	Total					DOVIH			EPF			LaboVIH		
	Nest	IC 95%	N_{obs}	Exh(%)	IC 95%	N_{obs}	Exh(%)	IC 95%	N_{obs}	Exh(%)	IC 95%	N_{obs}	Exh(%)	IC 95%
Année de diagnostic														
2003	107.6	72.4-142.7	60	55.8	42.0-82.9	30	27.9	21.0-41.4	28	26.0	19.6-38.7	30	27.9	21.0-41.4
2004	99.1	68.9-129.4	59	59.5	45.6-85.7	35	35.3	27.0-50.8	32	32.3	24.7-46.5	35	35.3	27.0-50.8
2005	91.7	62.4-120.9	53	57.8	43.8-85.0	27	29.5	22.3-43.3	27	29.5	22.3-43.3	34	37.1	28.1-54.5
2006	88.6	55.4-121.8	44	49.7	36.1-79.4	18	20.3	14.8-32.5	14	15.8	11.5-25.3	30	33.9	24.6-54.2
Pays de naissance														
France	152.6	100.3-204.9	86	56.4	42.0-85.7	37	24.2	18.1-36.9	47	30.8	22.9-46.8	55	36.0	26.8-54.8
Etranger	234.4	158.9-309.9	130	55.5	42.0-81.8	73	31.1	23.6-45.9	54	23.0	17.4-34.0	74	31.6	23.9-46.6
Région de diagnostic														
Ile de France	198.1	154.7-241.4	139	70.2	57.6-89.9	79	39.9	32.7-51.1	79	39.9	32.7-51.1	82	41.4	34.0-53.0
Province	188.9	101.0-276.8	77	40.8	27.8-76.2	31	16.4	11.2-30.7	22	11.6	7.9-21.8	47	24.9	17.0-46.5
Total	387	271-503	216	55.8	42.9-79.7	110	28.4	21.9-40.1	101	26.1	20.1-37.3	129	33.3	25.6-47.6

$\hat{N}$: estimation du nombre de diagnostics ; IC 95% : intervalle de confiance à 95% ; N_{obs} : nombre de cas observés ; Exh(%) : exhaustivité en %.

3.7. Discussion

Cette étude a permis d'estimer pour la première fois le nombre total de nouveaux diagnostics d'infection à VIH chez les enfants âgés de moins de 13 ans en France métropolitaine. Cette estimation, réalisée sur la période 2003-2006, s'élève à 369 (IC 95% [294 ;532]). Les trois sources combinées ont permis de recenser plus de la moitié des cas, alors que l'exhaustivité du système de déclaration obligatoire et de la cohorte EPF sont relativement faibles (<30%).

La variable pays de naissance n'était pas collectée dans la source LaboVIH, mais était presque complète pour les deux autres sources. L'approche standard dans les études capture-recapture consiste à ignorer les variables qui ne sont pas disponibles dans chacune des sources, ce qui peut mener à des estimations biaisées de la taille de la population [96]. Une approche classique pour traiter les bases de données incomplètes consiste à remplacer chaque donnée manquante par une donnée imputée et à analyser la base de données comme si elle était complète. Nous avons vu que de telles méthodes d'imputation simple ne sont pas valides statistiquement, puisqu'elles peuvent donner des estimations biaisées et qu'elles induisent une sous-estimation systématique des variances [12]. Deux méthodes sont couramment recommandées pour traiter les données manquantes de façon adéquate : l'estimation par la méthode de maximisation de la vraisemblance et l'imputation multiple. Il a été montré que ces deux méthodes sont asymptotiquement équivalentes, sachant qu'elles sont basées sur la même hypothèse que les données sont manquantes aléatoirement [17]. Dans cette étude, la variable pays de naissance est manquante du fait du schéma d'étude puisque non collectée dans la source LaboVIH. De ce fait, le mécanisme de données manquantes ne dépend pas des valeurs non-observées de la variable incomplète, et l'hypothèse MAR est vérifiée.

L'imputation de données manquantes dans des applications de la méthode capture-recapture a été rapportée dans quelques études seulement [96-98], et la méthode d'estimation par maximisation de la vraisemblance utilisant un algorithme d'Estimation-Maximisation (EM) avait été appliquée. Van der Heidjen et al. [98] ont estimé les données manquantes de variables d'hétérogénéité de capture non collectées dans toutes les sources telles que le sexe et la région de résidence. Cependant, les auteurs soulignent que l'algorithme EM implique souvent une phase d'intégration numérique complexe (étape E) et que l'imputation multiple présente l'avantage d'être plus simple

à appliquer dans la plupart des cas de figures, surtout pour des variables incomplètes à structure continue [99].

Au cours du processus d'analyse pas à pas descendante, la sélection du modèle final, incluant les variables d'hétérogénéité de capture, a été effectuée sur la base des critères AIC et DIC. Le modèle retenu est donc celui pour lequel les valeurs de l'AIC et du DIC sont les plus petites, et pour lequel l'adéquation aux données apparait correcte d'après les résultats du test du rapport de vraisemblances dit test de la déviance. Ce test a été réalisé d'après une approche proposée par Meng et Rubin, et a donné des p-valeurs légèrement plus basses que celle obtenues par l'approche naïve. Cependant, une limitation de cette étude est due au fait que les critères AIC et DIC ont été calculés en prenant la moyenne de leurs valeurs pour les 100 bases de données imputées et que les résultats doivent donc être interprétés avec précaution [100]. Ainsi, en se basant sur les valeurs de ces critères, les différences entre les modèles emboîtés pourraient être surestimées, ce qui pourrait mener à retenir un modèle trop complexe. Cependant, dans notre application, le modèle sélectionné selon l'AIC et le DIC (modèle 9) était plus parcimonieux que celui sélectionné selon la statistique du rapport de vraisemblances (modèle 8) et les estimations fournies par ces deux étaient identiques (387.0). Nous avons fait le choix de retenir le modèle le moins parcimonieux (modèle 8) de façon à s'assurer que l'adéquation du modèle était correcte selon le test du rapport de vraisemblance (p=0.07). De plus, le modèle 8 incluait une interaction entre la source EPF et l'année de diagnostic, qui nous paraissait pertinente d'un point de vue épidémiologique.

Notons que dans l'étude de Zwane et al. [96], la sélection du modèle final, basée sur les critères AIC et BIC après imputation multiple, a été réalisée à partir d'une seule base imputée supplémentaire (M+1), générée pour le processus de sélection du modèle. Par ailleurs, Zwane et al. ont souligné que l'approche par imputation multiple permettait d'obtenir directement les variances et intervalles de confiance des estimations. En revanche, l'approche par maximisation de la vraisemblance requiert de calculer les intervalles de confiance des estimations par bootstrap (paramétrique ou non). Notons cependant que l'incertitude liée au choix du modèle final n'a pas été prise en compte dans notre approche par imputation multiple. Des recherches complémentaires dans ce domaine pourraient permettre d'incorporer cette incertitude dans le calcul de la variance après imputation multiple.

Conclusion

Dans cette étude, une seule variable est incomplète et les données sont manquantes selon un schéma univarié et un mécanisme de type MCAR. La construction du modèle d'imputation s'est révélée simple puisque peu de variables étaient disponibles. Au cours d'analyses antérieures, les données manquantes avaient été estimées par réaffectation proportionnelle. L'imputation multiple permet de tenir compte des informations des variables d'hétérogénéité qui sont par la suite incluses dans les modèles d'analyses et ainsi d'éviter des biais. De plus, le calcul de la variance tient compte du processus d'estimation des données manquantes.

La phase d'analyse a nécessité une restructuration des bases permettant la création de tableaux de contingence pour chaque base imputée. Puis, les estimations à partir des modèles log-linéaires ont été obtenues par des commandes spécifiques à l'imputation. Cependant, cette application souligne le problème de l'étape de post-estimation après imputation multiple. En effet, la littérature récente précise que les statistiques dérivées de la vraisemblance ne peuvent être combinées simplement [100]. Lors d'une stratégie d'analyse par modèles emboités, le choix du meilleur modèle est essentiellement basé sur une approximation de Student permettant de sélectionner les variables à retenir dans le modèle. L'évaluation de l'adéquation d'un modèle aux données est plus difficile, même si l'approche de Meng et Rubin permet de comparer des modèles emboités entre eux.

CHAPITRE 3

PROCESSUS D'IMPUTATION MULTIPLE APPLIQUE A UNE ETUDE TRANSVERSALE DANS UN SYSTEME DE SURVEILLANCE DE L'HEPATITE C

L'infection par le virus de l'hépatite C (VHC) demeure un important problème de santé publique en France du fait de sa prévalence (1.05% en 1994 [101], 0.84% en 2004 [102]), du risque de complications hépatiques graves et de la transmission pérenne parmi les usagers de drogues intraveineuses (UDI).

De ce fait, un programme national a été mis en place en 1998 avec pour objectif de réduire la transmission du VHC, d'améliorer le dépistage dans les populations à risque ainsi que l'accès aux soins pour les patients infectés chroniquement par le VHC. Afin de contribuer à l'évaluation de ce programme, l'InVS a mis en place un système de surveillance depuis 2000.

Nous présenterons dans une première partie une étude visant à identifier les facteurs de risque de complications hépatiques graves de l'infection par le VHC au sein de ce système de surveillance, après estimation des données manquantes par imputation multiple sous l'hypothèse MAR. Dans une deuxième partie nous détaillerons une méthode d'analyse de sensibilité par pondération permettant de tester la robustesse de ces résultats à un éventuel non-respect de l'hypothèse MAR.

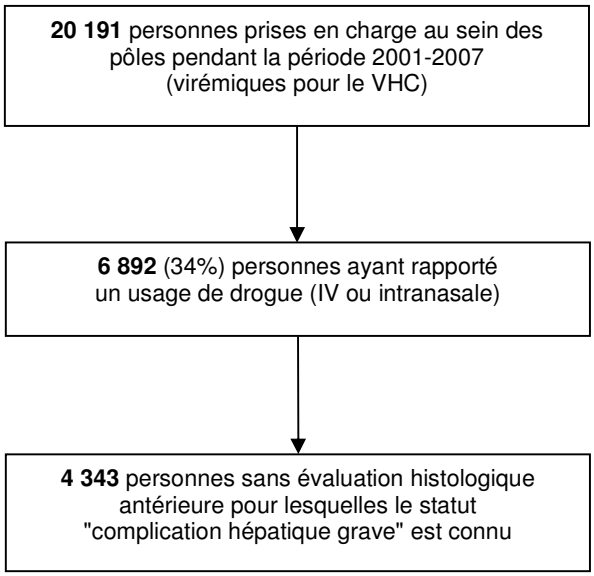
1. Analyse étiologique des facteurs de risque de complications hépatiques graves

1.1. Population d'étude et critères d'inclusion

Ce système de surveillance du VHC repose sur la participation volontaire de 26 des 30 pôles de référence existants, localisés dans des centres hospitaliers universitaires [103]. Son objectif est d'évaluer les modifications au cours du temps des caractéristiques épidémiologiques et cliniques des patients infectés par le VHC et nouvellement pris en charge dans un pôle de référence. Les patients peuvent être référés dans un pôle par un médecin généraliste, un spécialiste ou se présenter de leur propre initiative. Sont inclus dans le système de surveillance les patients porteurs d'Ac anti-VHC et admis en consultation ou hospitalisés pour la première fois dans un pôle de référence de l'hépatite C.

En 2000-2001, la proportion de complications hépatiques graves de type cirrhose était de 10% à la première visite [104], en baisse par rapport à 1993-1995 (20%). Par ailleurs, il a été montré que la prévalence du VHC chez les UDI variait de 55% à 60% [102;105]. L'objectif de notre étude était de déterminer les facteurs de risque de complications hépatiques graves (cirrhose, carcinome hépatocellulaire) dans une population de personnes virémiques pour le VHC et ayant rapporté un usage de drogue, par voie intraveineuse ou intranasale, au moins une fois dans leur vie. Ont été retenus les patients n'ayant pas eu d'évaluation histologique antérieure du foie et pour lesquels le statut de complication hépatique était connu. Au final, 4343 patients ont été inclus pour la période 2001-2007 (Figure 3.1).

Figure 3.1 - Inclusion des patients du système de surveillance dans l'étude



1.2. Recueil de données

1.2.1. Données collectées

Un formulaire standard a été utilisé pour collecter les données en routine. Les données disponibles étaient : l'année du dernier test négatif, la date du test, la charge virale qualitative (ARN), le génotype du VHC, une éventuelle co-infection par le VIH ou/et l'hépatite B (antigènes HBs), l'historique de consommation excessive d'alcool définie comme plus de 210 grammes par semaine d'éthanol pur chez les femmes et 280 grammes par semaine chez les hommes, et le nombre d'années de consommation excessive.

La sévérité de l'atteinte hépatique a été objectivée par une biopsie hépatique (score Metavir) ou/et par une évaluation clinique incluant un examen clinique, des dosages biochimiques (ALAT) et des examens d'imagerie médicale, essentiellement une échographie abdominale.

1.2.2. Examen des données manquantes

La base de données initiale contenait 20191 cas et 126 variables. Une base restreinte a été constituée à ce stade, limitée à 4343 usagers de drogues et à 27 variables.

- ***Variables construites***

La variable d'intérêt étudiée est la présence ou non de complications hépatiques graves (cirrhose ou carcinome hépatocellullaire) basée sur un examen clinique seul (pas de biopsie). Il a été choisi de ne retenir que les patients pour lesquels cette information était disponible. La variable d'intérêt est donc complète.

Les variables retenues pour être incluses dans l'analyse univariée sont présentées dans le Tableau 2.15 et détaillées ci-dessous :

- Données sociodémographiques : le sexe et l'âge ($\leq$ 40 ans, >40 ans)

- Historique de la maladie : la date de prise en charge du patient (2001-2007), le délai de prise en charge du patient défini comme le délai entre la sérologie positive réalisée dans un pôle de référence ou une autre structure et la prise en charge du patient au sein du pôle (<1an, $\geq$1an), la durée de l'infection par le VHC approximée par la durée entre la dernière sérologie VHC négative réalisée pendant la période d'usage de drogue ou l'année de première injection et la sérologie VHC positive (<18 ans, $\geq$18 ans, selon la médiane de la durée), et une consommation excessive d'alcool dans le passé (plus de 210 g d'éthanol par semaine pour la femme, et plus de 280g chez l'homme).

- Examens complémentaires de laboratoire : une co-infection par le VIH (statut sérologique VIH), une co-infection par le virus de l'hépatite B (VHB) (antigènes HBs), et le génotype du VHC.

Tableau 3.1 – Description des variables retenues pour l'analyse univariée

Variables	Codage	Données manquantes (%)
Année de prise en charge	2001-2007	0
Sexe	F/H	0
Age	≤40 ans/>40 ans	0
Durée de l'infection par le VHC	≤18 ans/>18 ans	14.6
Délai de prise en charge	≤1an/>1an	10.4
Consommation excessive passée d'alcool	non/oui	11.1
Co-infection par le VIH	non/oui	16.3
Co-infection par le VHB (AgHbs)	non/oui	15.7
VHC Génotype 3	non/oui	26.3

- ***Examen quantitatif***

A l'exception des variables sexe, âge et année d'inclusion, toutes les variables retenues pour l'analyse statistique sont incomplètes, avec une proportion de données manquantes variant de 10.4 à 26.3 % (Tableau 3.1).

En réalisant un examen de la répartition des données manquantes parmi les différentes combinaisons des 6 variables incomplètes, la répartition apparait arbitraire, c'est-à-dire que la proportion de données manquantes est du même ordre de grandeur (<3%) pour chacune des combinaisons de ces 6 variables. De ce fait, la perte d'effectifs attendue lors d'une analyse cas-complet est importante.

Ainsi, une analyse multivariée incluant les 6 variables incomplètes porterait sur seulement 1818 individus au lieu des 4343 initiaux, soit sur 41.8% de l'effectif. On peut donc craindre que le processus de sélection des variables lors d'une analyse multivariée cas-complet ne soit faussé au profit des variables les mieux renseignées.

- ***Examen qualitatif***

Les analyses prévues étaient à la fois descriptives et de nature étiologique. Pour l'identification des facteurs de risque, il est informatif d'effectuer un examen qualitatif du mécanisme de non-réponse, dépendant ou non de la variable d'intérêt, et ainsi d'identifier le risque de biais en analyse cas-complet. Une variable indicatrice binaire de données manquantes est générée pour chaque variable incomplète. Des analyses univariées et multivariées permettent d'expliquer chaque variable indicatrice à partir des 9 variables retenues. L'objectif de ces analyses est de

rechercher des liens significatifs entre, d'une part chaque variable indicatrice, et d'autre part la variable à expliquer et les variables explicatives. Ces analyses peuvent être réalisées à partir de la base incomplète, mais une perte d'effectifs importante est attendue en analyse multivariée. De ce fait, comme suggéré par White et al. [28], ces analyses ont été effectuées à partir des données imputées.

Tableau 3.2 – Analyse multivariée des variables indicatrices de données manquantes à partir des variables retenues pour l'analyse multivariée

	Rdurée	Rdélai	Ralcool	Rvih	Rvhb	Rgeno3
Complication hépatique grave		**x**	**x**	**x**		
Sexe					x	
Age	x		x	x		x
Année de prise en charge	x	x	x			x
Durée de l'infection VHC			x	x		x
Délai de prise en charge					x	
Consommation excessive d'alcool					x	
Co-infection par le VIH		x				
Co-infection par le VHB (AgHbs)						
VHC Génotype 3						

La zone en gris clair représente les croisements entre les indicatrices de données manquantes et la variable à expliquer, et les zones en gris foncé les intersections entre chaque variable d'exposition et son indicatrice de données manquantes.
X représente une liaison significative entre les variables définie comme un test de Wald significatif ($p \leq 0.05$).

Au vu des résultats présentés dans le Tableau 3.2, un mécanisme de type MCAR peut être exclu puisque des liens existent entre les indicatrices de données manquantes et les autres variables. De plus, cette analyse permet d'identifier un mécanisme de type MAR(ME) pour les variables délai de prise en charge, consommation d'alcool et co-infection par le VIH, et MAR(E) pour les autres variables incomplètes. La terminologie MAR(ME), définie dans le chapitre 1, signifie que le mécanisme de données manquantes dépend conjointement de la variable à expliquer (M) et d'au moins une variable d'exposition (E). Un risque de biais en analyse multivariée cas-complet doit donc être envisagé. Par ailleurs, un mécanisme de type MNAR ne peut être exclu puisque certaines indicatrices de données manquantes sont liées à des variables incomplètes. Par exemple, l'indicatrice de la variable sérologie VHB est significativement liée aux variables délai de prise en charge et consommation d'alcool qui sont incomplètes. Le mécanisme de données manquantes peut donc dépendre des valeurs non-observées de ces variables.

Au vu de l'impact attendu des données manquantes en analyse cas-complet, il a été décidé d'effectuer une analyse par imputation multiple.

1.3. Construction et validation du modèle d'imputation

Au cours de la première étape du processus d'imputation, il est important d'identifier les relations entre les variables par une analyse cas-complet. Lorsque l'analyse prévue après imputation est étiologique, l'analyse univariée cas-complet permet de sélectionner les variables qui seront incluses dans le modèle d'analyse multivariée. A ce stade sont identifiées les interactions éventuelles qui doivent être prises en compte lors de l'imputation. Il est également informatif de réaliser une analyse cas-complet multivariée afin d'obtenir des estimations qui pourront être comparées avec celles obtenues après imputation multiple (valeur centrale, variance).

1.3.1. Construction des équations de prédiction

- ***Sélection des variables***

Le processus de sélection des variables a été détaillé dans le chapitre 1. Ainsi, le modèle d'imputation doit être au moins aussi général que le modèle d'analyse c'est-à-dire qu'il doit au moins inclure la variable d'intérêt ainsi que les covariables retenues au terme de l'analyse univariée, soit 10 variables en tout (6 incomplètes et 4 complètes). Cependant, le modèle peut acquérir des qualités prédictrices supplémentaires par l'inclusion de variable dites auxiliaires qui n'interviennent que dans le processus d'estimation des données manquantes. Dans les bases de données de surveillance, la proportion de données manquantes est souvent élevée du fait du mode de recueil et, dans notre étude, seules deux variables partiellement renseignées ont pu être retenues comme variables auxiliaires potentielles. Plusieurs modèles d'imputation ont donc été construits, le plus complet incluant les variables auxiliaires pays de naissance et dosage des transaminases (ALAT). Cependant, ces variables étaient incomplètes (>10% données manquantes) et se sont révélées faiblement prédictrices, c'est-à-dire que l'on a observé des relations non-significatives entre ces deux variables et la plupart des indicatrices de données manquantes des variables à imputer. Au final, le modèle le plus simple incluant 10 variables et ne contenant aucune variable auxiliaire a été retenu.

- ***Fonction de lien***

Les variables incomplètes incluses dans le modèle d'imputation étaient toutes binaires et ont été imputées par régression logistique. La variable date d'inclusion était catégorielle et complète. Elle a été incluse telle quelle dans les équations de prédiction. Il a été montré plus récemment, après la finalisation de cette étude, qu'il était préférable de décomposer toutes les variables catégorielles, même complètes, en variables indicatrices binaires [56].

1.3.2. Choix du nombre de bases et de cycles

- ***Nombre de bases***

Les arguments liés à l'efficacité statistique ont été retenus (chapitre 1), c'est-à-dire que le nombre de bases imputées doit être du même ordre de grandeur que la fraction de cas incomplets pour tolérer une perte d'efficacité statistique inférieure à 5%, par rapport à un nombre infini d'imputations. Cette proportion de données manquantes varie de 10 à 26% selon les variables et 30 bases de données ont été imputées. Il faut noter que, compte-tenu de la petite taille de la base de données, il aurait été aisé d'imputer un plus grand nombre de bases sans trop alourdir les processus d'imputation et d'analyse.

- ***Nombre de cycles***

Le programme additionnel utilisé (ICE) spécifie 10 cycles par défaut car la convergence de l'échantillonneur de Gibbs est rapide si le modèle est correctement spécifié, et que les distributions de chaque variable incomplète convergent vers une distribution jointe. Nous avons fait le choix conservateur de spécifier 20 cycles car la plupart des variables incluses dans le modèle d'imputation sont incomplètes.

1.3.3. Analyse des bases de données imputées

Les 30 bases de données imputées sont analysées automatiquement de façon séparée puis combinée selon les règles de Rubin. Le modèle utilisé pour les analyses cas-complet et après imputation multiple est un modèle de régression logistique incluant les 6 variables binaires complétées ainsi que le sexe et l'âge en deux catégories ($\leq$ 40 ans, >40 ans). Pour les deux analyses, une stratégie d'analyse manuelle pas à pas descendante a été menée. La sélection des

variables a été effectuée par des tests de rapport de vraisemblances sur modèles emboités pour l'analyse cas-complet. Une stratégie différente a été utilisée pour l'analyse après imputation, puisque la vraisemblance est une statistique qui ne peut être combinée sur toutes les bases imputées [28]. De ce fait, la sélection des variables a été basée sur un test des coefficients de régression. Après imputation multiple, le test de Wald est approximé par un test de Student dont le nombre de degrés de liberté dépend des variances inter et intra-imputation et du nombre de bases.

1.3.4. Diagnostic de l'imputation

- ***Comparaisons des données observées et imputées***

Le Tableau 3.3 présente, pour chaque variable imputée, les proportions observées et estimées. Les résultats issus des deux bases de données sont proches, puisque les proportions observées de chaque variable appartiennent à l'intervalle de confiance à 95% des proportions estimées. Cet examen a été effectué pour les différents modèles d'imputation testés et les proportions estimées ne différent pas significativement selon les modèles. De ce fait, le choix du modèle le plus simple contenant 10 variables apparaît justifié.

Tableau 3.3 – Comparaison des proportions observées et imputées

	Durée de la maladie	Délai de prise en charge	Consommation d'alcool	Sérologie VIH	Sérologie VHB	Genotype3
	(≥18 ans)	(≥1an)	(excessive)	(positive)	(aghbs+)	(positif)
Observé (%)	53.95	55.6	47.82	8.09	2.43	34.91
Estimé (%)	52.20	58.3	46.12	8.19	2.76	35.20
IC 95%	46.4-58.0	51.3-65.0	38.1-54.2	5.1-11.3	0.7-4.9	31.1-39.3

- ***Choix du nombre de bases***

L'efficacité statistique obtenue avec l'imputation de 30 bases de données est supérieure à 99% pour toutes les variables, ce qui signifie que l'on observe une perte d'efficacité statistique inférieure à 1% par rapport à un nombre infini d'imputations. Des outils de diagnostic plus élaborés tels que l'examen de l'erreur de Monte-Carlo n'ont pas été appliqués à cette analyse, mais ils sont illustrés dans le chapitre 4, dans un cas de figure plus complexe.

1.4. Résultats et discussion

Le Tableau 3.4 présente les résultats comparés des deux analyses cas-complet et imputation multiple.

- ***Interprétation des résultats***

Les variables retenues dans les modèles finaux des deux analyses différent : la variable co-infection avec le VIH a été retenue après imputation alors que c'est la variable co-infection avec le VHB qui apparaît dans le modèle cas-complet. D'un point de vue épidémiologique et statistique, la co-infection avec le VIH apparaît plus justifiée comme facteur de risque de complications hépatiques. En effet, le nombre de cas de co-infections avec le VIH est de l'ordre de 8.1% (294 cas) versus 2.4% (89 cas) pour la co-infection avec le VHB. Même si la proportion de données manquantes est proche pour ces deux variables (16.3% versus 15.7%), la perte d'effectifs cumulés due aux données manquantes des 6 variables incomplètes a affecté davantage la variable co-infection par le VIH (c'est-à-dire que l'effectif global est plus faible pour le modèle multivarié incluant la co-infection avec le VIH). Le processus de sélection des variables a donc été faussé en faveur de la co-infection par le VHB. Il faut noter que la variable génotype 3, qui est en limite de significativité dans le modèle cas-complet final (p=0.05), apparaît nettement significative après imputation multiple (p=0.003). De plus, la variable génotype 3 n'était pas retenue lors d'une analyse cas-complet antérieure portant sur une période de surveillance plus restreinte (2001-2004, 3153 patients).

Tableau 3.4 – Analyse par régression logistique multivariée en cas-complet et par imputation multiple des facteurs de risques associés à des complications hépatiques graves

Variables		Patients (N=4343)	Complications hépatiques (%)	Données manquantes (%)	Analyse multivariée	
					Cas Complet (N*=2130) ORa (IC 95%)	Imputation Multiple (N*=4343) ORa (IC 95%)
Période d'inclusion						
	2001-2003	2330	7.0			
	2004-2007	2013	9.5		† NS	† NS
Sexe						
	femmes	993	4.2		1.0	1.0
	hommes	3350	9.3		1.8 (1.1-3.0)	2.0 (1.4-2.9)
Age						
	≤40 ans	2435	3.9		1.0	1.0
	>40 ans	1908	13.6		2.2(1.5-3.3)	2.3 (1.7-3.1)
Délai de prise en charge						
	<1an	1728	6.7			
	≥1an	2163	8.7		† NS	† NS
	manquant	452	11.5	10.4		
Durée de l'infection VHC						
	<18 ans	1709	3.0		1.0	1.0
	≥18 ans	2002	12.5		3.1 (2.0-5.1)	2.6 (1.8-3.7)
	manquant	632	8.2	14.6		
Historique de consommation excessive d'alcool						
	non	2015	4.5		1.0	1.0
	oui	1847	13.2		2.6 (1.8-3.7)	2.8 (2.2-3.7)
	manquant	481	4.4	11.1		
Statut AgHBs						
	négatif	3570	8.3		1.0	
	positif	89	13.5		2.4 (1.0-5.9)	† NS
	manquant	684	6.7	15.7		
Satut sérologique VIH						
	négatif	3342	8.2			1.0
	positif	294	14.0		† NS	1.8 (1.2-2.6)
	manquant	707	5.7	16.3		
VHC génotype 3						
	non	2083	7.2		1.0	1.0
	oui	1117	10.3		1.5 (1.1-2.0)	1.6 (1.3-2.1)
	manquant	1143	7.8	26.3		

* Nombre d'individus pris en compte dans le calcul des ORa dans le modèle final

† Odds ratio non significatif

Afin de pouvoir comparer les résultats des analyses cas-complet et imputation multiple, le même modèle incluant les 5 variables communes aux deux analyses a été appliqué aux deux bases de données (initiale et imputée). Les résultats sont présentés dans le Tableau 3.5. Les odds-ratios ajustés (ORa) ne diffèrent pas significativement entre les deux analyses puisque les ORa en analyse cas-complet appartiennent aux intervalles de confiance à 95% des ORa après imputation. Les écarts-types sont globalement réduits après imputation, puisque les effectifs sont restaurés par l'imputation (2265 à 4343), et ce malgré la prise en compte dans le calcul de la variance de la variabilité entre les bases imputées. Paradoxalement, la baisse d'écart type la plus marquée concerne une variable complète, l'âge, alors que pour la variable génotype 3 contenant une forte proportion de données manquantes, l'écart-type reste très stable entre les deux analyses. Lors de l'analyse cas-complet, les données manquantes ont induit une perte importante d'effectifs, de près de 50%, qui affecte proportionnellement davantage les variables originellement complètes comme l'âge, variable pour laquelle les effectifs varient donc beaucoup entre les deux analyses. Cet argument explique la réduction marquée d'écart-type pour l'âge, essentiellement due à une élévation artificielle de la valeur de l'écart-type en analyse cas-complet. Inversement, la variable genotype 3 induit la majorité des données manquantes en analyse cas-complet, et la valeur de son écart-type varie peu entre les deux analyses.

Tableau 3.5 – Résultats comparés de l'analyse multivariée (modèle final à 5 variables), analyse cas-complet et imputation multiple

	Cas Complet (N*=2465)			Imputation Multiple (N*=4343)		
	ORa (IC 95%)	SE	(SE/OR)100† (%)	ORa (IC 95%)	SE	(SE/OR)100† (%)
Sexe	1.93 (1.24 - 3.01)	0.44	22.5	1.94 (1.38 - 2.73)	0.34	17.4
Age	2.44 (1.67 - 3.57)	0.47	19.4	2.28 (1.70 - 3.06)	0.34	15.0
Durée de l'infection par le VHC	2.83 (1.82 - 4.41)	0.64	22.6	2.65 (1.84 - 3.81)	0.49	18.5
Consommation d'alcool	2.54 (1.85 - 3.48)	0.41	16.2	2.78 (2.14 - 3.60)	0.37	13.3
Génotype 3	1.50 (1.11 - 2.02)	0.23	15.2	1.62 (1.23 - 2.12)	0.22	13.7

* Nombre d'individus pris en compte dans le calcul des OR du modèle final
† Coefficient de variation associé à l'ORa

Ces résultats sont issus d'une imputation multiple utilisant un modèle simple, c'est à dire n'incluant pas de variables auxiliaires. Des modèles d'imputation plus complexes, incluant des variables auxiliaires et/ou des termes d'interaction (par exemple entre âge et durée de la maladie), n'ont pas apporté de changements dans les estimations. Comme pour toute procédure d'imputation, l'hypothèse MAR doit être discutée, ce qui sera fait dans la deuxième partie de cette étude.

- ***Discussion épidémiologique***

Les données ont été collectées pour un sous-échantillon d'usagers de drogues au sein d'un système de surveillance. Le type d'étude correspond donc à une enquête transversale et il est délicat d'identifier des facteurs de risque de complications hépatiques graves au moyen d'une analyse étiologique. Cependant, bien que les variables d'exposition aient été renseignées en même temps que la variable d'intérêt (complications hépatiques graves), la plupart de ces variables reflètent des expositions antérieures à l'issue de la maladie hépatique.

Les résultats de l'analyse étiologique après imputation multiple montrent que les variables suivantes sont significativement associées à des complications hépatiques graves chez les patients nouvellement diagnostiqués : sexe masculin, âge >40 ans, durée de l'infection ≥18ans, historique de consommation excessive d'alcool, co-infection par le VIH et infection par le VHC de génotype 3. A l'exclusion de la variable génotype 3, tous ces facteurs ont déjà été identifiés comme facteurs associés avec une fibrose hépatique [106].

De nombreuses études ont montré qu'une consommation excessive passée ou actuelle d'alcool était associée à une évolution plus rapide de la fibrose hépatique chez les patients porteurs du VHC [103;106]. Par ailleurs, l'infection par le VIH est connue pour modifier l'histoire naturelle de la maladie en accélérant la progression de la fibrose hépatique, que ce soit avant ou depuis l'avènement des traitements antirétroviraux [107]. Une étude a montré une association entre une infection par un virus de l'hépatite C de génotype 3 et une fibrose hépatique avancée chez des patients co-infectés par le VIH [108]. Par ailleurs, le portage du génotype 3, qui est le génotype de l'hépatite C le plus répandu parmi les usagers de drogue (>30%), a pu être lié à une stéatose hépatique et a également émergé comme un co-facteur de risque de progression de fibrose hépatique [109].

Ces trois facteurs de risque ont déjà été identifiés parmi les usagers de drogues. Ils doivent être pris en compte de façon à optimiser la prise en charge des patients ayant été soumis à ces expositions. L'identification du portage d'un virus de génotype 3 comme facteur de risque de fibrose hépatique est récente [109], et a été renforcée par ce travail.

- ***Hypothèse MAR***

La validité des résultats de l'imputation multiple repose, lors de l'utilisation de logiciels standards, sur l'hypothèse que le mécanisme de données manquantes est de type MAR.

La co-infection par le VIH et l'infection par le VHC de génotype 3 ont été identifiés comme facteurs de risque de fibrose hépatique après estimation des données manquantes par imputation multiple. La consommation excessive d'alcool est un facteur de risque bien établi de fibrose, mais sa déclaration peut être considérée comme sensible.

Même si les résultats des analyses cas-complet et imputation multiple sont proches, ce qui peut laisser supposer que les deux analyses sont valides (car non-biaisées), il parait important de tester la robustesse des résultats pour ces trois variables clés. De ce fait, une analyse de sensibilité par pondération, basée sur une approche proposée par Carpenter et al., a été appliquée à ces trois facteurs de risque.

2. Analyse de sensibilité par pondération

2.1. Contexte

L'imputation multiple, telle qu'elle est implémentée dans les logiciels standards d'analyse statistique [38], permet d'estimer des paramètres sous l'hypothèse que les données sont manquantes au hasard (MAR), impliquant qu'elles dépendent seulement des données observées des variables [3].

L'analyse cas-complet, appliquée par défaut par les logiciels d'analyses statistiques, restreint l'analyse aux individus pour lesquels les variables incluses dans l'analyse sont entièrement renseignées. Dans le cas d'analyses étiologiques, une approche cas-complet induit une perte de puissance, mais peut donner des estimations non-biaisées si le mécanisme de données

manquantes ne dépend pas de la variable à expliquer [110]. Dans le cas contraire, une analyse cas-complet donne des résultats systématiquement biaisés, même si le mécanisme de données manquantes est MAR. Par contre, une approche par imputation multiple permettra d'obtenir des estimations valides sous l'hypothèse MAR, si le modèle d'imputation est correct, et ce même si le mécanisme de données manquantes dépend de la variable à expliquer [3].

Les données manquantes peuvent également être dues à un mécanisme MNAR et il n'est pas possible en pratique de différencier statistiquement ces deux mécanismes [60]. Si les données manquantes sont MNAR, alors l'imputation multiple réalisée par le biais de procédures classiques donnera des estimations biaisées. L'importance des biais observées sous ce mécanisme est a priori proportionnelle à la dépendance entre le mécanisme de données manquantes et la variable à expliquer [111].

Dans de nombreuses analyses, le mécanisme de données manquantes d'au moins une covariable peut dépendre de la variable à expliquer et il est alors préférable d'appliquer une approche par imputation multiple. Cependant, le mécanisme de non-réponse peut dans ce cas dépendre également de données non-observées de certaines covariables, et les estimations des coefficients peuvent y être sensibles, et ce d'autant plus si des variables auxiliaires ne peuvent pas être incluses dans le modèle d'imputation afin de rendre l'hypothèse MAR plus plausible [14]. Il est donc important de tenir compte de l'impact d'un mécanisme MNAR potentiel sur les estimations lors de l'interprétation des résultats.

Dans la littérature biostatistique, deux types d'approches ont été proposées pour traiter des données manquantes selon un mécanisme MNAR : les modèles par sélection [112] et les modèles par mélange [113]. Ces approches nécessitent de faire des hypothèses fortes et invérifiables pour modéliser le mécanisme de données manquantes, même si cette incertitude peut être réduite en incluant dans les modèles des variables liées au mécanisme de données manquantes. Des procédures permettant d'adapter ces modèles pour réaliser des analyses de sensibilité ont été proposées [114;115]. Cependant, ces méthodes ne sont pas implémentées dans des logiciels statistiques standards et restent globalement d'une application complexe. Ceci peut expliquer qu'elles soient peu utilisées en pratique courante. Ainsi, d'après une revue de Sterne et al. [52], une analyse de sensibilité a été réalisée dans seulement 1 des 59 applications d'imputation multiple publiées dans des journaux majeurs entre 2002 et 2007.

La méthode présentée dans ce travail permet de réaliser rapidement une analyse de sensibilité après une imputation multiple classique sous l'hypothèse MAR. Cette approche plus directe est dérivée des modèles par sélection et a été proposée par Carpenter et al. [68;116]. Elle permet d'explorer la robustesse des estimations à un non-respect "local" de l'hypothèse MAR. Cela signifie que la sensibilité des estimations au non-respect de l'hypothèse MAR peut être estimée à partir des données observées, sans qu'il soit nécessaire de modéliser la non-réponse [117]. Les paramètres issus de l'analyse de bases de données imputées sous l'hypothèse MAR sont pondérés de façon à représenter la distribution des données imputées selon un mécanisme MNAR. De cette façon, les estimations obtenues sous les hypothèses MAR et MNAR peuvent être comparées, ce qui permet de tester la robustesse des estimations à un non-respect "local" de l'hypothèse MAR.

Cette méthode est attractive dans la mesure où elle est relativement facile à implémenter après imputation multiple. Elle n'a pas, à notre connaissance, été appliquée à des données épidémiologiques de type observationnel. Nous avons donc mis en œuvre cette méthode d'analyse de sensibilité à partir des données de surveillance de l'hépatite virale C en France, et proposé des règles d'application pratiques.

2.2. Arguments épidémiologiques pour le choix des variables

Nous avons choisi d'appliquer cette méthode aux données présentées en première partie de ce chapitre. L'analyse initiale visait à identifier les facteurs de risque de complications hépatiques graves dans une population d'usagers de drogues infectés par le VHC et nouvellement pris en charge au sein des pôles de références (période 2001-2007).

Les variables prédictrices incluses dans le modèle d'imputation ont été limitées à la variable à expliquer et aux 9 variables retenues après l'analyse univariée. Puisque le modèle d'imputation n'a pas pu être enrichi de variables auxiliaires qui auraient pu rendre l'hypothèse MAR plus plausible, une analyse de sensibilité a été envisagée. Parmi les facteurs de risque identifiés au terme de l'analyse après imputation multiple, trois variables ont été retenues pour l'analyse de sensibilité, sur la base d'hypothèses épidémiologiques sur leur mécanisme de données manquantes respectif : une consommation excessive passée d'alcool, une co-infection par le VIH et le portage d'un virus de l'hépatite C de génotype 3.

- ***Consommation excessive passée d'alcool***

Le fait de déclarer une consommation excessive d'alcool peut être ressentie comme sensible socialement, plus particulièrement dans une population de patients atteints par une hépatite, et ce même s'il s'agit d'une consommation d'alcool ancienne. Nous avons donc fait l'hypothèse que d'anciens gros consommateurs d'alcool étaient moins susceptibles que les autres de déclarer cette consommation.

- ***Co-infection par le VIH***

Le statut de co-infection par le VIH a pu être attesté soit par une sérologie antérieure, lorsque celle-ci était rapportée dans le dossier médical du patient, soit par un test effectué lors de la première présentation dans un pôle de référence. Il faut noter qu'en France, les patients co-infectés VHC-VIH sont généralement référés à un service hospitalier de maladies infectieuses. De ce fait, la prévalence de la co-infection VHC-VIH est d'environ 8% dans les pôles de référence, ce qui peut inciter les hépatologues à considérer leurs patients comme mono-infectés VHC quand aucune sérologie VIH antérieure n'est signalée. Nous avons donc fait l'hypothèse que l'information sur un test VIH est moins souvent rapportée pour des patients négatifs pour le VIH.

- ***Infection par un virus de l'hépatite C de génotype 3***

Le génotypage des souches virales a été réalisé par les laboratoires de virologie des pôles de référence. La demande de génotypage dépend du médecin prescripteur, sachant que le génotypage n'était pas réalisé en routine au début de la période de surveillance (2001-2003). Par la suite, la connaissance d'un impact du génotype sur la prise en charge du patient a conduit à un génotypage plus systématique. On observe d'ailleurs que la proportion de données manquantes de la variable genotype 3 diminue très nettement au cours du temps. Le génotype peut aussi ne pas être renseigné du fait du mode transversal de recueil des données, par exemple lorsque le test est réalisé à l'extérieur du pôle ou après le recueil de données. Il est cependant peu probable que le mécanisme de données manquantes dépende du génotype lui-même, puisque celui-ci est identifié par un processus de séquençage après amplification virale, indépendant du génotype lui-même. Nous avons cependant exploré l'hypothèse MAR pour cette variable du fait de sa proportion élevée de données manquantes et de l'importance de l'identification du génotype 3 comme facteur de risque de complications hépatiques graves.

2.3. Méthode d'analyse de sensibilité par pondération

Considérons une variable (covariable ou variable à expliquer) *Y* contenant des données manquantes. On définit par Y_i la valeur de *Y* pour l'individu *i*. Soit R_i une variable indicatrice de données manquantes égale à 1 si Y_i est observée et 0 sinon. Nous supposons un modèle logistique liant la probabilité d'observer Y à la valeur de *Y*, ajustée sur un vecteur *X* de covariables :

$$\text{logit}\,\Pr(R_i = 1) = \alpha + \beta X_i + \delta Y_i. \qquad (1)$$

Sous cette hypothèse de modèle paramétrique, si $\delta = 0$ et conditionnellement à l'ensemble des données observées, le mécanisme à l'origine des données manquantes de *Y* ne dépend pas de *Y*, de sorte que les données manquantes sont MAR. Inversement, si $\delta \neq 0$, le mécanisme de données manquantes dépend des données manquantes de *Y*, même en prenant en compte l'information contenue dans les données observées, et les données sont alors MNAR.

En pratique, la régression logistique ci-dessus ne peut pas être exécutée puisque, par définition, on ne peut pas observer Y_i quand $R_i = 0$. Cela implique qu'il faut choisir une valeur pour δ, puis explorer son impact sur les estimations obtenues à partir du modèle d'analyse. Pour la méthode que nous explorons, cela peut être fait en générant des poids qui sont fonction de δ et des données imputées. Nous donnons par la suite une explication intuitive de cette approche.

Supposons que *M* bases de données sont générées par une méthode d'imputation multiple sous l'hypothèse MAR. Pour chaque base de données, notons $\hat{\theta}_m$ l'estimation du paramètre d'intérêt (par exemple un coefficient de régression).

L'imputation multiple sous l'hypothèse MAR permet d'obtenir plusieurs estimations qui sont ensuite combinées selon les règles de Rubin pour obtenir une inférence finale $\hat{\theta}_{MAR}$ exprimée comme suit :

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

avec sa variance estimée $\hat{V}_{MAR}\left(\hat{\theta}_{MAR}\right) = \hat{V}_W\left(\hat{\theta}_{MAR}\right) + \left(1 + 1/M\right) \times \hat{V}_B\left(\hat{\theta}_{MAR}\right),$

où $\hat{V}_W\left(\hat{\theta}_{MAR}\right)=\frac{1}{M}\sum_{m=1}^{M}\hat{\sigma}_m^2$ et $\hat{V}_B\left(\hat{\theta}_{MAR}\right)=\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\theta}_m-\hat{\theta}_{MAR}\right)^2$.

L'approche de Carpenter consiste à remplacer cette moyenne simple par une moyenne pondérée, pour laquelle les estimations issues de l'imputation qui sont les plus susceptibles d'être affectées par un mécanisme MNAR vont être surpondérées par rapport aux autres. Une explication intuitive du processus de pondération est donnée dans le paragraphe suivant. Selon le modèle de régression logistique expliquant la probabilité de données manquantes décrit dans (1), Carpenter et al. montrent que les poids prennent une forme particulièrement simple [116].

Le modèle (1) fait l'hypothèse que, après ajustement sur les autres variables observées, la probabilité d'observer Y lorsque Y varie d'une unité est égale au log-odds ratio δ. Ainsi le poids, noté $\tilde{w}_m(\delta)$ pour l'imputation $m,(m=1,...,M)$, est égal à $\exp[-\delta\sum_{i\in I_Y}^{n_1}Y_i^m]$, où $Y_i^m\ (i\in I_Y)$ correspond à la valeur imputée de Y pour l'individu i dans la base de données m, et I_Y est l'ensemble des individus pour lesquels Y est manquant. La forme exponentielle des poids est dérivée de la fonction de lien logistique dans l'équation (1).

Les poids normalisés calculés pour chaque base de données imputées sont exprimés par :

$$w_m(\delta)=\frac{\tilde{w}_m(\delta)}{\sum_{k=1}^{M}\tilde{w}_k(\delta)}.$$

L'estimation MNAR de θ est alors définie comme :

$$\hat{\theta}_{MNAR}(\delta)=\sum_{m=1}^{M}w_m(\delta)\times\hat{\theta}_m\ ,$$

avec la variance estimée suivante :

$$\hat{V}_{MNAR}\left(\hat{\theta}_{MNAR}(\delta)\right)\approx\hat{V}_W\left(\hat{\theta}_{MNAR}(\delta)\right)+\left(1+1/M\right)\times\hat{V}_B\left(\hat{\theta}_{MNAR}(\delta)\right),$$

où $\hat{V}_W\left(\hat{\theta}_{MNAR}(\delta)\right)=\sum_{m=1}^{M}w_m(\delta)\times\hat{\sigma}_m^2$ et $\hat{V}_B\left(\hat{\theta}_{MNAR}(\delta)\right)=\sum_{m=1}^{M}w_m(\delta)\times\left(\hat{\theta}_m-\hat{\theta}_{MNAR}(\delta)\right)^2$.

Notons que si les données sont MAR, alors $\delta = 0$ et toutes les imputations sont pondérées de façon équivalente comme dans les règles de Rubin.

Pour donner une explication intuitive de ces poids, supposons que *Y* prenne seulement des valeurs positives et que δ soit positif, de façon à ce que la probabilité d'observer *Y* soit plus importante pour les valeurs positives de *Y* les plus élevées. Alors, les bases pour lesquelles la somme des valeurs de *Y* est petite seront sous-représentées parmi l'ensemble des bases imputées sous l'hypothèse MAR. Les poids permettent de corriger cela en surpondérant les estimations issues de bases ayant une somme des valeurs imputées de *Y* petite (par rapport aux autres bases de données imputées).

Nous présentons ci-dessous les estimations MAR pour les données de la base de surveillance de l'hépatite C. Nous explorerons la robustesse de ces estimations à un mécanisme MNAR lorsque δ s'éloigne de 0 et nous proposerons des règles d'application pratiques pour sélectionner une valeur de δ.

2.4. Application pratique de l'analyse de sensibilité par pondération

2.4.1. Problématique

Les analyses préliminaires sont présentées dans la première partie de ce chapitre. Elles ont porté sur 30 bases de données imputées avec un modèle incluant les 6 variables incomplètes, 3 variables complètes et la variable à expliquer.

Afin d'appliquer la méthode de Carpenter, il est recommandé d'imputer plus de 50 bases de données. Nous avons fait le choix d'imputer 1000 bases afin d'illustrer au mieux les particularités de la méthode. Le programme additionnel ICE de STATA a été utilisé pour l'imputation, en incluant les même variables que lors des analyses préliminaires, mais c'est le logiciel R et la librairie MICE qui ont permis de réaliser les analyses et les graphes afin de gérer les problèmes de calcul dus à la taille du fichier imputé.

Il est parfois possible de sélectionner la valeur du paramètre de sensibilité δ à partir de l'avis d'experts, aptes à proposer des valeurs plausibles du paramètre de sensibilité δ dans leur domaine de compétence. Peu d'informations sont disponibles en épidémiologie d'observation et

l'alternative consiste alors à explorer une étendue de valeurs de δ en tenant compte des hypothèses sur le mécanisme de données manquantes telles que présentées précédemment.

Nous proposons une approche combinant 4 étapes pour choisir une valeur appropriée de δ, illustrée en utilisant la variable infection par un virus de l'hépatite C de génotype 3, avant de l'appliquer aux deux autres variables consommation excessive d'alcool et co-infection par le VIH. Nous avons choisi la variable génotype 3 car elle permet de soulever les principales questions pratiques inhérentes à cette méthode.

Notre objectif est d'appliquer l'analyse de sensibilité au paramètre d'intérêt, c'est-à-dire le coefficient (odds ratio ajusté) de la variable génotype 3 dans la régression logistique multivariée après imputation multiple présentée dans le Tableau 3.4 (variable à expliquer : complications hépatiques graves). Selon les notations présentées précédemment, $\hat{\theta}_m$ est l'estimation du coefficient de régression logistique associée à la variable génotype 3 pour la base de données imputée $m, (m = 1...., M)$.

2.4.2. Processus de sélection du paramètre de sensibilité delta

- ***Etape 1 : Régression logistique explorant le mécanisme de données manquantes***

Cette étape consiste à générer une variable indicatrice de réponse pour la variable incomplète, puis à déterminer par une régression logistique multivariée les associations avec, d'une part la variable à expliquer, et d'autre part les autres covariables.

La variable génotype 3 étant utilisée pour l'illustration, on génère une variable indicatrice de données manquantes égale à 1 si la variable génotype 3 est observée et 0 sinon (notons que le codage retenu ici est celui utilisé par Carpenter et al. dans l'article princeps [116], inverse par rapport au reste de ce travail). Puis, en utilisant les données imputées pour toutes les variables à l'exclusion de la variable génotype 3, on ajuste un modèle de régression logistique multivariée pour expliquer l'indicatrice de données manquantes de la variable genotype 3. On inclut dans ce modèle la variable à expliquer (complications hépatiques graves) ainsi que toutes les covariables retenues dans le modèle d'analyse initial, à l'exclusion de la variable génotype 3. Les résultats sont présentés dans le Tableau 3.6 et montrent que le mécanisme de données manquantes de la

variable génotype 3 dépend de l'âge et de la durée de la maladie, mais pas de la variable à expliquer.

Tableau 3.6 – Régression multivariée expliquant l'indicatrice de données manquantes de génotype 3 à partir des covariables

Indicatrice de données manquantes de génotype 3	Coefficients de régression	SE	p
Complications hépatiques graves	**-0.05**	**0.13**	**0.72**
Age	0.17	0.09	0.06
Sexe	0.04	0.08	0.63
Durée de l'infection par le VHC	0.19	0.09	0.04
Délai de prise en charge	0.05	0.08	0.53
Consommation d'alcool	-0.005	0.07	0.94
Co-infection par le VIH	0.02	0.14	0.90
Co-infection par le VHB	-0.15	0.23	0.52

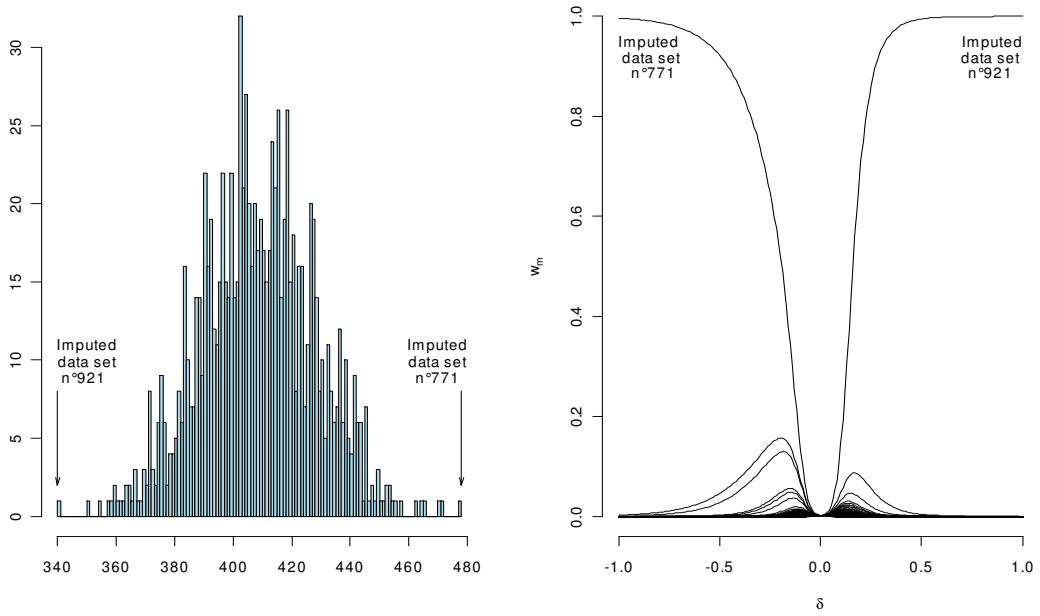
- ***Etape 2 : Détermination graphique d'une valeur de delta***

La justification théorique de la méthode repose sur le principe de l'échantillonnage préférentiel (importance sampling) [118]. Dans les applications d'échantillonnage préférentiel, il est recommandé de ne pas mettre tout le poids sur une ou très peu de valeurs. De ce fait, il est nécessaire de réduire l'étendue des valeurs prises par δ. Nous recommandons d'appliquer les critères suivants : les valeurs de δ doivent être choisies de façon à ce que la valeur du poids normalisé maximal se situe autour de 0.5 et qu'au moins 5 poids normalisés soient supérieurs à 1/M (c'est-à-dire la valeur des poids lorsque $\delta = 0$) [68]. Ainsi, l'estimateur MNAR obtenu est déduit d'informations issues d'au moins 5 bases de données imputées, ce qui correspond au nombre minimum usuellement recommandé en pratique.

Nous recommandons de représenter cette information sur un graphique tel que la Figure 3.2. La partie gauche de la Figure représente (pour chacune des 1000 bases imputées) l'histogramme de la somme des valeurs imputées de *Y* pour la variable génotype 3. Les valeurs extrêmes sont 340 pour la base imputée n°921 et 480 pour la base n°771. Sur la partie droite de la Figure sont représentés les poids normalisés en fonction de la valeur de δ, pour chacune des $M = 1000$ bases de données imputées. Le poids normalisé maximal correspond à la base de données pour laquelle la somme de valeurs imputées de *Y* est minimale (base n°921) quand $\delta > 0$ ou maximale (base

n°771) quand $\delta<0$. Lorsque $\delta=0$, le poids normalisé est égal à $1/M$ car tous les $\tilde{w}_m(0)$ sont égaux à 1.

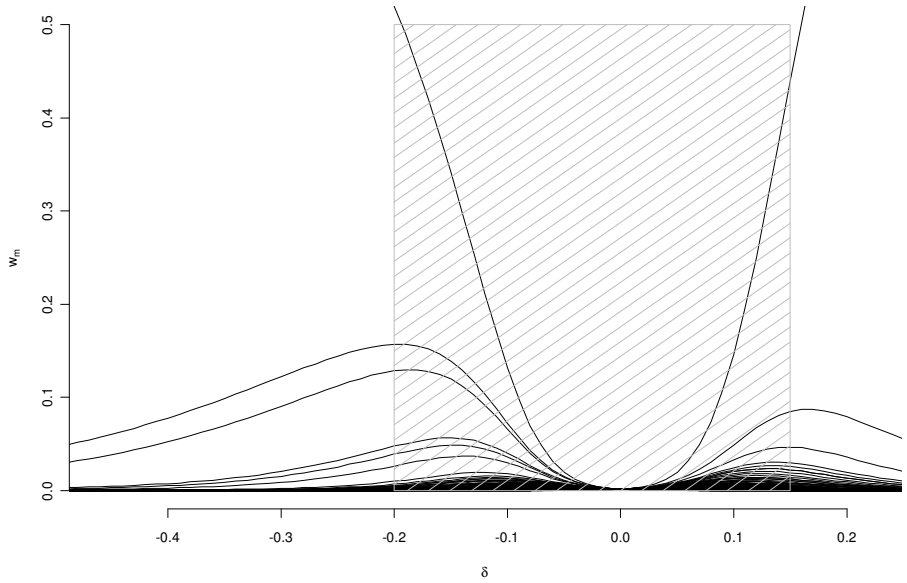
Figure 3.2 – Détermination graphique d'une valeur de δ pour la variable génotype 3



Partie gauche : histogramme de la somme des valeurs imputées de génotype 3 pour chaque base de données imputée parmi M=1000 bases.
Partie droite : poids normalisés (w_m) selon la valeur de δ pour chaque base de données imputée.

La Figure 3.3 montre la partie centrale de la Figure 3.2. En suivant les recommandations citées précédemment, nous retenons des valeurs positives de δ correspondant à un poids normalisé d'environ 0.5, ce qui donne un intervalle pour δ de [-0.2;0.15]. Même pour les valeurs extrêmes de cet intervalle, on note que plus de 5 poids normalisés sont supérieurs à 0.001. Notons que la partie centrale de la zone hachurée correspond à des valeurs de δ induisant des écarts à l'hypothèse MAR trop réduits pour être utilisés en pratique, puisque tous les poids normalisés décroissent vers $1/M$.

Figure 3.3 – Poids normalisés (w_m) selon δ pour chaque base de données imputée



La zone hachurée délimite les valeurs de δ correspondant à des poids normalisés égaux à 0.5 au maximum.

- ***Etape 3 : Choix du signe de delta***

Nous déterminons lors de cette étape si la valeur de δ retenue correspond à la borne haute ou basse de l'intervalle identifié au cours de l'Etape 2.

Pour le génotype 3 de l'hépatite C, l'équation (1) montre la relation entre le signe de δ et le mécanisme de données manquantes supposé : pour des valeurs positives de δ, la probabilité d'observer le génotype augmente si l'individu porteur de l'hépatite C est infecté par une souche virale de génotype 3, et inversement pour des valeurs négatives de δ. Dans notre exemple et tenant compte des résultats de l'Etape 2 (Figure 3.3), nous sélectionnons $\delta = 0.15$. Cela signifie en pratique qu'il est 1.2 (exp(0.15)) fois plus fréquent d'observer des données manquantes pour le génotype 3 parmi les individus infectés par une souche virale de génotype 3 que parmi ceux infectés par des souches virales d'autres génotypes.

Dans le cas particulier de cette variable, l'expérience ne permet pas de suggérer fortement une valeur positive ou négative pour δ, et les résultats sont présentés pour les deux valeurs.

- ***Etape 4 : Diagnostic graphique***

La méthode de pondération est applicable seulement pour une analyse de sensibilité 'locale', ce qui signifie que les distributions du paramètre d'intérêt obtenues sous les hypothèses MAR et MNAR doivent se chevaucher partiellement, même si elles ont des valeurs moyennes différentes. Ceci ne sera généralement pas vrai pour une analyse de sensibilité "non locale". Pour vérifier si cette condition est respectée pour la valeur de δ choisie, nous proposons de représenter (i) la distribution des poids normalisés w_m en fonction de $\hat{\theta}_m$, $m = 1,...,M$, et (ii) la distribution de l'estimateur MNAR en fonction d'un nombre croissant d'imputation. Pour ces deux représentations, si la méthode donne des résultats valides, la distribution de l'estimateur MNAR doit être contenue dans la distribution des $\hat{\theta}_m$ obtenus par imputation multiple sous l'hypothèse MAR, sans être positionnée aux extrêmes de la distribution MAR.

Pour la variable génotype 3, les résultats de cet examen graphique sont présentés sur la Figure 3.4. La partie gauche de la Figure représente, pour $\delta = 0.15$, les poids normalisés selon $\hat{\theta}_m$ pour chaque base de données imputées (sachant que $\hat{\theta}_m$ est l'estimation du coefficient de régression obtenu sous l'hypothèse MAR pour chaque base de données imputée). La partie droite de la figure représente l'estimateur MNAR, calculé pour *n* imputations en fonction du nombre de bases de données imputées noté $n, (n = 10,...,M)$, et défini comme :

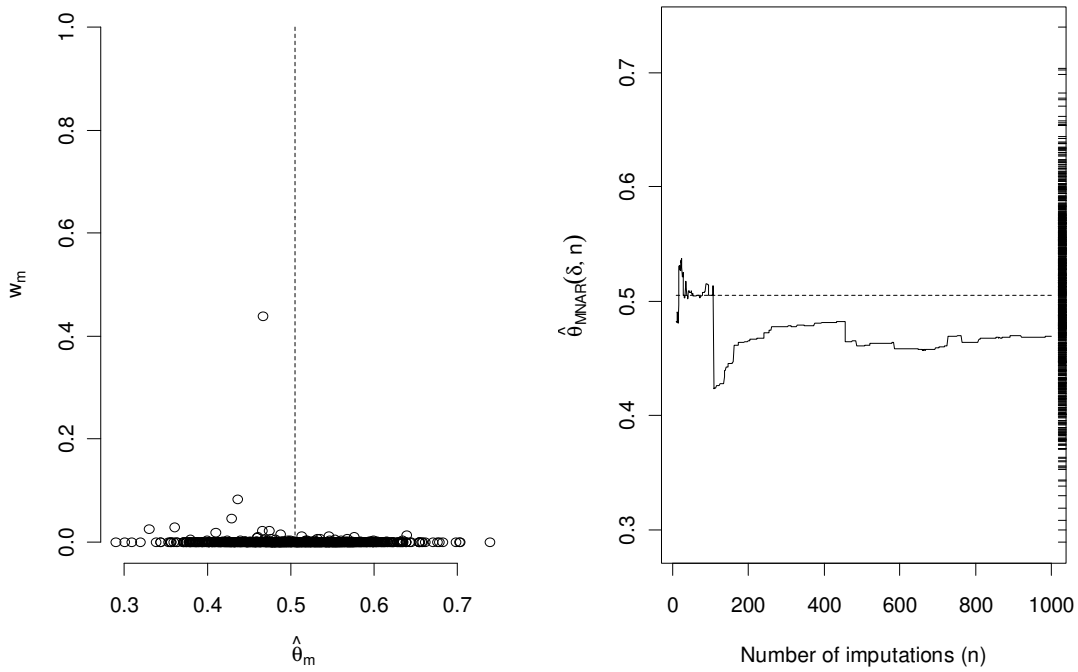
$$\hat{\theta}_{MNAR}(\delta, n) = \frac{\sum_{m=1}^{n} w_m(\delta) \times \hat{\theta}_m}{\sum_{m=1}^{n} w_m(\delta)}.$$

Par exemple :

$$\hat{\theta}_{NMAR}(\delta,1) = w_1(\delta) \times \hat{\theta}_1 / w_1(\delta), \hat{\theta}_{NMAR}(\delta,2) = \left(w_1(\delta) \times \hat{\theta}_1 + w_2(\delta) \times \hat{\theta}_2\right) / \left(w_1(\delta) + w_2(\delta)\right)$$

et ainsi de suite.

Figure 3.4 – Analyse de la variable genotype 3 pour δ=0.15



Partie gauche : poids normalisés (w_m) selon $\hat{\theta}_m$ (coefficient de régression estimé de genotype 3 pour la base de données imputées *m*).

Partie droite : estimateur pondéré, calculé comme la moyenne mobile des $\hat{\theta}_{MNAR}$ selon le nombre de bases de données imputées. A l'extrémité droite du graphe est représentée la distribution des 1000 estimations de $\hat{\theta}_m$, une pour chaque base de données imputée.

Pour les deux graphes, la ligne pointillée représente $\hat{\theta}_{MAR}$ (moyenne des $\hat{\theta}_m$ sur les 1000 bases de données imputées).

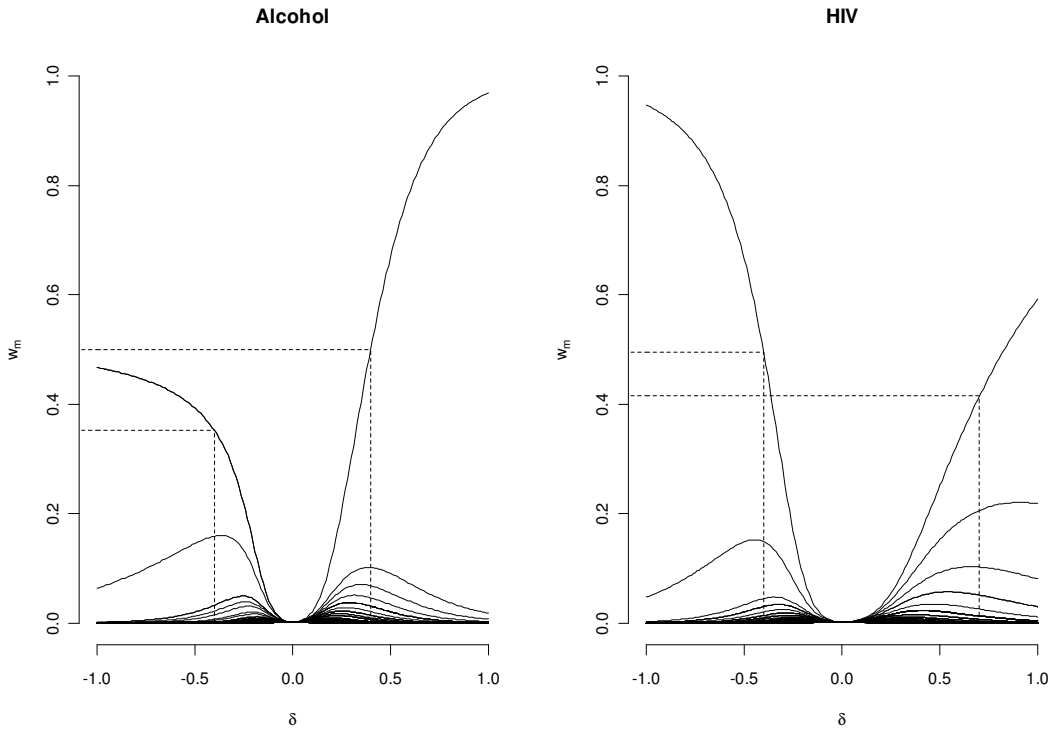
Sur le graphe nous observons que (i) l'estimateur MNAR paraît se stabiliser lorsque le nombre d'imputations augmente et que (ii) la distribution de l'estimateur MNAR ne se situe pas dans les valeurs extrêmes de la distribution MAR (représentée par les traits sur la partie droite du graphe).

2.5. Résultats

Les résultats des analyses cas-complet et imputation multiple sous l'hypothèse MAR sont présentés dans le Tableau 3.7. Nous présentons également dans ce tableau les résultats de l'analyse de sensibilité pour les trois variables suivantes : génotype 3 du VHC, sérologie VIH et historique de consommation excessive d'alcool.

Pour le génotype 3 du VHC, nous utilisons la valeur de δ telle qu'elle a été sélectionnée selon la procédure en 4 étapes proposée. Nous appliquons la même approche pour les variables sérologie du VIH et consommation d'alcool. La Figure 3.5 représente les graphes de l'Etape 2 pour ces deux variables.

Figure 3.5 – Poids normalisés (w_m) selon δ pour chaque base de données imputée, pour les variables consommation d'alcool et statut sérologique VIH



L'intervalle retenu pour δ est [-0.4;0.4] pour l'alcool (partie gauche) et [-0.4;0.7] pour le VIH (partie droite).

Pour la variable consommation d'alcool, l'Etape 1 montre que la probabilité d'avoir des observations pour cette variable dépend de la variable à expliquer (complications hépatiques graves). L'Etape 2 permet d'identifier une étendue de valeurs de δ de [-0.4;0.4] (partie gauche de la Figure 3.5). Tenant compte des résultats de l'Etape 1 et des hypothèses épidémiologiques formulées précédemment, nous supposons que la probabilité d'observer la consommation d'alcool est plus faible si cette consommation est excessive et nous choisissons $\delta = -0.4$. L'interprétation de ce choix est que, après ajustement sur les autres covariables, la probabilité d'observer l'historique de consommation d'alcool est réduite d'un facteur $0.7 = \exp(-0.4)$ pour les individus ayant un historique de consommation d'alcool excessive.

Pour la variable co-infection par le VIH, l'Etape 1 montre que la probabilité d'avoir des observations pour cette variable dépend de la variable à expliquer (complications hépatiques graves). L'Etape 2 permet d'identifier une étendue de valeurs pour δ de [-0.4;0.7] (partie droite de la Figure 3.5). Selon les résultats de l'Etape 1, et sachant que, dans un contexte similaire, la probabilité d'observer une co-infection par le VIH est plus élevée pour les individus positifs pour le VIH, nous retenons $\delta = 0.7$. L'interprétation de cette valeur est que, après ajustement sur les autres covariables, la probabilité d'observer une sérologie VIH est augmentée d'un facteur $2.0 = \exp(0.7)$ pour les individus positifs pour le VIH.

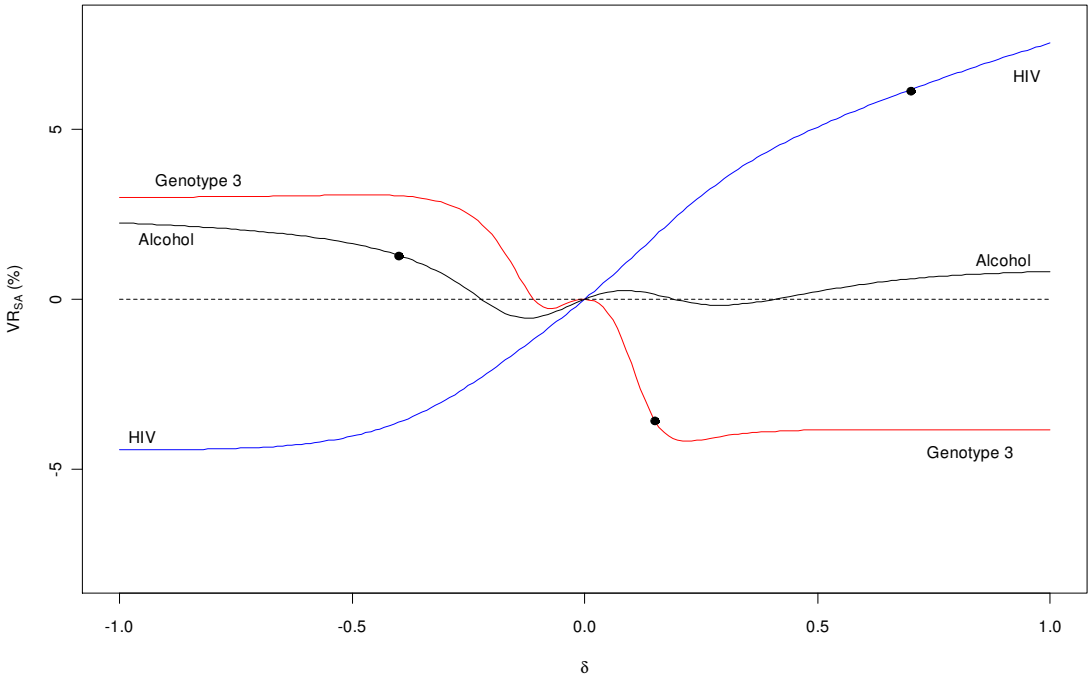
Pour ces trois variables, on peut considérer le diagnostic exposé en Etape 4 comme acceptable. Les odds ratios ajustés sont présentés dans le Tableau 3.7. Notons que le même modèle multivarié incluant les variables sexe, âge, durée de l'infection par le VHC, consommation excessive d'alcool, génotype du VHC et co-infection par le VIH a été appliqué pour chaque analyse (cas-complet, imputation multiple sous MAR, analyse de sensibilité). L'analyse de sensibilité a été réalisée pour chacune des 3 variables tour à tour.

Deux critères ont été retenus pour interpréter les odds ratios ajustés dans chacune des trois analyses :

(1) Le coefficient de variation (CV) de l'odds ratio qui donne sa mesure de dispersion normalisée. Pour les trois variables, il est nettement réduit après imputation et reste stable après pondération.

(2) Le taux de variation (TV) permet d'estimer la variation relative entre l' $\hat{OR}_{MNAR}$ et l' $\hat{OR}_{MAR}$, et est défini par $TV_{AS} = 100 \times (\hat{OR}_{MNAR} - \hat{OR}_{MAR}) / \hat{OR}_{MAR}$. De la même façon, nous avons défini un taux de variation (TV_{IM}) qui rend compte de la variation relative des odds ratios ajustés obtenus lors des analyses cas-complet et imputation multiple sous l'hypothèse MAR. Le TV_{IM} varie de 9.7% pour le genotype 3 à 15.5% pour la co-infection par le VIH et 22% pour la consommation d'alcool. Le TV_{AS} est calculé pour la valeur de δ choisie pour chaque variable. Sa valeur est relativement faible pour la consommation d'alcool (1.3%) et pour le génotype 3 (3.5%) mais augmente pour la co-infection par le VIH (6.6%). La Figure 3.6 montre que le TV_{AS} est relativement stable pour des valeurs de δ variant de [-1;1] pour les variables consommation d'alcool et génotype 3, mais augmente pour la co-infection par le VIH pour des valeurs de δ supérieures à celle retenue.

Figure 3.6 - Taux de variation selon δ après analyse de sensibilité (TV_{SA}) pour les variables génotype 3, consommation d'alcool et co-infection par le VIH



Les points noirs correspondent au TV_{SA} calculé pour la valeur de δ retenue pour chaque variable (génotype 3 : δ=0.15 ; alcool : δ=-0.4 ; VIH : δ=0.7)

Tableau 3.7 - Analyse multivariée de type cas complet, imputation multiple et analyse de sensibilité, pour M = 1000 bases de données imputées

Variables	Cas Complet (CC)				Imputation multiple (IM)				Analyse de sensibilité (AS)				
	Données Manquantes (%)	ORa (IC 95%)	SE	CV* (%)	ORa (IC 95%)	SE	CV* (%)	TV_{IM}† (IM versus CC) (%)	δ	ORa (IC 95%)	SE	CV* (%)	TV_{AS}‡ (AS versus IM) (%)
Consommation d'alcool	11.1	2.32 (1.66 - 3.23)	0.39	17	2.82 (2.18 - 3.66)	0.37	13	21.86	-0.40	2.86 (2.21 - 3.70)	0.37	13	1.29
Génotype 3	26.3	1.51 (1.10 - 2.07)	0.24	16	1.66 (1.27 - 2.16)	0.23	14	9.70	0.15	1.60 (1.23 - 2.06)	0.21	13	3.56
Co-infection par le VIH	16.3	1.56 (0.92 - 2.62)	0.41	27	1.80 (1.24 - 2.61)	0.34	19	15.52	0.70	1.91 (1.32 - 2.76)	0.36	19	6.12

* Coefficient de variation de l'aOR

† Taux de variation exprimant la variation relative entre l'ORa obtenu par analyse de sensibilité et l'ORa après imputation multiple

‡ Taux de variation exprimant la variation relative entre l'ORa obtenu après imputation multiple et l'ORa par analyse cas-complet.

2.6. Discussion

En présence de données manquantes, toutes les analyses et estimations correspondantes sont basées sur des hypothèses sur le mécanisme de données manquantes que l'on ne peut tester directement. Des analyses de sensibilité permettant d'explorer la robustesse des estimations selon les différentes hypothèses sont donc essentielles.

Ainsi, des approches ont été proposées afin d'élaborer des modèles prenant en compte des données MNAR selon diverses hypothèses sur les paramètres de modélisation, et de tester l'impact de ces paramètres sur les inférences clés. Van Buuren [2] propose d'adapter la méthode d'imputation par équations chaînées en incorporant dans le processus d'imputation un facteur d'ajustement δ, générant ainsi plusieurs modèles d'imputation sous l'hypothèse MNAR selon la valeur de δ retenue. Troxel et al. [117] proposent de créer un "index de sensibilité à un mécanisme MNAR" (index of sensitivity to non-ignorability) calculé à partir des données observées. Si l'impact sur les inférences est modéré, alors l'analyse est robuste à la spécification du mécanisme de données manquantes, et les résultats de l'imputation multiple sont valides.

La méthode présentée dans ce travail permet d'appliquer rapidement une analyse de sensibilité afin de tester la robustesse des estimations obtenues par imputation multiple sous l'hypothèse MAR. Cette approche consiste à surpondérer des imputations qui sont alors plus plausibles sous un mécanisme MNAR. En modélisant le mécanisme de données manquantes par une régression logistique, ces poids prennent alors une forme particulièrement simple. Même si l'analyse de sensibilité demeure à un niveau local, elle fournit cependant d'importantes informations sur l'impact des déviations de l'hypothèse MAR sur les estimations, tout en évitant la complexité d'une modélisation du mécanisme de données manquantes (full joint modeling). La pertinence de cette approche a été confirmée par d'autres études [68;116].

Nous avons donc développé et illustré la mise en application de cette approche, et proposé une procédure en 4 étapes afin de sélectionner une valeur pour le paramètre de sensibilité δ.

Pour la variable génotype 3, l'étape 1 de notre procédure montre que la probabilité d'observer le génotype ne semble pas liée à la variable à expliquer de notre modèle d'intérêt (complications hépatiques graves), après ajustement sur les autres covariables (Tableau 3.6). L'analyse de sensibilité permet à la probabilité d'observer le génotype de dépendre en plus de la valeur du

génotype. Le Tableau 3.7 et la Figure 3.6 montrent que les estimations ne sont pas sensibles à cette dépendance. Ainsi, on peut en déduire que cette dépendance MNAR ne modifie pas la relation entre la probabilité d'observer cette variable génotype 3 et la variable à expliquer.

En ce qui concerne la variable consommation d'alcool, nous avons posé l'hypothèse que les patients pouvaient être réticents à déclarer une consommation excessive passée car cette question pouvait être ressentie comme sensible socialement. Cependant, le mode de déclaration de la consommation d'alcool et sa relation avec la consommation réelle observée n'est pas si claire dans la littérature. Ainsi, Pernanen [119] a rapporté des taux de non-réponse plus élevés dans des populations incluant une forte proportion de gros consommateurs d'alcool. Van Oers [120] a observé une sous-déclaration de la consommation d'alcool plus importante chez les femmes que chez les hommes, mais pas spécifiquement parmi les gros consommateurs. Un pourcentage plus élevé de non-consommateurs et de consommateurs excessifs parmi des non-répondants a été rapporté par Knibbe [121], tandis que Lemmens [122] n'a pas mis en évidence une consommation plus importante parmi les non-répondants que parmi les répondants. Finalement, Lahaut [123] a observé des proportions de non-réponse importantes parmi les non-buveurs, et beaucoup moins élevées parmi les buveurs excessifs.

Le fait de rapporter une consommation d'alcool est fortement lié aux caractéristiques sociodémographiques des individus, caractéristiques qui peuvent être incluses dans le modèle d'imputation de façon à réduire le biais de non-réponse. Nous avons seulement pu inclure l'âge et le sexe dans le modèle d'imputation car les autres variables n'étaient pas liées au mécanisme de données manquantes. Notre Etape 1 montre que la probabilité d'observer la consommation d'alcool dépend de la variable à expliquer, après avoir pris en compte les autres covariables. L'analyse de sensibilité permet à cette probabilité de dépendre en plus de la valeur de la consommation d'alcool. Ainsi, on peut en déduire que cette dépendance MNAR ne modifie pas la relation entre la probabilité d'observer cette variable consommation d'alcool et la variable à expliquer.

Pour la variable co-infection avec le VIH, les hépatologues des pôles de référence ont probablement tendance à considérer leurs patients comme étant mono-infectés par le VHC car les patients co-infectés VHC-VIH sont habituellement référés dans les services hospitaliers de maladies infectieuses en France. Par ailleurs, les patients infectés par le VHC auraient dû être testés plus fréquemment pour le VIH depuis la conférence de consensus de 2002 qui spécifie un

traitement plus long pour les patients co-infectés [124]. Etonnamment, aucun effet de la période sur le mécanisme de données manquantes n'a été détecté puisque la proportion de données manquantes reste à peu près constante au cours du temps. De plus, lorsque le statut VIH a été mieux rapporté, la proportion de patients négatifs pour le VIH a augmenté. Nous avons donc fait l'hypothèse par défaut que le statut VIH était plus susceptible d'être observé si le statut sérologique était positif, d'où notre choix d'un signe positif pour δ. L'Etape 1 montre que la probabilité d'observer le statut VIH dépend de la variable à expliquer, après ajustement sur les autres covariables. L'analyse de sensibilité permet à cette probabilité de dépendre en plus du statut VIH. Le Tableau 3.7 et la Figure 3.6 montrent que les résultats sont sensibles à cela. On peut en déduire que, si le mécanisme de données manquantes est MNAR, avec une probabilité plus élevée que le statut VIH soit renseigné pour les patients positifs pour le VIH, alors l'association entre cette variable et la variable à expliquer est en réalité plus forte que ce qui est observé dans l'analyse réalisée sous l'hypothèse MAR.

Conclusion

On peut conclure de l'analyse de ces données que l'analyse de type cas-complet est potentiellement biaisée puisque les données suggèrent une dépendance entre la probabilité d'observer les valeurs et la variable à expliquer, après prise en compte des covariables. Une analyse réalisée par imputation multiple sous l'hypothèse MAR est donc préférable. Notre analyse de sensibilité montre que, pour des déviations locales de l'hypothèse MAR, les estimations pour les variables génotype 3 et consommation d'alcool sont peu affectées, alors que l'impact sur l'odds ratio associé à la variable co-infection par le VIH est sous-estimé si la probabilité d'observer le statut VIH est plus importante lorsque ce statut sérologique est positif.

CHAPITRE 4

PROCESSUS D'IMPUTATION MULTIPLE PERENNE : APPLICATION AU SYSTEME DE SURVEILLANCE DU VIH

Le traitement des données manquantes dans les bases de données de surveillance de santé publique est peu abordé dans la littérature récente [82;125;126]. Dans le domaine des maladies infectieuses à déclaration obligatoire et plus spécifiquement du VIH, peu de systèmes de surveillance font l'objet d'un traitement raisonné des données manquantes. Le système de surveillance du VIH mis en place aux Etats-Unis récolte des données qui sont incomplètes. Les données manquantes, traitées en routine par une méthode d'imputation simple [127], font l'objet d'un processus d'imputation multiple dans le cadre de l'estimation de l'incidence du VIH [128;129]. Il a été également proposé d'appliquer systématiquement une méthode d'imputation multiple aux données du VIH et du SIDA comme méthode alternative à l'imputation simple (risk factor redistribution) [130].

En Europe, des systèmes de surveillance du VIH ont été mis en place depuis les années 2000 avec la prise de conscience par les autorités de santé publique des limites des systèmes nationaux. Il a donc été proposé d'étendre la surveillance aux nouveaux diagnostics de l'infection au VIH (et plus seulement aux cas au stade sida), de surveiller les comportements à risque et de diversifier les systèmes de surveillance sentinelles. Au niveau européen, la surveillance du VIH se fait par le biais de systèmes hétérogènes, avec des variations importantes dans la qualité des données collectées. L'European Center for Disease Prevention and Control (ECDC) assure la centralisation de ces données. La France a mis en place un système de surveillance du VIH avec d'emblée un grand nombre de variables recueillies, ce qui explique la proportion non-négligeable de données manquantes. Au fil du temps, il est apparu préférable d'estimer ces données manquantes par une méthode adaptée. Nous présentons dans ce chapitre l'application de l'imputation multiple au système de déclaration obligatoire du VIH en France, c'est-à-dire les étapes de sa mise en œuvre ainsi que les phases de diagnostic et de validation.

1. Système de surveillance du VIH

1.1. Déclaration obligatoire des diagnostics d'infection à VIH

Les objectifs de la déclaration obligatoire (DO) de l'infection à VIH sont de connaître le nombre et les caractéristiques des personnes découvrant leur séropositivité au VIH, d'en suivre l'évolution au cours du temps [131], et de fournir des données permettant d'estimer le nombre de nouvelles contaminations, c'est-à-dire l'incidence du VIH [71].

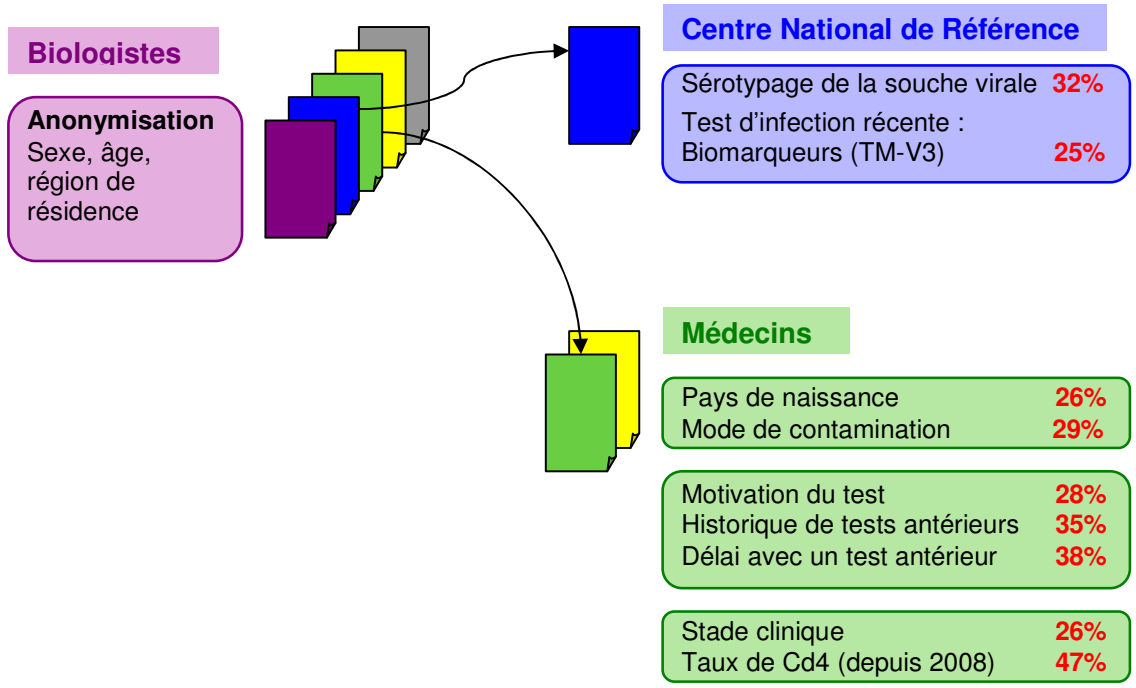
Mise en place en 2003 [132], cette notification est initiée par les biologistes qui doivent déclarer toute personne dont la sérologie VIH est confirmée positive, et cela pour la première fois dans le cadre de leur laboratoire. Un code d'anonymat unique est créé pour chaque individu à partir de la date de naissance, du prénom, de l'initiale du nom et du sexe. La fiche de notification complétée d'abord par le biologiste comporte cinq feuillets : le biologiste en conserve un et envoie les autres respectivement au médecin de l'Agence Régionale de Santé (ARS) (feuillet 1), au médecin prescripteur (feuillets 2 et 3), et au Centre National de Référence du VIH (feuillet 4) (Figure 4.1).

Le feuillet médical est complété par le médecin prescripteur du test qui renseigne les informations sociodémographiques et épidémiologiques telles que le pays de naissance, le mode de contamination ainsi que les raisons ayant motivé le test VIH. Le médecin doit également consigner l'historique des tests VIH antérieurs à la notification, qu'ils aient été positifs ou négatifs. Cet historique se décompose en deux questions sur le fait d'avoir eu ou non un test VIH antérieur, positif ou négatif. En cas de sérologie antérieure, la durée en mois entre cette sérologie antérieure et celle donnant lieu à la notification doit être précisée. Le médecin complète également la fiche par des informations cliniques telles que le stade clinique et, depuis 2008, le taux de lymphocytes T4 (CD4).

Les feuillets de déclaration (feuillets biologiques et médicaux) sont ensuite adressés aux médecins inspecteurs des Agences Régionales de Santé (ARS) qui doivent les coupler et les transmettre à l'InVS. Les doublons, définis comme plusieurs déclarations pour une même personne, sont alors détectés grâce au code d'anonymat et permettent le plus souvent de compléter la notification initiale.

Figure 4.1 – Représentation schématique du système de déclaration obligatoire du VIH

La proportion moyenne de données manquantes est indiquée en % pour les variables incomplètes



1.2. Surveillance virologique

La surveillance virologique réalisée par le Centre National de Référence (CNR) du VIH a été initiée en même temps que la notification obligatoire en 2003 [133]. Les biologistes envoient volontairement un échantillon de sérum au CNR du VIH. Cet échantillon est déposé sur buvard à partir du "fond de tube" ayant permis de diagnostiquer l'infection à VIH. L'objectif de la surveillance virologique est double : la réalisation d'un test d'infection récente et le sérotypage de la souche virale.

Le test d'infection récente permet d'estimer la part des contaminations récentes, c'est-à-dire datant de moins de 6 mois en moyenne, parmi les découvertes de séropositivité. Ce test EIA-RI (Enzyme Immuno-Assay Recent Infection) a été initialement développé afin de détecter les infections récentes par le biais d'un algorithme combinant les mesures par densité optique de deux marqueurs spécifiques [134]. Ces paramètres biologiques sont des anticorps anti-VIH définis par leur liaison avec d'une part l'épitope immunodominant (IDE ou TM) de la protéine gp41 et d'autre part la région V3 de la protéine gp120.

Le sérotypage permet de suivre l'évolution des sous-types viraux circulant actuellement en France. Des tests sérologiques permettent la détermination des types viraux, ainsi que des groupes puis des sous-types. Les échantillons sont d'abord classés en type VIH-1 ou VIH-2, puis pour les types VIH-1 en groupe M ou O, et enfin parmi les groupes M en sous-type B ou non-B. Sachant que le VIH-1 est très largement majoritaire, la proportion des sous-types B versus les sous-types non-B est informative car elle permet d'évaluer la dynamique de l'épidémie en termes de migrations [135].

Les résultats de la surveillance virologique sont transmis à l'InVS, où ils sont couplés aux informations de la déclaration obligatoire grâce au code d'anonymat.

1.3. Objectifs du processus d'imputation multiple

L'objectif du processus d'imputation multiple de la base de déclaration obligatoire du VIH est de fournir à l'équipe de surveillance une base de données complète utilisée pour (i) réaliser les analyses descriptives permettant de caractériser les personnes nouvellement contaminées par le VIH et (ii) estimer l'incidence du VIH.

Il s'agit donc d'un processus d'imputation pérenne appliqué à une base de surveillance importante (>40000 cas), effectué sur une période définie (2003-2010) et en vue d'analyses spécifiques. Les variables qui sont retenues pour ce processus d'imputation doivent donc être identifiées au préalable par l'équipe de surveillance. Puis, à l'issue de l'imputation, la base de données produite est mise à disposition de l'équipe pour analyse. Elle sera utilisée comme base de travail pendant un an afin de produire les diverses communications sur le VIH au niveau national et international.

La base de déclaration obligatoire du VIH est imputée annuellement depuis 2008. Nous présenterons dans ce travail les différentes étapes du processus d'imputation effectué à partir de la base de données du 31 décembre 2009.

2. Examen de la base de données

Parmi les variables construites à partir des données issues de la notification obligatoire et de la surveillance virologique du VIH, 13 variables sont retenues pour l'imputation.

2.1. Examen quantitatif

Les variables incomplètes sélectionnées sont 3 variables binaires, 5 variables catégorielles et 5 variables continues. Parmi ces 13 variables incomplètes, 12 contiennent une proportion de données manquantes variant de 23 à 44% (Tableau 4.1).

L'examen du motif de répartition des données manquantes montre qu'elles suivent un motif aléatoire. Une illustration est donnée dans le Tableau 4.2 pour 4 variables catégorielles à partir d'une commande dédiée à l'examen des données manquantes sous STATA (misschk). Ce tableau montre que les 4 variables contiennent 17% de données manquantes communes, mais que les autres combinaisons de données manquantes apparaissent selon des fréquences faibles. On peut également déduire du Tableau 4.2 que 18028 individus parmi l'effectif total de 41049 individus ont des données manquantes pour au moins une des 4 variables considérées. Une analyse cas-complet incluant les 4 variables catégorielles porterait donc seulement sur 23021 des 41049 individus soit sur seulement 56.08 % des observations.

Tableau 4.1 – Description des 13 variables incomplètes traitées par imputation

Variables	Type	Codage	Données manquantes (%)
Données socio-démographiques			
Pays de naissance	Catégorielle	France, Afrique sub-saharienne (AFSS), Europe, Amérique/Haïti, Autres	26.20
Mode de contamination	Catégorielle	Homo/bisexuel, Usage de drogue intraveineuse (UDI), Hétérosexuel, Autres	29.01
Historique de dépistage			
Motif de dépistage	Catégorielle	Symptômes cliniques et biologiques (SCB), Exposition (EXP), Bilan/Grossesse, Dépistage orienté (DOR), Prise en charge (PEC)/Autres	27.25
Sérologie négative antérieure	Binaire	Non, Oui	26.20
Durée entre sérologie négative antérieure et sérologie de la notification	Continue	Durée en mois	44.90
Sérologie positive antérieure	Binaire	Non, Oui	35.81
Durée entre sérologie positive antérieure et sérologie de la notification	Continue	Durée en mois	33.93
Données cliniques			
Stade clinique	Catégorielle	Asymptomatique (ASY), Primo-infection virale (PIV), Symptomatique non-Sida (SNS), Sida	25.73
Taux de lymphocytes T4	Continue	Nombre de CD4 par mm3	43.36
Surveillance virologique			
Type viral	Catégorielle	Type 1, Type 2, Type 1 et 2	1.50
Sérotype viral	Binaire	Sérotype B, Sérotype non-B	33.93
Biomarqueurs V3 et Tm	Continue	Mesure par densité optique	34.14

Tableau 4.2 – Examen du motif de répartition des données manquantes

Variables examinées pour leurs données manquantes		
Variables	Données manquantes (Effectifs)	Données manquantes (%)
1 - Pays de naissance	10740	26.2
2 - Mode de contamination	11907	29.0
3 - Stade clinique	10560	25.7
4 - Motif de dépistage	11187	27.3

Données manquantes pour quelles variables ?			
Variables	Fréquence	%	% cumulé
1234	7117	17.34	17.34
123_	348	0.85	18.19
12_4	229	0.56	18.74
12__	616	1.5	20.24
1_34	210	0.51	20.76
1_3_	292	0.71	21.47
1__4	524	1.28	22.74
1___	1404	3.42	26.16
_234	194	0.47	26.64
_23_	395	0.96	27.6
_2_4	357	0.87	28.47
_2__	2651	6.46	34.93
__34	869	2.12	37.04
__3_	1135	2.76	39.81
___4	1687	4.11	43.92
____	23021	56.08	100
Total	41049	100	

Données manquantes pour combien de variables ?			
Nombre de variables incomplètes	Fréquence	%	% cumulé
0	23021	56.08	56.08
1	6877	16.75	72.83
2	3053	7.44	80.27
3	981	2.39	82.66
4	7117	17.34	100.00
Total	41049	100.00	

2.2. Examen qualitatif

Afin de réaliser cet examen qualitatif, on génère des indicatrices binaires de données manquantes pour chaque variable incomplète. Ces indicatrices permettent d'identifier le mécanisme de données manquantes pour chaque variable, par une régression multivariée expliquant l'indicatrice de données manquantes de chaque variable incomplète en fonction des covariables identifiées dans la base de DO.

Notons que, dans ce cas particulier d'imputation de données de surveillance, les analyses réalisées après imputation sont essentiellement descriptives. Il n'est donc pas nécessaire de distinguer une variable à expliquer et des variables explicatives. L'examen du mécanisme de données manquantes de chaque variable incomplète fournit des informations utiles pour l'interprétation des résultats des comparaisons entre données observées et imputées lors de l'étape de diagnostic. Par exemple, le mécanisme de données manquantes de la variable mode de contamination dépend de la variable pays de naissance. On peut donc s'attendre à une distribution différentielle des données imputées de la variable mode de contamination selon les différentes modalités de la variable pays de naissance.

Les hypothèses sur les mécanismes causant des données manquantes varient selon les variables recueillies. Ainsi, les données issues de la surveillance virologique sont essentiellement manquantes si le biologiste ne souhaite pas participer à cette surveillance (40% des données manquantes), ou bien pour des raisons techniques, si l'échantillon est inexploitable ou que l'infection est à un stade trop récent pour que le sérotype soit identifiable. Le médecin prescripteur renseigne un volet clinique à partir de l'examen clinique et d'examens complémentaires (stade clinique, nombre de CD4), mais rapporte également des informations impliquant un entretien avec le patient, au cours duquel le patient peut décider de répondre ou non. Alors que les questions concernant l'historique de tests sérologiques peuvent être considérées comme anodines, d'autres questions portant sur les raisons ayant motivé le test VIH ou sur le mode probable de contamination sont plus sensibles et le patient peut faire le choix de ne pas y répondre.

On peut donc proposer les hypothèses épidémiologiques suivantes pour le mécanisme de données manquantes des variables citées : (i) les données des variables de la surveillance virologique sont de type MAR ou MCAR, (ii) les données des variables cliniques et déclaratives sont de type MAR et (iii) les données déclaratives sensibles sont à risque d'être MNAR. L'examen des mécanismes de données manquantes à partir des variables indicatrices de données manquantes permet d'exclure un mécanisme MCAR, puisque des liens existent entre indicatrices et covariables pour toutes les variables incomplètes examinées. Des mécanismes MAR dépendant de certaines variables sont identifiés, mais cet examen ne permet pas d'exclure un mécanisme MNAR puisque le mécanisme de données manquantes de toutes les variables dépend de variables incomplètes, et donc potentiellement des valeurs non-observées de ces variables.

3. Processus d'imputation en deux phases

3.1. Historique de sérologies VIH

Cet historique se décompose en deux questions binaires sur l'existence ou non d'un test VIH antérieur, positif ou négatif, à partir desquelles sont créées deux variables binaires, sérologie antérieure positive et sérologie antérieure négative. En cas de sérologie antérieure connue, positive ou négative, une durée entre cette sérologie et celle dont découle la notification peut être déduite des informations contenues dans le dossier médical ou déclarées par le patient. Une terminologie simplifiée sera adoptée dans la suite de ce travail pour faire référence aux deux variables binaires sur une sérologie antérieure, serpo (sérologie antérieure positive) et serneg (sérologie antérieure négative) ainsi que pour les variables quantitatives indiquant les durées en mois entre une sérologie antérieure positive ou négative et la sérologie de la notification, délais positifs et délais négatifs.

Ces 4 variables sont incomplètes et sont déterminantes pour les analyses des données de la DO du VIH pour les raisons suivantes :

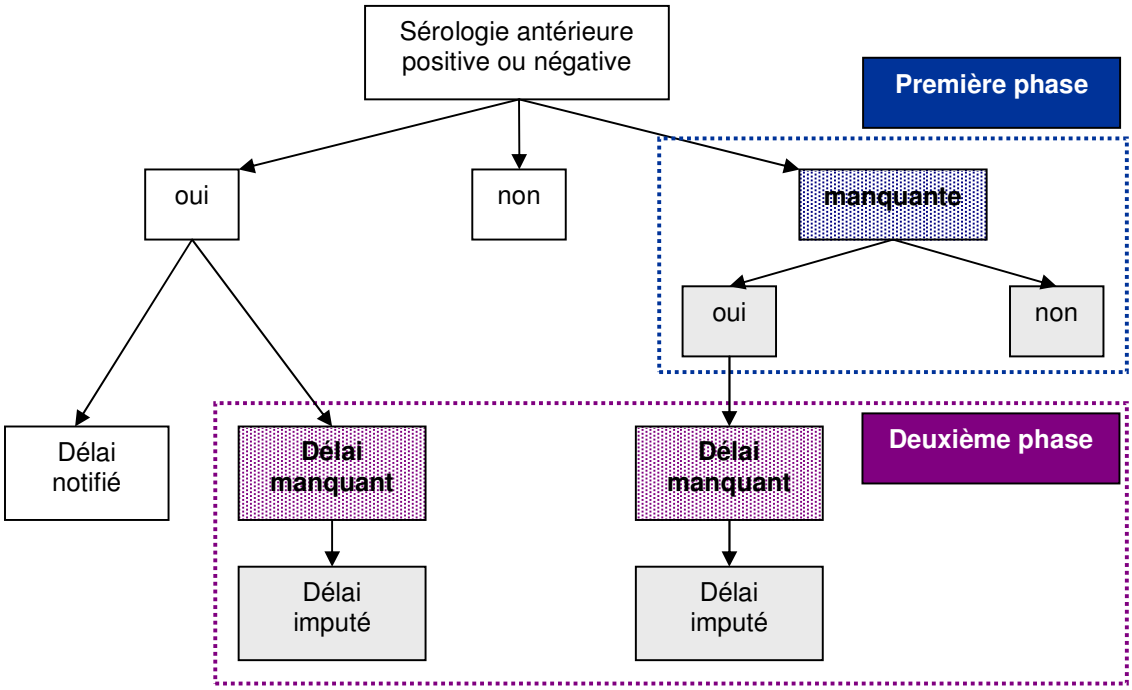
- Les analyses descriptives de la DO ont pour but d'établir le nombre annuel de nouveaux diagnostics pour le VIH, ainsi que de décrire les caractéristiques de cette population. On parle de nouveau diagnostic ou de découverte de séropositivité pour un individu lorsqu'aucune sérologie positive antérieure n'a été rapportée ou bien qu'elle date de plus de 11 mois.

- L'estimation de l'incidence du VIH prend en compte la proportion de patients ayant eu une sérologie antérieure négative ainsi que le délai associé.

3.2. Imputation conditionnelle

Du fait du lien entre les variables binaires rapportant une sérologie antérieure, positive ou négative, et les délais correspondants, il est nécessaire d'estimer les données manquantes des variables binaires avant d'estimer celles des délais positifs ou négatifs. De ce fait, comme illustré par la Figure 4.2, l'imputation doit être réalisée en deux temps.

Figure 4.2 – Imputation en deux phases



En effet, les variables délais positifs et négatifs dépendent de deux variables filtres, serpo et serneg, et ne peuvent donc être renseignées que pour la modalité "oui" de ces variables. Cependant ces variables filtres sont incomplètes. Il est donc nécessaire d'estimer d'abord les données manquantes des variables serpo et serneg afin d'identifier tous les individus avec une réponse positive. Cette estimation doit donc avoir lieu lors d'une première phase d'imputation.

Puis, les variables délais positifs et négatifs peuvent être complétées conditionnellement aux variables serpo et serneg, c'est-à-dire seulement pour les réponses positives pour ces variables, que celles-ci soient observées dans la base de départ ou imputées au cours de la première phase. Les variables de délais sont donc complétées au cours d'une deuxième phase d'imputation.

Nous verrons dans la suite de ce chapitre que cette particularité impacte tout le processus d'estimation des données manquantes puisque deux modèles d'imputation doivent être élaborés.

4. Etapes préliminaires

4.1. Base de données de départ

La base de données de déclaration obligatoire contient 44067 notifications depuis l'année 1982 et 116 variables issues de la saisie ou reconstruites. Tenant compte du début de la notification obligatoire, la base de données retenue pour toutes les analyses de 2011 a été restreinte à la période du 01 juin 2003 au 31 décembre 2010 et inclut 41049 notifications. Nous n'avons pas inclus le premier semestre 2003 car cette période correspond au début de la déclaration obligatoire et que les données ne nous ont pas semblé suffisamment fiables.

Les variables retenues pour le processus d'imputation sont sélectionnées en fonction des analyses programmées par l'équipe de surveillance. Du fait de la méthode d'imputation retenue (imputation multiple par équations chaînées), les variables peuvent être conservées sous leur forme originelle (binaire, catégorielle ou continue). Un recodage des variables catégorielles est cependant parfois nécessaire afin de limiter le nombre de modalités à 5 et d'équilibrer au mieux les effectifs [56].

La base de données contient donc les variables d'origine incomplètes, les variables recodées pour l'imputation et leurs indicatrices de données manquantes. Compte tenu de la taille de la base de données, nous avons créé une base restreinte, en écartant les variables contenant une proportion de données manquantes trop élevée (de 60 à 85%). Cette base finale est stabilisée à environ 100 variables.

4.2. Recodages liés à des problématiques spécifiques

4.2.1. Biomarqueurs V3 et TM

Les valeurs des biomarqueurs biologiques du test d'infection récente sont d'une interprétation délicate pour les patients au stade sida. En effet, en relation avec l'histoire naturelle de la maladie, les valeurs de ces biomarqueurs peuvent être artificiellement basses pour ces patients, ce qui pourrait amener à les classer à tort en infectés récents. De ce fait, les valeurs observées des

biomarqueurs sont recodées en données manquantes pour les patients au stade sida et le processus d'imputation prend en compte cette singularité.

4.2.2. Sérologie antérieure négative

Pour les fiches de DO utilisées pendant la période juin 2003-juin 2007, les questions concernant l'existence d'une sérologie antérieure, positive ou négative, ont fait l'objet d'une formulation ambiguë : "Le médecin a-t-il eu connaissance d'une sérologie négative antérieure ?". Les réponses possibles étaient oui, non ou inconnu. Au vu de la proportion élevée de réponses négatives, nous avons émis l'hypothèse que le médecin pouvait cocher la modalité non au lieu d'inconnu s'il n'avait pas connaissance de cette information. La formulation de la question était identique sur le feuillet du biologiste. De ce fait, une certaine proportion de réponses pourrait être renseignée à tort par le biologiste ou/et le médecin prescripteur en réponses négatives. Ce processus serait plus marqué pour la variable serneg du fait de la structure de la fiche (la réponse sur le feuillet biologiste est autocopiée sur le feuillet médical).

Nous avons supposé que ce problème de formulation avait induit un déséquilibre entre les réponses positives et négatives de serneg (27% de réponses positives en 2003). Cette hypothèse a été renforcée par une comparaison avec des données externes issues d'enquêtes ou de cohortes (60% de réponses positives pour serneg dans Copana, une cohorte de patients séropositifs nouvellement diagnostiqués [136]). Par ailleurs, ce déséquilibre en faveur des réponses négatives a évolué au cours du temps et s'est atténué avec l'introduction en juillet 2007 d'une nouvelle version de la fiche de notification contenant une formulation plus claire des deux questions. La proportion de réponses positives passe ainsi pour serneg de 24% en 2003 à 65% en 2010. La proportion des modalités oui-non de la variable serneg impacte le calcul de l'incidence. Puisqu'un retour à la réponse d'origine n'est pas possible, nous avons fait le choix d'élaborer un filtre afin de sélectionner les réponses "valides" de serneg.

La première étape consiste à sélectionner des variables contenant des données que l'on considère sensibles et dont on pense qu'elles sont recueillies si l'entretien mené par le médecin prescripteur est approfondi. L'hypothèse de départ est qu'un remplissage correct de serneg est lié à la qualité de l'entretien médical, évaluée par le degré de remplissage global du questionnaire. Ne sont retenues à l'issue de cette étape que les variables contenant une proportion de données manquantes inférieure à 50% (8 variables catégorielles).

La deuxième étape consiste à construire des modèles prédictifs de la variable serneg en utilisant l'aire sous la courbe ROC. Ces analyses sont effectuées sur une sous-partie de la base de notification correspondant à la période 2008-2010, c'est-à-dire la période au cours de laquelle les nouvelles fiches ont été introduites. Ces nouvelles fiches sont identifiables par le recueil d'une nouvelle variable, le taux de CD4. La stratégie d'analyse consiste tout d'abord à sélectionner, parmi les 8 modèles bivariés prédictifs de serneg, le modèle avec la meilleure aire sous la courbe ROC. Puis, à partir du meilleur modèle, on construit 7 autres modèles en incluant chacune des covariables restantes. L'aire sous la courbe ROC n'est cependant pas améliorée par l'ajout de ces variables. Le modèle final retenu inclut donc les variables catégorielles mode de contamination et motivation pour le test, avec une aire sous la courbe ROC de 0,74.

On sélectionne ensuite sur l'ensemble de la base de données un sous-échantillon de fiches de notification pour lequel ces 2 variables sont entièrement renseignées. Pour les fiches qui ne font pas partie de ce sous-échantillon, les valeurs observées de serneg sont recodées en données manquantes, quelles que soient ces modalités. La proportion de données manquantes de serneg passe ainsi de 26% à 61%, et la répartition des réponses oui-non passe de 27%-73% à 40%-60% sur l'ensemble de la période.

La variable sérologie positive antérieure (serpo) est sujette au même problème de remplissage, mais le déséquilibre des réponses oui-non est moins sensible et il a été décidé de ne pas le traiter.

5. Construction des modèles d'imputation

L'imputation de la base de notification est réalisée de façon pérenne, c'est-à-dire que le processus d'imputation est réactualisé chaque année. Nous présentons la construction du modèle d'imputation pour chacune des deux phases d'imputation.

5.1. Constructions des équations de prédiction

5.1.1. Contexte méthodologique

Les variables retenues pour le processus d'imputation suivent des distributions différentes (binaire, catégorielle et continue) et la méthode d'imputation par équations chaînées permet de

conserver la forme originelle de ces variables. Une fonction de lien est spécifiée pour chaque variable à imputer. L'essentiel de la construction des équations de prédiction réside donc dans la sélection des variables prédictrices pour chacune des deux phases. Rappelons que les variables retenues pour les analyses ultérieures de la base de données sont systématiquement incluses dans le modèle d'imputation (variables principales) et que des variables supplémentaires peuvent être identifiées pour améliorer les capacités prédictrices du modèle d'imputation (variables auxiliaires).

Chaque variable incomplète est modélisée par défaut dans la commande ICE à partir de toutes les covariables incluses dans le modèle d'imputation. Cependant, il est possible de sélectionner un jeu de prédicteurs différent pour chaque variable à imputer. Cette sélection repose en partie sur l'examen des relations entre chaque indicatrice de données manquantes et toutes les covariables disponibles. Il est recommandé de réaliser cet examen par régression logistique multivariée de chaque indicatrice de données manquantes sur l'ensemble des covariables. Une variable est dite prédictrice si le test de Wald du coefficient de régression associé à cette variable est significatif ($p \leq 0.05$).

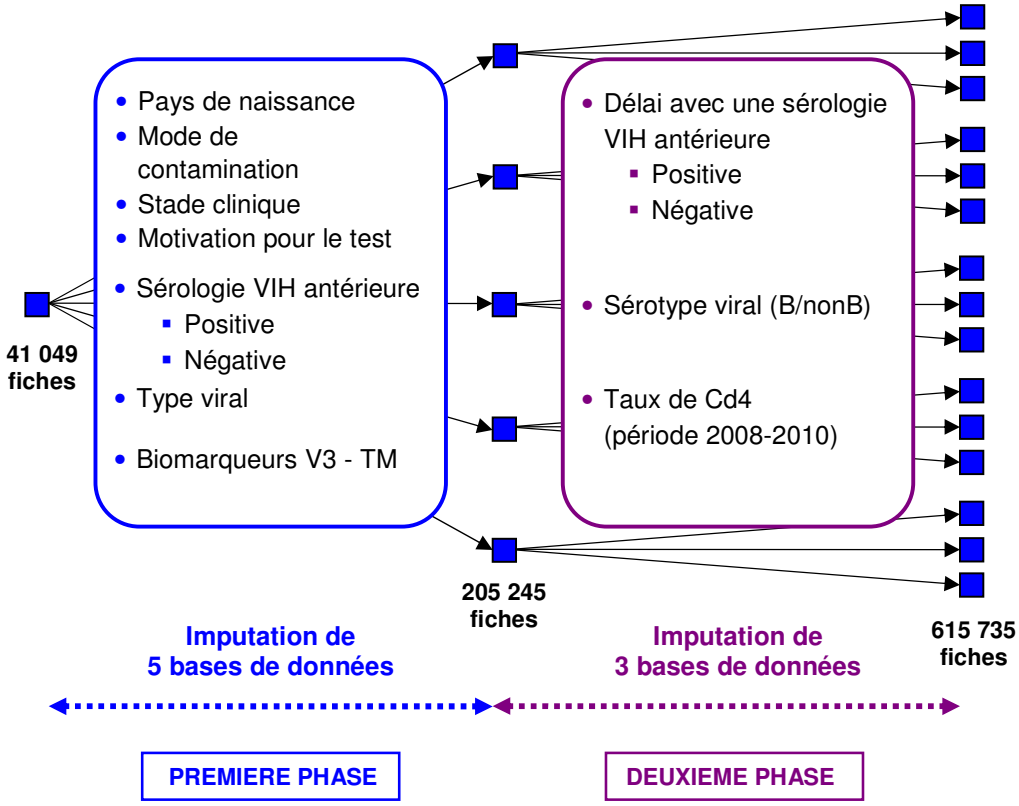
Comme décrit dans le chapitre 1, certains auteurs proposent de sélectionner les variables prédictrices en se basant sur les données imputées de façon à éviter le biais de sélection dû à la perte d'effectifs en analyse cas complet [28;82]. Nous avons appliqué cette stratégie pour évaluer les relations entre variables à imputer et variables prédictrices.

5.1.2. Première phase

- ***Variables principales***

Les 9 variables imputées lors de la première phase sont 5 variables catégorielles (pays de naissance, mode de contamination, stade clinique, motivation pour le test et type viral), 2 variables binaires (sérologie antérieure positive ou négative) et 2 variables continues (biomarqueurs TM et V3) (Figure 4.3). Ces 9 variables constituent les variables principales et sont donc incluses d'emblée dans le modèle d'imputation.

Figure 4.3 – Description du processus d'imputation en deux phases



Le modèle d'imputation doit être le plus général possible, et nous avons fait le choix dans un premier temps d'inclure chacune des 9 variables dans les équations de prédiction des autres variables, en faisant l'hypothèse que leurs capacités prédictrices étaient correctes. Dans un deuxième temps, nous avons exclu deux variables de toutes les équations de prédiction. La variable serneg a été recodée et contient 61% de données manquantes. Il nous a donc paru préférable de l'exclure des équations de prédiction. La variable type viral est presque complète avec 1.4% de données manquantes mais contient 3 modalités très déséquilibrées avec 98% de sous-type 1, 1.7% de sous-type 2 et 0.8% de sous-type 1 et 2. De ce fait, elle induit une instabilité du modèle d'imputation et a été exclue des équations de prédictions.

Par ailleurs, les biomarqueurs TM et V3 sont imputés conditionnellement au stade clinique observé de chaque patient. En effet, les valeurs de ces marqueurs ne doivent pas être retenues pour les individus au stade sida et les mesures disponibles sont donc recodées en données manquantes. Une variable binaire spécifiant le statut sida est utilisée pour limiter l'imputation des biomarqueurs aux patients non-sida.

- ***Variables auxiliaires***

Du fait du mode de recueil des informations dans le domaine de la surveillance, la proportion de données manquantes est souvent élevé et il peut donc s'avérer difficile d'identifier des variables auxiliaires suffisamment bien renseignées. Dans le cas de la notification du VIH, nous avons retenu comme variables auxiliaires potentielles 2 variables binaires, le sexe et le type de déclarant (médecin de ville, médecin hospitalier), 2 variables catégorielles, l'année de diagnostic (2003-2010) et l'inter-région (6 catégories), et 1 variable continue, l'âge. Ces 5 variables auxiliaires sont complètes.

Une première imputation a donc été réalisée en incluant les variables principales et ces 5 variables auxiliaires ; 5 bases ont été générées. L'examen des relations entre les variables indicatrices de données manquantes de 8 des 9 variables principales (on ne retient pas la variable type viral) et des 13 covariables (variables principales et variables auxiliaires potentielles) a été réalisé en appliquant un modèle de régression multivariée à l'ensemble des données imputées, par les commandes spécifiques de l'imputation. Les résultats, présentés dans le Tableau 4.3, montrent que les variables année de diagnostic, inter-région et type de déclarant sont prédictrices de toutes les variables incomplètes. Les variables sexe et âge sont respectivement liées à 4 et 2 indicatrices de données manquantes. Cependant, ces variables peuvent être utilisées lors d'analyses sur les données imputées et elles ont été retenues dans les équations de prédiction de toutes les variables.

Le Tableau 4.3 montre également que les 4 variables catégorielles (pays de naissance, mode de contamination, stade clinique et motif de dépistage), la variable sérologie positive antérieure et le biomarqueur TM ont des capacités prédictrices correctes (3 à 5 liens significatifs avec les 8 variables indicatrices). Le biomarqueur V3 n'est lié à aucune indicatrice, mais a été maintenu dans le modèle pour conserver son niveau de lien avec le biomarqueur TM. Après imputation des données manquantes, la variable sérologie négative antérieure n'est effectivement prédictrice d'aucune variable incomplète, ce qui justifie son retrait des équations de prédictions.

Le modèle de la première phase se compose donc de 14 variables, 9 variables principales et 5 variables auxiliaires.

Tableau 4.3 – Analyses multivariées des indicatrices de données manquantes à partir des données imputées au cours de la première phase (M=5)

	Rpnais	Rmcont	Rstclin	Rmotif	Rserneg	Rserpo	RV3	RTM
Pays de naissance		**$<10^{-3}$**	**$<10^{-3}$**	**$<10^{-3}$**	**$<10^{-3}$**	0.26	**$<10^{-3}$**	**$<10^{-3}$**
Mode de contamination	**$<10^{-3}$**		0.05	**0.02**	**$<10^{-3}$**	**0.01**	**$<10^{-3}$**	**$<10^{-3}$**
Stade clinique	0.39	**0.01**		**$<10^{-3}$**	**0.02**	0.18	0.14	0.14
Motif de dépistage	**$<10^{-3}$**	**$<10^{-3}$**	**0.02**		**0.01**	**$<10^{-3}$**	0.40	0.40
Sérologie négative	**0.04**	0.39	0.18	0.91		**0.03**	0.08	0.09
Sérologie positive	0.16	0.12	0.13	**$<10^{-3}$**	**$<10^{-3}$**		**0.01**	**0.01**
Biomarqueur V3	0.35	0.38	0.26	0.84	0.42	0.79		0.52
Biomarqueur TM	**0.04**	**0.02**	**0.09**	**$<10^{-3}$**	**$<10^{-3}$**	**0.02**	0.52	
Année de diagnostic	**$<10^{-3}$**	**$<10^{-3}$**	**$<10^{-3}$**	**$<10^{-3}$**	**$<10^{-3}$**	**$<10^{-3}$**	**$<10^{-3}$**	**$<10^{-3}$**
Inter-région	**$<10^{-3}$**	**$<10^{-3}$**	**$<10^{-3}$**	**$<10^{-3}$**	**$<10^{-3}$**	**$<10^{-3}$**	**$<10^{-3}$**	**$<10^{-3}$**
Sexe	**0.03**	**0.04**	**0.01**	0.19	0.30	**0.02**	0.30	0.30
Age	0.09	**$<10^{-3}$**	**0.04**	0.17	0.65	0.65	0.07	0.07
Type de déclarant	**$<10^{-3}$**	**$<10^{-3}$**	**$<10^{-3}$**	**$<10^{-3}$**	**$<10^{-3}$**	**$<10^{-3}$**	**$<10^{-3}$**	**$<10^{-3}$**

Les variables indicatrices de données manquantes sont les variables à expliquer de chaque modèle. Ces variables sont abréviées comme suit : Rpnais (pays de naissance), Rmcont (mode de contamination), Rstclin (stade clinique), Rmotif (motif de dépistage), Rserneg (sérologie négative), Rserpo (sérologie positive), RV3 (biomarqeur V3), RTM (biomarqueur TM). Les p-valeurs associées à chacune des variables explicatives sont présentées, avec en caractères gras les p-valeurs ≤0.05. Les zones grisées représentent les intersections entre chaque variable et son indicatrice de données manquantes.

5.1.3. Deuxième phase

- ***Variables principales***

Les variables principales imputées lors de la deuxième phase sont par construction imputées conditionnellement à des variables complétées lors de la première phase (Figure 4.3).

Il s'agit tout d'abord des variables continues indiquant la durée en mois entre la sérologie à l'origine de la notification et une éventuelle sérologie antérieure, positive ou négative, nommées délai positif et délai négatif. On spécifie dans les équations de prédiction qu'elles sont imputées conditionnellement aux variables binaires serpo et serneg, que les modalités de celles-ci soient imputées ou observées. Puisque la quantité d'individus avec une modalité oui pour les variables serpo et serneg augmente après imputation, la proportion de données manquantes des variables de délai s'accroit également : de 4.6% à 35.7% pour la variable délai positif et de 4.1% à 43.2% pour la variable délai négatif.

De la même façon, la variable sérotype viral (sérotype) est imputée au cours de cette phase conditionnellement à la variable type viral imputé à la première phase. La variable sérotype est reconstruite en deux modalités, sérotype B versus non-B, et imputée seulement lorsque le virus est de type 1.

Le taux de lymphocytes T4 (CD4), qui est une variable continue, a été ajoutée au modèle d'imputation de la deuxième phase à partir de 2009. Elle est recueillie par les nouvelles fiches de notification depuis 2008 seulement et elle est donc imputée conditionnellement à la période 2008-2010. Du fait de sa forte proportion de données manquantes (>50% en 2009 et 43% en 2010), nous avons fait le choix de l'inclure dans le modèle d'imputation de la deuxième phase car celui-ci est plus stable. Ce modèle ne contient en effet pas de variable catégorielle et bénéficie de l'apport d'informations issues des 9 variables renseignées lors de la première phase (et 15 bases de données sont imputées au total).

Ces 4 variables contiennent une proportion élevée de données manquantes (de 33 à 43%), et il est donc préférable de ne pas les retenir comme variables prédictrices. Chacune d'entre elles est donc exclue de l'équation de prédiction des autres variables.

- ***Variables auxiliaires***

Les variables auxiliaires potentielles peuvent être identifiées parmi des variables originellement complètes ou complétées lors de la première phase d'imputation. De ce fait, il parait cohérent de les sélectionner à partir de la base imputée issue de la première phase. Le Tableau 4.4 montre les résultats des régressions multivariées expliquant les indicatrices de données manquantes des 4 variables principales par les 11 covariables retenues.

Les résultats des régressions montrent que, en dehors de l'âge et des biomarqueurs, les variables retenues à partir de la première phase ont de bonnes capacités prédictrices. La variable âge est maintenue dans les équations de prédiction car elle est utilisée lors des analyses. Les biomarqueurs sont prédicteurs pour au moins la moitié des variables imputées et sont donc retenus.

Tableau 4.4 – Analyses multivariées des indicatrices de données manquantes à partir des données imputées au cours de la première phase (M=3)

	Rdelpos	Rdelneg	Rserotype	Rcd4
Pays de naissance	0.01	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$
Mode de contamination	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$
Stade clinique	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$
Motif de dépistage	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$
V3	0.02	$<10^{-3}$	$<10^{-3}$	**0.84**
TM	**0.30**	$<10^{-3}$	**0.27**	$<10^{-3}$
Année de diagnostic	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	_
Inter-région	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$
Sexe	0.01	$<10^{-3}$	0.05	0.01
Age	**0.13**	$<10^{-3}$	**0.51**	**0.18**
Type de déclarant	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$

Les variables indicatrices de données manquantes sont les variables à expliquer de chaque modèle. Ces variables sont abréviées comme suit : Rdelpos (délais positifs), Rdelneg (délais négatifs), Rserotype (sérotype), Rcd4 (taux de CD4). Sont présentées les p-valeurs associées à chacune des variables explicatives, avec en caractères gras les valeurs >0.05.

Bien que les 4 variables imputées au cours de la deuxième phase contiennent de fortes proportions de données manquantes, le modèle d'imputation est très complet du fait de l'inclusion de 11 variables auxiliaires à fortes capacités prédictrices.

Notons que dans le cas d'un modèle d'imputation complexe, le processus de sélection des variables n'est pas linéaire mais qu'il fait appel à des informations recueillies au cours du processus de validation postérieur à l'imputation. Nous présentons donc la sélection finale des variables mais plusieurs modèles qui ne seront pas détaillés ont été construits et testés.

5.2. Traitement des variables catégorielles et continues

5.2.1. Traitement des variables catégorielles

Les variables catégorielles sont bien prises en compte par la méthode d'équations chaînées. Pour chaque variable catégorielle incomplète, nominale dans le cas de cette imputation, une régression multinomiale est donc spécifiée. Il est souhaitable de ne pas dépasser 5 catégories pour des raisons de stabilité du modèle et d'équilibrer les effectifs des différentes modalités [56;125].

Lorsqu'une variable catégorielle, complète ou incomplète, est utilisée comme variable prédictrice, elle est décomposée en un jeu de variables binaires qui se substituent à la variable catégorielle d'origine dans les équations de prédiction des autres variables. L'usage montre que, pour les variables catégorielles complètes, la seule contrainte est que le nombre de variables binaires reste faible par rapport à la taille de l'échantillon [31]. Dans notre application, la variable inter-région a été utilisée comme variable prédictrice catégorielle. Nous avons dû restreindre le nombre de zones géographiques à 6 pour des raisons de stabilité du modèle.

Lors de l'imputation de variables discrètes, un biais dit de prédiction parfaite peut apparaitre. Par exemple, dans une régression logistique, un prédiction parfaite se produit si, pour une modalité d'une variable catégorielle explicative, la variable à expliquer prend toujours la valeur 1 (ou 0). La vraisemblance tend alors vers sa limite et un ou plusieurs paramètres de cette régression tendent vers moins ou plus l'infini. En conséquence, les variances associées à ces paramètres sont très élevées [137]. Ce biais de prédiction parfaite n'est traité que depuis peu et seulement par certains logiciels d'imputation multiple [48;56;138]. Une des solutions préconisées pour éviter ce biais, et implémentée dans le programme additionnel ICE, est d'ajouter à la base de données quelques observations supplémentaires avec un poids faible, ce qui permet d'éviter la prédiction parfaite. L'option AUGMLOGIT, appliquée par défaut par le programme additionnel ICE, a été utilisée durant la première phase pour l'imputation de certaines variables catégorielles.

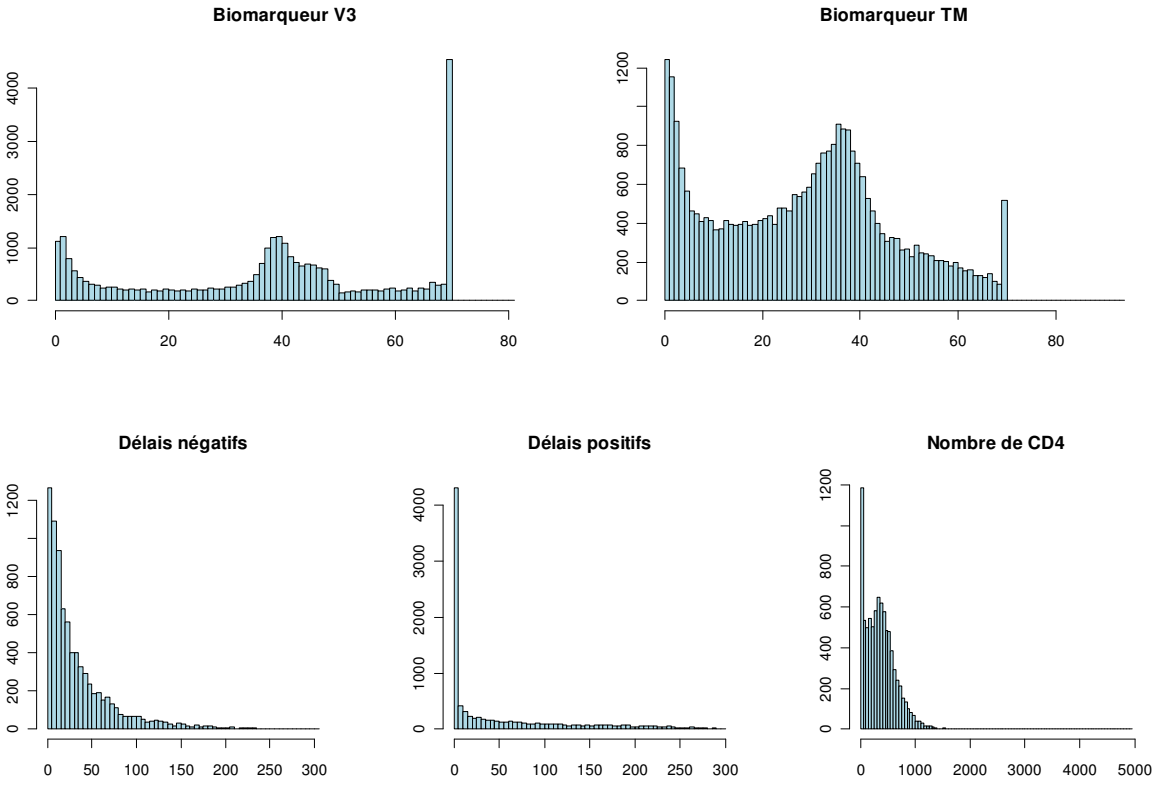
Il faut noter que le traitement de plusieurs variables catégorielles incomplètes dans un même modèle d'imputation peut être problématique en termes de convergence du modèle [28]. L'étape de diagnostic est donc importante pour ces variables.

5.2.2. Traitement des variables continues

L'imputation des variables continues repose sur l'hypothèse d'une distribution normale. Il a été démontré que les résultats de l'imputation étaient sensibles au non-respect de la normalité, aussi bien pour l'approche multivariée normale que pour l'approche par équations chaînées [45]. Il est donc essentiel de transformer chaque variable dont la distribution est nettement asymétrique afin de s'assurer de la validité de l'imputation de cette variable et ainsi d'éviter d'introduire des biais dans les estimations des coefficients des autres variables du modèle d'imputation.

Les graphes de la Figure 4.4 représentent les distributions des valeurs observées des variables continues. Les distributions des biomarqueurs V3 et TM sont nettement asymétriques avec une distribution en U et des valeurs censurées à droite. Les variables délais négatifs, délais positifs et CD4 ont une distribution plus classique même si la forte proportion de 0 est problématique pour les variables délais positifs et CD4.

Figure 4.4 – Représentation graphique des distributions observées des biomarqueurs, des délais positifs et négatifs, et des CD4



Dans le cas des biomarqueurs, des essais d'imputation ont été réalisés avec des transformations classiques (logarithme, racine carrée) ou plus élaborées en utilisant des fonctions dédiées (commandes boxcox, lnskew sous STATA) mais l'adéquation entre les distributions des données imputées et observées s'est révélée très insuffisante. Au vu de ces résultats, une transformation proposée par Nevalainen et al. [63] a été appliquée : la transformation par scores de quantiles.

Les étapes de cette transformation sont les suivantes :

Notons v_i les données de la variable à laquelle on s'intéresse.

- Les valeurs v_i sont ordonnées.

- Les ex-aequos sont départagés (zéros ou valeurs de censure) c'est-à-dire que les valeurs v_i sont légèrement transformées en les remplaçant par une valeur comprise entre $v_i - \varepsilon$ et v_i à partir de tirages aléatoires dans une distribution uniforme $(v_i - \varepsilon, v_i)$. Nous avons choisi arbitrairement $\varepsilon = 0.04$.

- Les nouvelles valeurs sont réordonnées et une variable de rang notée R_i est générée pour chaque valeur $i, (i = 1,...,n)$.

- Des valeurs z_i sont calculées à partir des R_i en utilisant la relation suivante :

$$z_i = \Phi^{-1}\left(\frac{R_i}{n+1}\right),$$

où Φ^{-1} est la fonction de répartition d'une distribution normale centrée réduite.

- Les valeurs z_i sont utilisées pour la phase d'imputation à la place des valeurs v_i.

- Les valeurs z_i imputées sont retransformées pour générer les valeurs imputées v_i en utilisant la fonction de répartition empirique des valeurs observées des v_i.

Cette transformation par scores de quantiles a été retenue pour les 5 variables quantitatives (biomarqueurs V3 et TM en phase 1, délais positifs/négatifs et CD4 en phase 2).

De plus, l'imputation de ces variables continues a été réalisée en utilisant une régression par intervalle. Pour chaque variable, des bornes inférieures et supérieures ont été spécifiées afin de contraindre les valeurs imputées à se situer dans l'intervalle des valeurs observées. Les variables et leurs bornes ont été incluses dans les modèles d'imputation. La régression par intervalle est implémentée sous le programme additionnel ICE [139] et permet d'obtenir après retransformation des valeurs strictement positives.

5.3. Choix du nombre de bases et de cycles

5.3.1. Nombre de bases

Les critères de choix du nombre de bases ont été développés dans le chapitre 1. En se basant uniquement sur le critère d'efficacité statistique relative, il faudrait générer $M = 5$ bases pour une fraction d'information manquante (FMI) de l'ordre de 25%, ce qui revient à tolérer une perte d'efficacité statistique de 5% par rapport à un nombre infini d'imputations. Si l'on approxime la FMI par la fraction de cas incomplets, il est alors recommandé d'imputer plutôt $M = 10$ bases si la proportion de cas incomplets varie de 30 à 50%, pour une même perte d'efficacité statistique de 5%. En prenant en compte des critères supplémentaires tels que la perte potentielle de puissance statistique et la répétabilité des analyses, il faut alors, comme proposé par Bodner [57], approximer la valeur de *M* par la fraction de cas incomplets, soit $M = 30$ bases de données pour une proportion de données manquantes de 30%.

Dans notre application, des contraintes pratiques doivent être prises en compte dans le choix de M : (i) le fichier de départ est volumineux (41049 observations, 100 variables), (ii) l'imputation est réalisée en deux phases et (iii) le fichier final est utilisé pour des analyses répétées au fil du temps.

Soit M_1 le nombre de bases générées pendant la première phase, M_2 le nombre de réplications de chacune des M_1 bases au cours de la deuxième phase et $M = M_1 \times M_2$ le nombre total de bases imputées pour l'ensemble du processus. Il aurait été indiqué d'imputer au moins $M_1 = 10$ bases puisque la proportion moyenne de données manquantes est d'environ 30%. Cependant, pour $M_1 = 10$ et $M_2 = 3$, la taille de la base finale aurait atteint les limites du logiciel utilisé.

Nous avons donc décidé de tolérer une perte d'efficacité statistique de plus de 5% pour la première phase et d'imputer $M_1 = 5$ bases puis $M_2 = 3$ bases.

Du fait de ce processus en deux phases, l'imputation pour les variables de la deuxième phase revient à générer 15 bases de données. Les variables imputées lors de la deuxième phase contiennent des informations identiques dans les $M_1 = 5$ bases. A partir de chacune de ces 5 bases seront générées 3 bases différentes, soit au final 15 bases potentiellement différentes pour les variables imputées au cours de la deuxième phase. Ces variables contiennent en moyenne 40% de données manquantes et le choix de $M = 15$ permet de tolérer une perte de moins de 5% d'efficacité statistique par rapport à un nombre infini d'imputations.

Lors de la phase de validation interne du processus d'imputation présentée dans le paragraphe 7, des critères tels que la FMI, l'efficacité statistique et l'erreur de Monte Carlo, calculés pour chacune des 14 variables, permettent de discuter les avantages et les limites du processus retenu.

5.3.2. Nombre de cycles

Du fait de la convergence rapide de l'échantillonneur de Gibbs, 10 cycles sont spécifiés par défaut dans le programme additionnel ICE. En lien avec la complexité des modèles d'imputation construits dans cet exemple, particulièrement pour le modèle de la première phase qui contient 5 variables catégorielles, nous avons spécifié 20 cycles pour chacune des deux phases pour nous assurer de la convergence de l'échantillonneur.

6. Imputation et analyse

Le fichier final délivré après imputation contient les 15 bases complétées et la base originelle incomplète. Chaque variable imputée est accompagnée d'une indicatrice permettant d'identifier les données d'origine. Des variables sont construites à partir des variables imputées, et sont générées automatiquement dans chacune des 15 bases. La base de données est transmise annuellement depuis 2008 à l'équipe de surveillance.

Les analyses sont réalisées sur chaque base de données puis les résultats sont combinés selon les règles de Rubin. Deux implémentations sont disponibles pour l'analyse de données imputées sous STATA : le programme additionnel MIM de Carlin et Royston [64;140] et le package MI ESTIMATE de la version STATA 11 [65], plus récent et plus complet. Cependant, les analyses de routine sont réalisées à partir de MIM en raison d'une particularité de la base de notification. En effet, les analyses descriptives concernent seulement les découvertes de séropositivité et cette notion de découverte dépend de deux variables imputées, serpo et délais positifs. Les analyses restreintes aux découvertes de séropositivité portent donc sur des effectifs qui varient selon les bases imputées et il est plus aisé de gérer cette particularité avec le programme MIM.

L'équipe de surveillance est donc autonome pour les analyses de la base imputée. Ces analyses prennent également en compte l'exhaustivité de la notification obligatoire ainsi que les délais de notification. Cette étape intervient après l'imputation multiple et ses résultats doivent être combinés avec ceux de l'imputation pour produire les résultats finaux (valeur centrale, variance).

7. Validation interne

Afin d'obtenir des estimations valides après imputation multiple, deux conditions sont requises : (i) le modèle d'imputation doit être correctement spécifié en termes de sélection et de type de variables, (ii) l'hypothèse que les données sont MAR doit être vérifiée.

L'étape de validation interne permet de vérifier la première condition, c'est-à-dire la validité du modèle, sachant que, si des données sont manquantes selon un mécanisme MNAR, les résultats de l'imputation seront biaisés même si le modèle est correctement spécifié. Cette validation interne a été effectuée à deux niveaux : des approches simples de diagnostics graphiques et numériques, puis une étude par simulation permettant d'évaluer les capacités prédictrices du modèle de la première phase.

7.1. Diagnostic de l'imputation

7.1.1. Problématique

Cette étape de diagnostic consiste à rechercher des incohérences marquées dans les résultats de l'imputation. A un stade initial où le modèle d'imputation n'est pas encore stabilisé, différentes équations de prédictions ainsi que différents codages et/ou transformations peuvent être testés.

Comme suggéré par Abayomi et al. et Raghunatan et al. [61;62], un critère de validité interne réside dans la comparaison des données imputées et observées via les variables indicatrices de données manquantes. En effet, sous l'hypothèse que les données sont manquantes aléatoirement, des différences peuvent être détectées entre données observées et imputées, mais elles doivent rester modérées et explicables par des mécanismes identifiés.

Abayomi et al. proposent d'effectuer des comparaisons numériques et graphiques, appliquées consécutivement à plusieurs bases de données imputées. Dans notre étude, les comparaisons numériques ne sont pas facilement interprétables en raison (i) de la différence d'effectifs entre valeurs imputées et observées et (ii) de la taille de la base de données de départ. Ainsi, des tests statistiques réalisés sur une seule base imputée (41049 observations) sont dans cette étude

artificiellement significatifs. Une comparaison directe des proportions, complétée par des analyses graphiques, permettent cependant d'effectuer une première étape de validation.

7.1.2. Variables discrètes

Les proportions estimées à partir des données imputées sont assorties d'un intervalle de confiance à 95% (IC 95%). L'examen qui nous a paru le plus adapté a consisté à vérifier si les proportions observées étaient ou non contenues dans les intervalles de confiance des proportions estimées. Cet examen a également été effectué après ajustement sur certaines covariables, identifiées à partir de l'analyse des mécanismes de données manquantes présentée dans les Tableaux 4.3 et 4.4.

- ***Variables binaires***

Sont présentées dans le Tableau 4.5 les proportions observées et estimées pour les 3 variables binaires serneg, serpo et sérotype.

- **Sérologie négative antérieure : variable serneg**

Rappelons que la variable serneg, du fait d'un problème de codage initial, a été reconstruite et contient au final 61% de données manquantes. Les proportions observées pour toute la période ainsi que selon l'année de diagnostic n'appartiennent pas aux IC 95% des proportions estimées. Pour toute la période, l'écart entre ces proportions est cependant modéré avec une différence de moins de 3% sur 40%, sachant que l'impact des valeurs imputées sur la proportion globale (données observées et imputées) est de 1%.

Tableau 4.5 – Proportions observées et estimées des variables binaires

SEROLOGIE NEGATIVE ANTERIEURE									
Année de diagnostic	2003	2004	2005	2006	2007	2008	2009	2010	2003-2010
Données observées	**24.37**	**28.79**	31.07	34.01	38.09	**54.86**	**58.97**	**64.92**	**40.90**
Données imputées	17.77	21.2	28.09	34.08	40.19	43.39	50.27	55.09	38.09
(IC 95%)	(14.3-21.2)	(18.6-23.8)	(25.0-31.1)	(30.8-37.3)	(37.7-42.7)	(41.4-45.4)	(47.8-52.7)	(52.0-58.1)	(37.2-38.9)
Stade clinique	PIV*	ASY*	SNS*	SID*					
Données observées	62.78	40.62	24.41	22.12					
Données imputées	62.80	40.28	26.42	23.72					
(IC 95%)	(59.2-66.4)	(39.0-41.5)	(24.3-28.5)	(18.3-29.0)					
SEROLOGIE POSITIVE ANTERIEURE									
Année de diagnostic	2003	2004	2005	2006	2007	2008	2009	2010	2003-2010
Données observées	36.14	33.43	**36.02**	**36.55**	**37.48**	**42.68**	**43.87**	**45.22**	**38.31**
Données imputées	32.76	31.19	30.48	32.26	32.76	32.52	33.39	34.24	32.53
IC 95%	(28.9-36.6)	(28.2-34.2)	(27.1-33.9)	(29.4-35.1)	(30.2-35.3)	(29.2-35.8)	(31.4-35.4)	(31.7-36.8)	(31.3-33.8)
Mode de contamination	HOMO/BI†	UDI†	HETERO†	AUTRES†					
Données observées	**33.10**	**8.23**	**56.66**	**2.01**					
Données imputées	30.22	6.85	60.07	2.86					
(IC 95%)	(28.2-32.4)	(6.0-7.7)	(57.8-62.3)	(2.3-3.4)					
Motif de dépistage	SCB‡	EXP‡	BIL/GROS‡	DOR‡	PEC/AUT‡				
Données observées	26.33	14.83	27.96	20.70	79.88				
Données imputées	27.19	14.32	25.54	19.92	76.76				
IC 95%	(25.2-29.2)	(12.5-16.1)	(22.3-28.8)	(15.2-24.7)	(73.2-80.3)				
SEROTYPE B									
Année de diagnostic	2003	2004	2005	2006	2007	2008	2009	2010	2003-2010
Données observées	**51.61**	**56.08**	**60.48**	**59.44**	**61.22**	**58.54**	**57.26**	**55.72**	**58.05**
Données imputées	65.32	64.81	67.52	66.46	67.28	65.89	66.62	69.6	66.65
(IC 95%)	(61.3;69.4)	(61.9;67.7)	(63.9;71.2)	(62.9;70.1)	(64.0;70.6)	(62.7;69.1)	(63.2;70.0)	(66.0;73.2)	(65.2;68.1)
Pays de naissance	France	AFSS§	EUROPE	AME/HAITI§	AUTRES				
Données observées	**78.30**	28.40	67.53	85.76	62.48				
Données imputées	81.28	31.16	71.91	87.50	64.51				
(IC 95%)	(80.0;82.5)	(28.2;34.1)	(66.7;77.2)	(84.7;90.4)	(58.6;70.4)				

Une proportion observée est notée en caractères gras si elle n'appartient pas à l'IC 95% de la proportion estimée.
* PIV : primo-infection virale ; ASY : asymptomatique ; SNS : symptomatique non sida ; SID : sida.
† HOMO/BI : mode homosexuel/bisexuel ; UDI : usage de drogue intraveineuse ; HETERO : mode hétérosexuel. ‡ SCB : symptômes cliniques et biologiques ; EXP : exposition ; BIL/GROS : bilan/grossesse ; DOR : dépistage orienté ; PEC/AUT : prise en charge/autre. § AFSS : Afrique sub-Saharienne ; AME/HAITI : Amérique/Haïti

Notons que, selon le Tableau 4.3, le modèle de prédiction de la variable serneg est riche puisque les 4 variables catégorielles sont prédictrices de serneg, c'est-à-dire qu'elles sont significativement liées à l'indicatrice de données manquantes de cette variable. On observe d'ailleurs que les différences entre les proportions observées et estimées sont plus faibles après ajustement sur les variables pays de naissance et mode de contamination. Après ajustement sur la variable stade clinique, les différences entre proportions observées et estimées ne sont plus significatives. On peut donc conclure que le mécanisme de données manquantes de la variable serneg est complexe, mais que le modèle de prédiction inclut sans doute les principaux prédicteurs. Une variabilité résiduelle est attendue du fait de la proportion élevée de données manquantes.

- **Sérologie positive antérieure : variable serpo**

Comme pour la variable serneg, la qualité de remplissage de la variable serpo a été altérée par la formulation de la question dans la fiche de notification. L'impact sur l'équilibre des réponses oui-non est visible jusqu'en 2008, avec une surreprésentation de réponses négatives. Avec la mise en circulation des nouvelles fiches, la proportion de réponses négatives diminue au profit des données manquantes. De ce fait, la proportion de données manquantes augmente nettement au fil du temps, de 26% en 2003 à 36% en 2007 et presque 50% en 2010. Notons que, pour la variable serpo, un processus de tri des fiches n'a pas été appliqué, car l'impact sur l'équilibre des réponses oui-non était moindre que pour la variable serneg.

Les résultats du Tableau 4.5 montrent que les proportions observées pour la période 2003-2010 et par année n'appartiennent pas aux IC 95% des proportions estimées. L'écart entre les proportions observées et imputées de la période complète est important, de l'ordre de 6% sur 36%, même si l'impact des valeurs imputées sur la proportion globale (valeurs observées et imputées) n'est au final que de 2 % (38.3% de réponses positives pour les valeurs observées et 36.2% en global). Par année de diagnostic, l'impact des valeurs imputées sur les proportions globales est inférieur à 1% jusqu'en 2007, puis s'accroit et atteint 5 % en 2010.

Les résultats de l'imputation apparaissent donc moins fiables pour la variable serpo que pour la variable serneg malgré une proportion de données manquantes inférieure (36% versus 61%). Cependant, le modèle d'imputation de la variable serpo est moins riche que celui de la variable serneg, puisque seulement 2 des 4 variables catégorielles sont prédictrices de serpo. Après

ajustement sur ces deux variables, mode de contamination et motif de dépistage, l'écart entre proportions observées et estimées est réduit, et devient même non-significatif après ajustement sur la variable motif de dépistage.

La variable serpo est une variable clé car elle permet de déterminer, en association avec la valeur des délais positifs, la proportion de nouveaux diagnostics. Rappelons qu'un nouveau diagnostic est posé quand aucune sérologie positive antérieure n'est rapportée ou qu'elle date de plus de 11 mois. L'imputation a pour effet de diminuer la proportion de réponses positives pour serpo, et donc d'augmenter la proportion de nouveaux diagnostics, sans tenir compte de l'imputation des délais positifs. Ainsi, la proportion de nouveaux diagnostics passe pour la période 2003-2010 de 79.7% à partir des données observées à 81.4% à partir des données globales après imputation, et pour l'année 2010 de 77.1% à 80.8%.

- **Sérotype viral B versus non-B pour les souches virales de type 1 : variable sérotype**

Sont présentées dans le Tableau 4.5 les proportions de souches virales de type 1 et de sous-type B, les autres types et sous-types étant regroupés dans la catégorie non-B. Les proportions observées de sérotype B, par année de diagnostic et sur toute la période, ne sont pas incluses dans les IC 95% des proportions estimées. Les proportions estimées de sérotype B sont systématiquement plus élevées que les proportions observées, avec un écart global de 9% sur 60%, et des écarts par année variant de 6% en milieu de période pour atteindre 14% sur 69% en 2010. L'impact des données imputées sur les données globales n'est cependant que de 2.8% pour l'ensemble de la période et atteint 5% en 2010.

L'imputation de la variable sérotype est réalisée au cours de la deuxième phase, conditionnellement à la variable type viral imputée lors de la première phase. L'équation de prédiction de la variable sérotype contient 8 variables prédictrices et 15 bases sont imputées. Un problème de spécification du modèle parait donc peu probable. De plus, les variations observées après imputation peuvent être expliquées, au moins en partie, par le lien entre les variables sérotype et pays de naissance. Ainsi, le sérotype est un des indicateurs de l'évolution de l'épidémie en France en matière de flux migratoires, en particulier en provenance d'Afrique sub-Saharienne (AFSS). Historiquement, le sérotype B est majoritaire en France et les sérotypes non-B prédominent en AFSS. Ainsi, après ajustement sur la variable pays de naissance, la proportion

de souches virales B reste plus élevée pour les valeurs imputées, mais cette différence n'est plus significative qu'en France.

Afin d'interpréter cette différence en France, nous avons généré une variable indicatrice permettant d'identifier après imputation des catégories de données manquantes d'origine différente. Les 3 catégories de données manquantes correspondent (1) aux buvards non-réalisés ou inexploitables, (2) aux souches virales de type M (recodées en données manquantes pour obtenir le sous-type) et (3) aux souches virales définies comme non-typables, très majoritairement parce que l'infection est trop récente. On observe que la proportion de sérotype B estimée passe de 62.1% pour (1) à 75.0% pour (2) et 78.0% pour (3). Par ailleurs la proportion de nouveaux diagnostics passe de 81.4% pour (1), à 89.7% pour (2) et à 93.6% pour (3). La proportion de sérotype viral B est donc plus élevée parmi les infections récentes. Ce constat peut être lié aux pratiques de dépistage selon le mode de contamination, puisque les individus ayant des pratiques homosexuelles se testent plus fréquemment que les hétérosexuels (d'où un taux d'infection récente plus élevé) et sont majoritairement nés en France, où le sérotype B prédomine. Ce mécanisme peut expliquer la proportion plus élevée de sérotype B parmi les infections récentes.

En règle générale, l'imputation de variables binaires par la méthode d'équations chaînées donne des résultats fiables, si les conditions d'application de la méthode sont remplies. Dans cette application, les variables serneg et serpo contiennent une forte proportion de données manquantes et le mécanisme à l'origine de ces données manquantes varie selon les individus et avec le temps. Cet effet semble paradoxalement moins marqué pour serneg, variable pour laquelle une partie des données manquantes a été générée a posteriori et selon un mécanisme défini. Dans le cas du sérotype viral, cette variable binaire est en fait de structure initiale complexe, avec des données manquantes découlant de mécanismes très différents.

La conclusion de cette étape de diagnostic est que les variations observées sont contrôlées pour les variables serneg et sérotype, mais que l'imputation de la variable serpo gagnerait à être améliorée, éventuellement par l'application d'un filtre comme dans le cas de serneg.

- ***Variables catégorielles***

Le tableau 4.6 présente les proportions des variables pays de naissance et mode de contamination, et le Tableau 4.7 détaille les variables stade clinique, motif de dépistage et type viral.

Tableau 4.6 - Proportions observées et estimées des variables catégorielles pays de naissance et mode de contamination

PAYS DE NAISSANCE	France	AFSS*	EUROPE	AME/HAITI*	AUTRES
Données observées	**48.08**	36.27	3.56	**7.22**	4.87
Données imputées	46.46	36.46	3.89	8.21	4.98
IC 95%	(45.4-47.5)	(35.2-37.7)	(3.3-4.5)	(7.4-9.0)	(4.3-5.9)
MODE DE CONTAMINATION	HOMO/BI†	UDI†	HETERO†	AUTRES	
Femmes					
Données observées		2.78	94.81	2.42	
Données imputées		2.35	95.73	1.92	
IC 95%		(1.6-3.1)	(94.8-96.6)	1.4-2.4	
Hommes					
Données observées	**52.41**	4.76	**41.48**	1.35	
Données imputées	46.75	4.52	47.30	1.43	
IC 95%	(44.9-48.6)	(3.9-5.2)	(45.5-49.1)	(1.1-1.8)	
Selon le pays de naissance (Hommes)	HOMO/BI†	UDI†	HETERO†	AUTRES	
FRANCE					
Données observées	**69.39**	4.65	**24.79**	1.17	
Données imputées	66.80	4.88	27.36	0.96	
IC 95%	(65.1-68.5)	(4.0-5.7)	(25.6-29.2)	(0.5-1.5)	
AFSS*					
Données observées	5.72	**0.86**	91.42	1.99	
Données imputées	5.35	0.71	91.56	2.39	
IC 95%	(3.7-7.0)	(0.1-1.4)	(89.5-93.6)	(1.6-3.1)	
EUROPE					
Données observées	49.75	23.94	24.82	1.50	
Données imputées	48.44	19.90	30.33	1.33	
IC 95%	(42.9-54.0)	(13.4-26.4)	(23.3-37.4)	(0-3.8)	
AME/HAITI*					
Données observées	44.67	0.84	53.39	1.11	
Données imputées	40.46	0.84	57.18	1.51	
IC 95%	(36.2-44.7)	(0-2.0)	(52.6-61.8)	(0-3.3)	
AUTRES					
Données observées	41.67	9.95	47.19	1.20	
Données imputées	38.69	9.34	50.27	1.70	
IC 95%	(30.7-46.6)	(5.2-13.4)	(40.8-59.9)	(0-3.4)	

Une proportion observée est notée en caractères gras si elle n'appartient pas à l'IC 95% de la proportion estimée.
* AFSS : Afrique sub-Saharienne ; AME/HAITI : Amérique/Haïti ;
† HOMO/BI : mode homosexuel/bisexuel ; UDI : usage de drogue intraveineuse ; HETERO : mode hétérosexuel.

- **Pays de naissance (France, Afrique sub-Saharienne, Europe, Amérique/Haïti, Autres)**

Les proportions estimées à partir des données imputées sont proches des valeurs observées, avec cependant une différence significative pour la France (Tableau 4.6). L'impact des valeurs imputées sur les valeurs totales reste inférieur à 0.5%. L'équation de prédiction de la variable pays de naissance contient 7 variables prédictrices (Tableau 4.3).

- **Mode de contamination (homosexuel/bisexuel, usage de drogue intraveineuse, hétérosexuel, autres)**

L'examen des proportions a été réalisé selon le sexe. Le Tableau 4.6 montre des différences parmi les hommes puisque, pour les deux modalités principales (modes homo/bisexuel et hétérosexuel), les proportions observées n'appartiennent pas aux IC 95% des proportions estimées. La proportion estimée est plus élevée que la proportion observée parmi les hommes se contaminant par voie homosexuelle ou bisexuelle (écart de 6% sur 60%), et on observe la relation inverse pour la voie hérétosexuelle (écart de 3% sur 25%).

La variable mode de contamination est liée à la variable pays de naissance. Chez les hommes, après stratification sur le pays de naissance, des différences entre données imputées et observées ne sont plus significatives qu'en France. La proportion observée est supérieure à la proportion estimée pour le mode de contamination homo/bisexuel en France, avec un impact inférieur à 1% sur 70% pour les données totales. Il est cohérent d'obtenir une proportion estimée plus faible que la proportion observée pour le mode de contamination homosexuel. En effet, une hiérarchisation des réponses lors de la saisie des fiches de notification privilégie la modalité contamination par voie homo/bisexuelle par rapport aux autres modalités, soit essentiellement la modalité contamination par voie hétérosexuelle (le mode de saisie est basé sur l'hypothèse que, si le mode de contamination homosexuel/bisexuel est renseigné, il constitue la cause probable de contamination, ce qui n'est pas le cas pour le mode hétérosexuel).

Par ailleurs, la variable mode de contamination est renseignée par voie déclarative par le patient à la demande du praticien. Cette question peut être considérée comme sensible socialement, en particulier pour la catégorie des homo/bisexuels, ce qui pourrait provoquer des données manquantes de type MNAR (par exemple si un homme contaminé par voie homosexuelle préfère

ne pas divulguer cette information). Cependant, l'examen des proportions avant et après imputation montre des résultats proches, particulièrement après ajustement sur le pays de naissance. Le modèle d'imputation de la variable mode de contamination est riche (10 variables prédictrices) et inclut probablement les variables les plus importantes, permettant ainsi de réduire l'effet sur les estimations d'un éventuel mécanisme MNAR.

Tableau 4.7 - Proportions observées et estimées des variables catégorielles stade clinique, motif de dépistage et type viral

STADE CLINIQUE	PIV*	ASY*	SNS*	SID*	
Données observées	**8.12**	58.71	13.50	**19.67**	
Données imputées	6.80	62.40	13.91	16.88	
IC 95%	(5.9-7.7)	(61.3-63.5)	(12.9-14.9)	(15.9-17.9)	
Selon le motif de dépistage					
SCB†					
Données observées	15.82	16.01	24.58	**43.59**	
Données imputées	14.06	17.84	27.46	40.64	
IC 95%	(12.0-16.1)	(15.8-19.8)	(25.3-29.6)	(38.1-43.2)	
EXP†					
Données observées	7.75	85.31	4.81	2.13	
Données imputées	6.47	86.46	5.11	1.97	
IC 95%	(4.5-8.5)	(83.8-89.1)	(3.9-6.4)	(0.9-3.0)	
BIL/GROS†					
Données observées	1.38	89.25	4.77	4.60	
Données imputées	1.14	89.95	4.91	4.00	
IC 95%	(0.4-1.9)	(87.8-92.1)	(3.5-6.3)	(2.5-5.5)	
DOR†					
Données observées	4.15	79.60	13.25	3.00	
Données imputées	4.11	78.99	14.14	2.76	
IC 95%	(1.3-7.0)	(69.5-88.3)	(7.5-20.9)	(0-7.2)	
PEC/AUTRES†					
Données observées	2.53	74.15	11.14	12.18	
Données imputées	2.44	75.58	11.30	10.67	
IC 95%	(1.7-3.2)	(73.2-78.0)	(9.6-13.0)	(7.8-13.5)	
MOTIF DE DEPISTAGE	SCB†	EXP†	BIL/GROS†	DOR†	PEC/AUT†
Données observées	**33.84**	**20.78**	19.40	2.57	23.41
Données imputées	36.82	19.15	19.31	2.31	22.41
IC 95%	(35.6-38.1)	(18.4-19.9)	(18.4-20.2)	(2.0-2.6)	(21.0-23.8)
TYPE VIRAL	Type 1	Type 2	Type 1 et 2		
Données observées	98.23	1.70	0.08		
Données imputées	98.59	1.34	0.07		

Une proportion observée est notée en caractères gras si elle n'appartient pas à l'IC 95% de la proportion estimée.
* PIV : primo-infection virale ; ASY : asymptomatique ; SNS : symptomatique non sida ; SID : sida.
† SCB : symptômes cliniques et biologiques ; EXP : exposition ; BIL/GROS : bilan/grossesse ; DOR : dépistage orienté ; PEC/AUT : prise en charge/autres.

- **Stade clinique (asymptomatique, primo-infection virale, symptomatique non-sida, sida)**

Pour 3 catégories sur 4 de la variable stade clinique, les proportions observées n'appartiennent pas aux IC 95% des proportions estimées, avec des écarts modérés entre les proportions, de 2.5% sur 7% à 3% sur 18% selon les modalités (Tableau 4.7). Les capacités prédictrices du modèle d'imputation pour la variable stade clinique paraissent satisfaisantes, puisque les 3 autres variables catégorielles ainsi que 4 variables auxiliaires sont prédictrices. Ainsi, un ajustement sur les variables pays de naissance et mode de contamination permet d'atténuer les écarts entre les proportions. C'est cependant la variable motif de dépistage qui est la plus nettement prédictrice puisque, après ajustement sur cette variable, les proportions observées et estimées de la variable stade clinique ne différent pas significativement.

- **Motif de dépistage (symptômes cliniques et biologiques, exposition, bilan/grossesse, dépistage orienté, prise en charge/autres)**

Les résultats du Tableau 4.7 montrent que les proportions observées et estimées sont proches. Ainsi, les valeurs observées sont incluses dans l'IC 95% des valeurs imputées pour 3 catégories sur 5, avec un écart maximal de 2% sur 35% pour les 2 autres catégories. Par ailleurs, l'équation de prédiction de la variable motif de dépistage contient 9 variables prédictrices sur les 12 incluses dans le modèle d'imputation.

- **Type viral (type 1, type 2, type 1 et 2)**

La variable type viral reprend le codage originel et contient donc des effectifs très déséquilibrés entre les modalités, puisque le type viral 1 est très peu fréquent et donc a fortiori la co-infection par les types viraux 1 et 2. Malgré ce déséquilibre, les proportions estimées sont très proches des proportions observées, sachant que cette variable a dû être exclue de l'équation de prédiction des autres variables lors de la première phase d'imputation car elle induisait une instabilité du modèle. Notons que les IC 95% des proportions estimées n'ont pu être calculés pour 2 catégories sur 3 du fait des faibles effectifs.

Les données de la littérature montrent que les variables catégorielles sont les plus problématiques à imputer, même si l'utilisation de la méthode par équations chaînées permettant de spécifier une

fonction de lien multinomiale limite les risques de biais. La décomposition des variables catégorielles en variables binaires incluses dans les équations de prédiction des autres variables a permis d'améliorer nettement les qualités prédictrices des modèles d'imputation si l'on compare les résultats au cours du temps (depuis 2008).

Par ailleurs, l'inclusion de plusieurs variables catégorielles incomplètes dans un même modèle d'imputation peut induire une instabilité de ce modèle. Nous avons pu observer ce problème avec la variable type viral, mais les 4 autres variables catégorielles, globalement très liées les unes aux autres, ont pu être maintenues dans toutes les équations de prédiction, ce qui a permis d'améliorer les qualités prédictrices du modèle d'imputation de la phase 1. Enfin, la variable mode de contamination, considérée a priori comme étant à risque de contenir des données manquantes de type MNAR, ne présente pas de décalage majeur et inexpliqué entre les résultats de l'imputation et les données d'origine. Nous en concluons que le modèle d'imputation pour cette variable est suffisamment riche en prédicteurs pour rendre peu perceptible l'effet d'un mécanisme MNAR potentiel.

Cette étape de diagnostic montre donc que, pour les variables catégorielles, les différences entre valeurs observées et imputées sont faibles, ou/et peuvent être expliquée par un lien avec une ou plusieurs autres variables, ce qui définit un mécanisme MAR plus ou moins complexe selon les variables.

7.1.3. Variables continues

Cinq variables continues sont imputées au total, les biomarqueurs V3 et Tm à la première phase, les délais positifs et négatifs ainsi que le taux de Cd4 à la deuxième phase. Les procédés de diagnostic pour l'imputation des variables continues proposés dans la littérature sont des tests numériques et graphiques.

Pour les bases de données importantes, les tests numériques tels que le test de Kolmogorov-Smirnov, indiqué pour tester l'égalité de deux distributions, sont surtout utilisés pour dépister les variables pour lesquelles un examen plus approfondi doit être mené [61]. Dans notre application, le test de Kolmogorov-Smirnov, appliqué à chaque base de données imputée une à une, conclut pour chaque variable testée à une différence significative entre données imputées et observées.

Cependant, l'importance des effectifs (>40000) peut donner à tort des résultats significatifs du fait de la puissance du test [62].

De ce fait, nous avons complété cette approche par une représentation graphique. Pour cette analyse graphique, nous avons retenu des graphes quantiles-quantiles et des histogrammes représentant les valeurs observées versus les valeurs imputées. Les Figures 4.5, 4.6 et 4.7 montrent que, pour les 5 variables considérées, les distributions des données observées et imputées sont proches.

Figure 4.5 – Comparaison graphique des données observées et imputées des délais positifs et négatifs

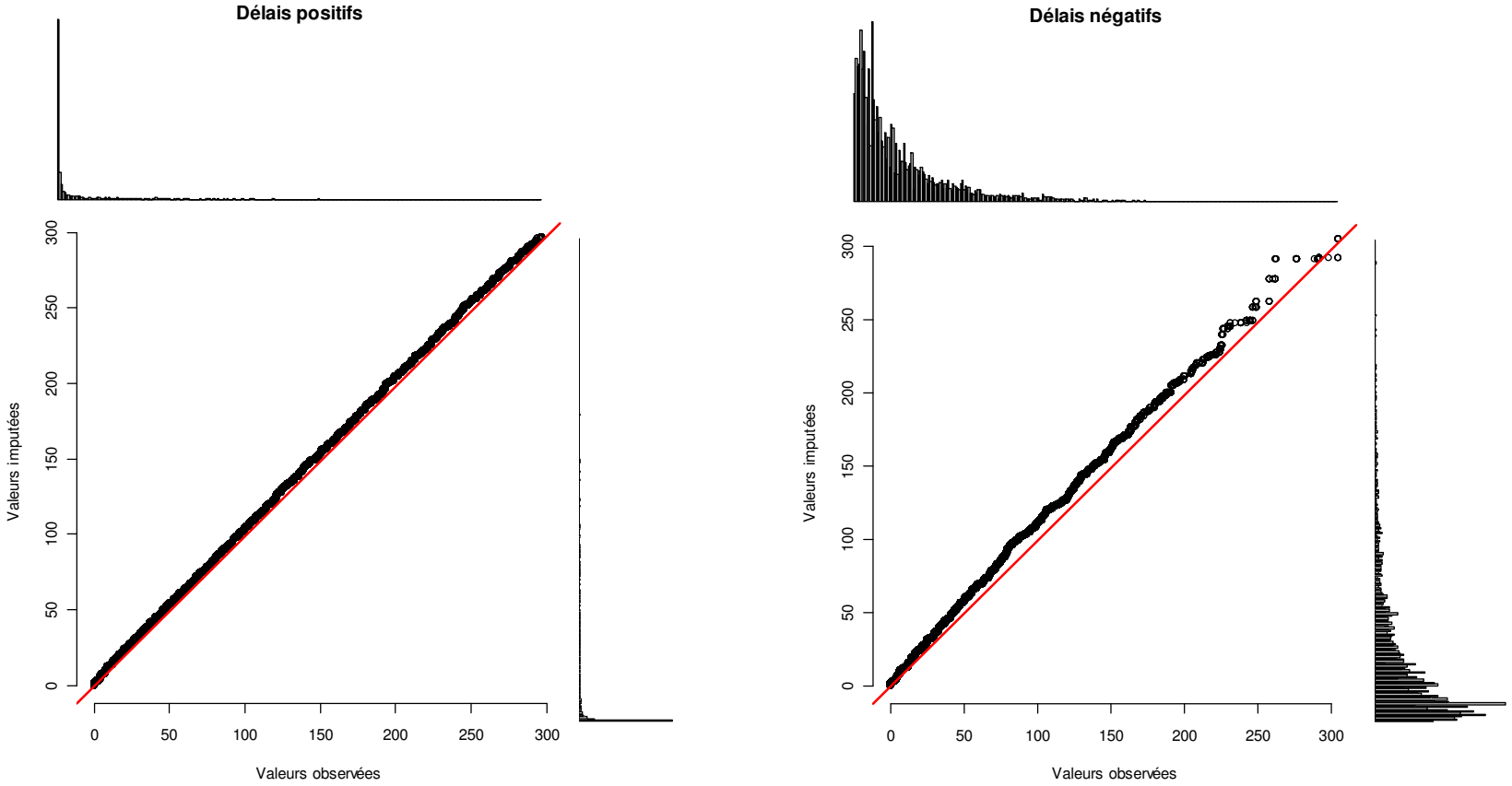


Figure 4.6 – Comparaison graphique des données observées et imputées des biomarqueurs V3 et TM

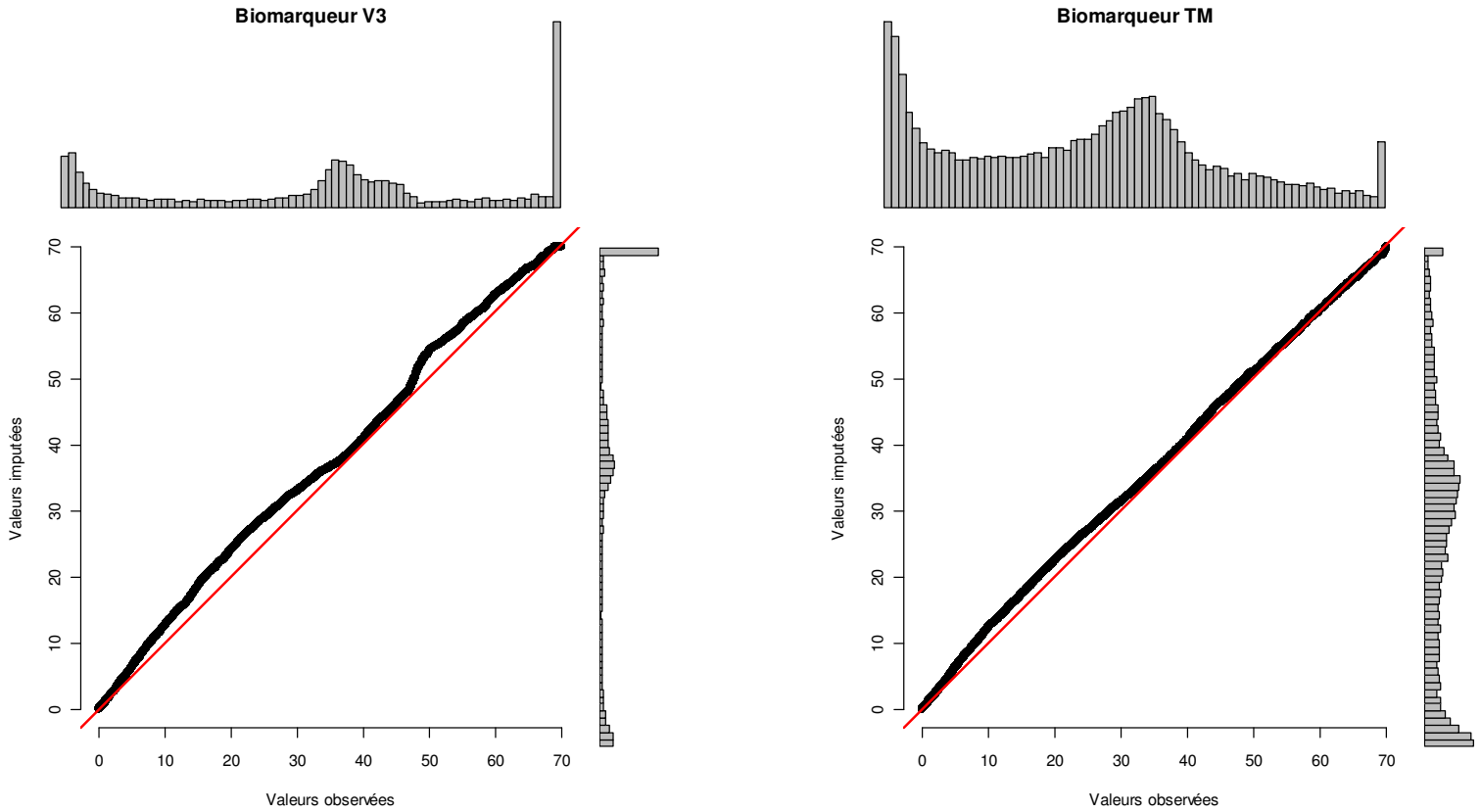
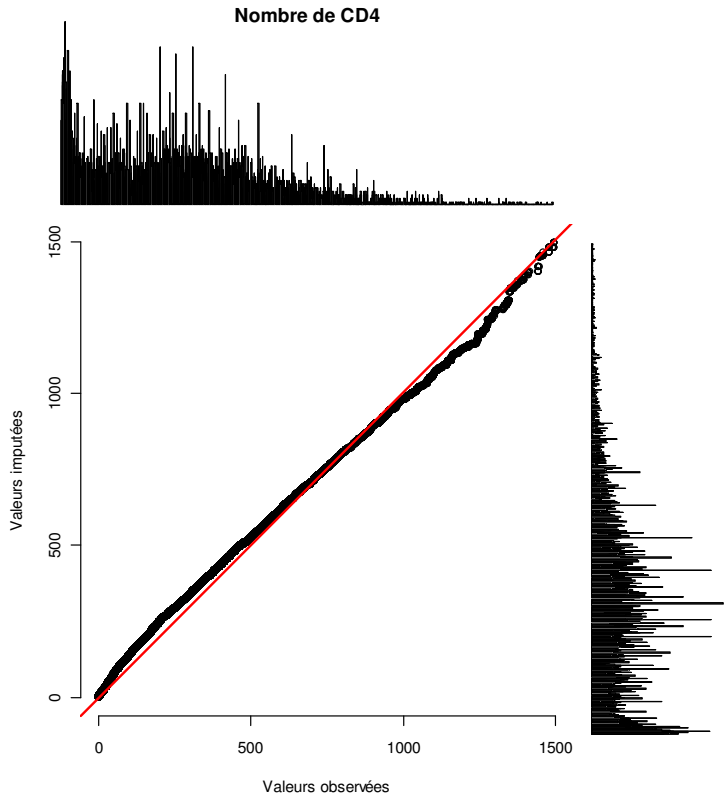


Figure 4.7 – Comparaison graphique des données observées et imputées du taux de CD4



Afin de compléter cette approche graphique simple, nous avons comparé les distributions observées et imputées selon leurs quantiles, leurs moyennes et leurs variances. Nous avons appliqué un diagnostic proposé par Stuart et al. [82] et qui définit deux indicateurs de discordance entre les distributions : (i) une différence en valeur absolue entre les moyennes des données observées et imputées supérieure à 2 écarts types (imputés), (ii) un ratio des variances observées et estimées inférieur à 0.5 (variance estimée/variance observée) ou supérieur à 2 (variance observée/variance imputée).

Tableau 4.8 – Comparaisons numériques des données observées et imputées des variables quantitatives

	Quantiles					moy*	|Δ moy|†	SE	|Δmoy| / 2SE	var	var_{obs} / var_{imp} ‡	var_{imp} / var_{obs} §	FMI (%)
	1%	25%	50%	75%	99%								
BIOMARQUEUR V3													
Observé	0.4	20.5	39.8	55.3	70.0	38.10	0.32	0.17	0.9	0.03	1.0	1.0	42.6
Imputé	0.4	19.0	39.7	55.7	70.0	37.78		0.17		0.03			
BIOMARQUEUR TM													
Observé	0.3	12.9	30.0	40.0	70.0	28.51	0.51	0.10	2.3	0.01	0.9	1.1	33.2
Imputé	0.3	11.6	29.5	39.8	70.0	28.00		0.11		0.01			
DELAIS POSITIFS													
Observé	0.0	1.0	12.0	90.0	266.0	54.04	0.60	0.75	0.4	0.56	0.9	1.1	23.9
Imputé	0.0	1.0	13.0	91.0	268.0	54.64		0.80		0.64			
DELAIS NEGATIFS													
Observé	0.0	9.0	21.0	47.0	196.0	35.79	2.08	0.46	2.2	0.21	1.0	1.0	47.7
Imputé	0.0	9.0	21.0	45.0	180.0	33.71		0.47		0.22			
TAUX DE CD4													
Observé	3.0	143.0	333.0	524.0	1226.0	366.84	8.60	3.14	1.2	9.86	0.8	1.3	20.8
Imputé	4.0	160.0	350.0	534.0	1163.0	375.44		3.56		12.67			

* Moyenne
† Valeur absolue de la différence entre moyenne des valeurs observées et imputées
‡ Rapport de la variance de la moyenne des valeurs observées sur celle des valeurs imputées
§ Rapport de la variance de la moyenne des valeurs imputées sur celle des valeurs observées

Les résultats sont présentés dans le Tableau 4.8. La comparaison des valeurs par quantiles montre des valeurs très proches entre les valeurs observées et imputées pour les biomarqueurs et les délais, mais dénote plus de variabilité pour le taux de CD4. L'indicateur défini comme la différence des moyennes divisée par 2 écarts types est supérieur à 1 pour le biomarqueur TM, les délais négatifs et les CD4. Notons cependant que, pour ces variables, la différence entre les moyennes des proportions observées et imputées est faible, mais que l'écart type est également

petit, ce qui affecte la valeur du ratio. Cela est confirmé par les indicateurs définis à partir des rapports de variances (proportions observées et estimées) qui sont très proches de 1 pour 4 variables sur 5, et incluses dans l'intervalle prédéfini (0.5 ;2.0) pour les CD4.

Les distributions des données observées et imputées sont donc proches pour les variables considérées. Les équations de prédiction varient selon les variables, avec 6 variables prédictrices pour les biomarqueurs, 8 pour les Cd4, et respectivement 9 et 11 pour les délais positifs et négatifs. Le modèle d'imputation apparait donc efficace pour les variables continues, avec une variabilité supplémentaire pour la variable taux de CD4, imputée seulement sur une période de 2 ans.

Posons quelques hypothèses sur les mécanismes générant des données manquantes pour ces variables. Les valeurs des biomarqueurs sont manquantes lorsque l'échantillon n'est pas transmis au laboratoire, qu'il n'est pas utilisable techniquement ou que le patient refuse le test. Ce mécanisme ne dépend donc pas des valeurs non-observées des biomarqueurs. Un mécanisme similaire peut être envisagé pour le taux de CD4, mesuré également à partir d'un prélèvement sanguin. Pour la valeur des délais, son recueil est déclaratif ou basé sur le dossier médical du patient. Cette valeur pourrait être volontairement omise, mais sans rapport identifiable avec la valeur du délai.

Cette étape de diagnostic des variables imputées continues est donc satisfaisante puisqu'elle permet de conclure que (i) le modèle d'imputation apparaît correctement spécifié, en termes de sélection de prédicteurs mais aussi de normalisation des variables continues et (ii) qu'un mécanisme de type MNAR est peu probable.

7.1.4. Critères diagnostiques du nombre de bases

Le Tableau 4.9 présente pour chaque variable imputée la proportion de données manquantes globale, la fraction d'information manquante (*FMI*) par modalité, l'efficacité statistique relative et l'erreur de Monte Carlo. Les résultats sont présentés par phase, sachant que 5 bases sont générées à la première phase et 15 après la deuxième phase.

Tableau 4.9 – Critères diagnostiques du nombre de bases imputées

	Proportion Moyenne	Données manquantes (%)	FMI (%)	Efficacité Statistique (%)	SE	Mcerror*	Mcerror/SE (%)
PREMIERE PHASE							
Pays de naissance							
FRANCE	47.66%	26.16	9.36	98.16	2.58E-03	3.16E-04	12.23
AFSS	36.32%		24.43	95.34	2.70E-03	5.23E-04	19.39
EUROPE	3.64%		**38.82**	**92.80**	1.15E-03	2.78E-04	**24.20**
AME/HAITI	7.48%		**33.11**	**93.79**	1.55E-03	3.48E-04	**22.41**
AUTRES	4.90%		**32.37**	**93.92**	1.27E-03	2.81E-04	**22.15**
Mode de contamination*							
HOMO/BI†	32.10%	29.01	**32.09**	**93.97**	2.74E-03	6.05E-04	**22.09**
UDI†	3.94%		29.56	**94.42**	1.12E-03	2.39E-04	**21.27**
HETERO†	62.26%		28.65	**94.58**	2.79E-03	5.83E-04	**20.93**
AUTRES	1.70%		23.82	95.45	7.23E-04	1.38E-04	19.09
Stade clinique							
PIV‡	7.78%	25.73	**43.40**	**92.01**	1.70E-03	4.34E-04	**25.59**
ASY‡	59.66%		13.80	97.31	2.60E-03	3.83E-04	14.75
SNS‡	13.61%		29.74	**94.39**	1.98E-03	4.22E-04	**21.28**
SID‡	18.95%		21.71	95.84	2.16E-03	3.97E-04	18.34
Motif de dépistage‡							
SCB§	34.65%	27.25	25.94	95.07	2.69E-03	5.37E-04	19.95
EXP§	20.34%		1.90	99.62	2.01E-03	1.12E-04	5.58
BIL/GROS§	19.38%		14.73	97.14	2.10E-03	3.20E-04	15.22
DOR§	2.50%		13.52	97.37	8.25E-04	1.20E-04	14.55
PEC/AUTRES§	23.14%		**44.74**	**91.79**	2.70E-03	7.00E-04	**25.96**
Sérologie positive	36.24%	35.81	**44.18**	**91.88**	3.06E-03	7.89E-04	**25.78**
Sérologie négative	27.35%	60.96	12.54	97.55	2.93E-03	4.14E-04	14.11
Biomarqueur V3	37.78	34.14	**42.55**	**92.16**	1.54E-01	3.90E-02	**25.32**
Biomarqueur TM	28.00	34.14	33.24	**93.77**	1.18E-01	2.64E-02	**22.45**
DEUXIEME PHASE							
Sérotype B	60.82%	33.45	**37.51**	97.56	3.05E-03	4.59E-04	15.06%
Délais positifs	54.64	33.93	23.86	98.43	6.93E-01	8.36E-02	12.06%
Délais négatifs	33.71	44.90	**47.67**	96.92	4.46E-01	7.58E-02	16.98%
Taux de CD4	375.44	43.36	20.79	98.63	2.56E+00	2.89E-01	11.27%

Sont notées en caractères gras les valeurs de FMI supérieures à la proportion de données manquantes, d'efficacité statistique inférieures à 95% et d'erreur de Monte Carlo supérieures à 20%.
* Erreur de Monte Carlo
†HOMO/BI : mode homosexuel/bisexuel ; UDI : usage de drogue intraveineuse ; HETERO : mode hétérosexuel.
‡ PIV : primo-infection virale ; ASY : asymptomatique ; SNS : symptomatique non sida ; SID : sida.
§ SCB : symptômes cliniques et biologiques ; EXP : exposition ; BIL/GROS : bilan/grossesse ; DOR : dépistage orienté ; PEC/AUT : prise en charge/autres

- ***Fraction d'information manquante (FMI)***

La FMI est une fonction des variances inter et intra-imputation. Bodner [57] a proposé de l'approximer par la proportion de données manquantes. Nos résultats montrent que sa valeur peut en différer nettement, et peut être supérieure (valeur en gras) ou nettement inférieure à la proportion de données manquantes initiale.

Il a été suggéré par Bodner que la FMI peut être inférieure à la proportion de données manquantes si le mécanisme de données manquantes est de type MCAR [28;57]. Par ailleurs, la FMI étant une fonction des variances intra et inter-imputation, elle est un indicateur de la qualité de l'imputation, dépendant conjointement des propriétés du modèle d'imputation et des mécanismes induisant les données manquantes.

Le premier déterminant de la valeur de la FMI est l'effectif de chaque modalité pour les variables catégorielles, qui impacte la valeur de la variance intra-imputation. La variabilité est effectivement réduite si les effectifs de la modalité sont importants. Cela est nettement visible pour les modalités principales des variables pays de naissance (France et AFSS) et stade clinique (Asymptomatique). Ce critère ne permet cependant pas d'expliquer les variations observées entre la FMI et la proportion de données manquantes pour les autres variables.

Pour certaines modalités des variables discrètes, le processus de récolte de l'information pourrait être plus aisé que pour d'autres, rendant le mécanisme de données manquantes plus aléatoire car indépendant de facteurs extérieurs. Cela pourrait être le cas, par exemple, de la modalité "France" pour la variable pays de naissance (FMI=9.4%), ou des modalités "exposition" (FMI=1.9%), "bilan/grossesse" et "dépistage orienté" pour la variable motif de dépistage. La variable serneg contient des données manquantes qui ont été générées artificiellement d'où un mécanisme plus aléatoire (FMI=12.5%) que dans le cas de serpo (FMI=44.2%). Enfin, le taux de CD4 est manquant en partie car les anciennes fiches ne collectent pas cette information (FMI=21%).

Cependant, pour certaines variables, un autre déterminant doit être envisagé pour interpréter les écarts entre proportion de données manquantes et FMI. En effet, les biomarqueurs ont des données manquantes pour les mêmes individus et donc selon le même mécanisme, et cependant on observe seulement pour V3 une FMI plus élevée que la proportion de données manquantes. Les qualités du modèle d'imputation sont par ailleurs équivalentes pour les deux biomarqueurs (6 prédicteurs, 5 bases imputées). Une explication pourrait être liée à la distribution originelle du

biomarqueur V3, très éloignée de la normalité et avec une proportion élevée de données censurées à droite.

- ***Efficacité statistique relative***

L'efficacité statistique relative est une fonction de la FMI et du nombre de bases de données imputées.

Pour les variables imputées lors de la première phase, 5 bases de données ont été générées, ce qui revient à tolérer une perte d'efficacité statistique variant de 5 à 10 % par rapport à un nombre infini d'imputations. Pour cette première phase, l'efficacité statistique varie nettement en fonction de la FMI.

Ainsi, les performances sont meilleures, avec une efficacité statistique supérieure à 95%, lorsque la FMI est inférieure à la proportion de données manquantes, et inversement. La variable motif de dépistage illustre bien ce mécanisme puisque les valeurs extrêmes de la FMI sont obtenues pour deux modalités de cette même variable : la modalité "exposition" avec une efficacité de 99.6% (FMI=1.9%) et la modalité "prise en charge/autres" avec une efficacité de 91.8%.

Logiquement, l'effet de la valeur de la FMI sur l'efficacité statistique est moins sensible pour les variables imputées lors de la deuxième phase puisque les estimations sont cette fois pondérées par 15 bases au lieu de 5. Il en résulte que l'efficacité statistique est globalement meilleure, avec une valeur dépassant 95% pour les 4 variables imputées.

- ***Erreur de Monte Carlo***

L'erreur de Monte Carlo est un indicateur de la répétabilité des analyses. Pour un paramètre donné, le ratio de l'erreur de Monte Carlo et de l'écart type du paramètre doit être inférieur à 10% [28]. Les résultats montrent que ce ratio est très corrélé à la FMI pour les variables imputées au cours de la phase 1, avec des valeurs variant de 5.6% à 26%. Globalement, ce critère de répétabilité des analyses n'est pas rempli pour la première phase puisque sa valeur se situe aux environ de 20% pour la majorité des variables. Pour la phase 2, l'effet du plus grand nombre de bases imputées est là encore perceptible puisque le ratio varie de 12 à 17%, avec une corrélation toujours nette avec la FMI.

White et al. proposent une règle empirique à partir de l'interprétation de ce ratio qui implique d'imputer environ $M = 100 \times FMI$. Or nos analyses montrent qu'avec seulement 15 bases et une FMI de 48%, le ratio obtenu est 1.5 plus élevé que la valeur recherchée. En toute rigueur, il aurait donc fallu imputer 10 bases à la première phase et atteindre 30 bases globalement. Ce résultat était attendu, sachant que le choix du nombre de bases a dû être restreint pour des raisons pratiques.

7.1.5. Conclusion

Les examens diagnostiques proposés interviennent dans la phase finale de validation des données imputées mais ils constituent également une étape essentielle dans la construction du modèle d'imputation puisqu'aucun test d'adéquation du modèle d'imputation n'est à ce jour applicable en routine. Il est ainsi recommandé de tester plusieurs modèles d'imputation [34] et d'évaluer l'impact du choix du modèle sur la cohérence des résultats obtenus, ce qui peut constituer une forme simple d'analyse de sensibilité. Au vu de la complexité des modèles d'imputation élaborés ainsi que du type de variables, la méthode d'imputation multivariée normale apparaît moins adaptée et n'a pas été testée.

Plusieurs modèles ont été élaborés au cours du temps pour chacune des deux phases, et les étapes diagnostiques ont permis d'identifier des différences plus marquées entre valeurs observées et imputées pour certains d'entre eux. Pour la phase 1, une dizaine de modèles ont été construits. Les résultats obtenus pour les variables binaires sont restés stables selon les modèles. L'imputation des variables catégorielles a été nettement améliorée par la décomposition des variables à imputer ainsi que des variables prédictrices en variables indicatrices. Concernant les variables continues, la normalisation de la distribution des biomarqueurs s'est révélée problématique et la transformation par scores de quantile a permis d'obtenir des distributions de données observées et imputées nettement plus proches. Pour la phase 2, plusieurs modèles ont été élaborés, faisant varier la composition des équations de prédiction ainsi que la fonction de lien pour les variables délais positifs et taux de CD4, initialement imputées sous une forme discrète. De plus, l'examen des critères liés au nombre de bases imputées aurait pu mener à une modification du choix initial si le contexte d'imputation et d'analyse l'avait permis.

Cette étape de diagnostic permet de sélectionner les modèles d'imputation, de proposer des critères de validité interne du modèle d'imputation retenu ainsi que d'évaluer la robustesse des

résultats à l'influence d'un éventuel mécanisme MNAR. Cette étape permet ainsi de rechercher conjointement des incohérences dues à la spécification du modèle d'imputation et/ou des biais dans les estimations liées à des données manquantes selon un mécanisme MNAR. Donc, si la prise en compte des valeurs observées permet de supprimer des différences systématiques entre valeurs observées et imputées [23], l'impact de données manquantes de type MNAR sur les résultats de l'imputation apparaît négligeable, ce qui est le cas dans notre application.

Nous pensons qu'il est cependant indiqué d'appliquer des mesures de validation plus élaborées, internes par des études par simulation, et externes par comparaison avec des données extérieures.

7.2. Validation croisée

Nous avons effectué une étude de simulation afin d'évaluer les capacités prédictrices du modèle d'imputation de la première phase en termes de construction du modèle et de nombre de bases imputées.

7.2.1. Etude de simulation

Cette étude est inspirée d'une illustration de Nevalainen et al. [63]. Nous avons sélectionné la variable mode de contamination imputée au cours de la première phase, codée en 4 modalités (homo/bisexuel, UDI, hétérosexuel et autres) et contenant 29% de données manquantes. L'estimation d'intérêt est la proportion de la modalité "homosexuel/bisexuel". Pour cette étude, la base de données de départ est restreinte aux 29142 individus pour lesquels la variable mode de contamination est entièrement renseignée.

On réalise une simulation de Monte Carlo en répliquant 500 fois l'algorithme suivant :

- On tire un échantillon avec remise de 5000 individus
- On calcule la proportion observée (homosexuel/bisexuel)
- On génère des données manquantes (30%)
- On impute les données manquantes en utilisant le modèle construit pour la première phase, et 20 bases sont générées
- On compare la vraie proportion observée et la proportion estimée après imputation

Les données manquantes sont générées selon deux scénarios :

- MCAR avec 30% de données manquantes
- MAR dépendant de l'âge en 3 catégories (0-24 ans, 25-49 ans et 50-90 ans). Comme la distribution par classe d'âge varie selon les échantillons, on choisit de conserver dans chaque échantillon respectivement 550 individus pour les moins de 24 ans, 2425 individus pour les 25-49 ans et 525 individus pour les 50-90 ans, soit au total 3500 individus, et de générer des données manquantes pour 1500 individus. Ainsi, 30% de données manquantes sont générées globalement ce qui donne en moyenne pour chaque catégorie 15, 34 et 20% de données manquantes.

L'imputation est réalisée en appliquant le modèle construit lors de la première phase. La variable mode de contamination y est incluse sous une forme catégorielle. Les covariables incluses dans le modèle sont celles retenues pour la première phase. Elles peuvent contenir des données manquantes et sont donc à la fois imputées et prédictrices.

7.2.2. Résultats et conclusions

L'objectif de cette étude est de calculer, pour une variable donnée, trois statistiques :

- Le biais relatif défini comme :

$$BR = 100 \times \frac{moy(P) - P_{obs}}{P_{obs}}.$$

- La probabilité de couverture de l'intervalle de confiance correspondant à la proportion d'intervalles de confiance à 95% qui contiennent la vraie valeur de *P* parmi les 500 échantillons.

- L'efficacité statistique relative définie ici comme la moyenne de l'efficacité statistique sur les 500 échantillons.

Les résultats sont présentés selon les mécanismes MCAR et MAR, et pour 5 ou 20 bases (Figure 4.8, Tableau 4.10).

Figure 4.8 – Probabilité de couverture de l'IC, biais relatif et efficacité statistique relative pour des mécanismes MCAR et MAR, pour les 500 échantillons, et pour 5 ou 20 bases imputées

Sont représentés en rouge les intervalles de confiance à 95% ne contenant pas la vraie valeur de P.

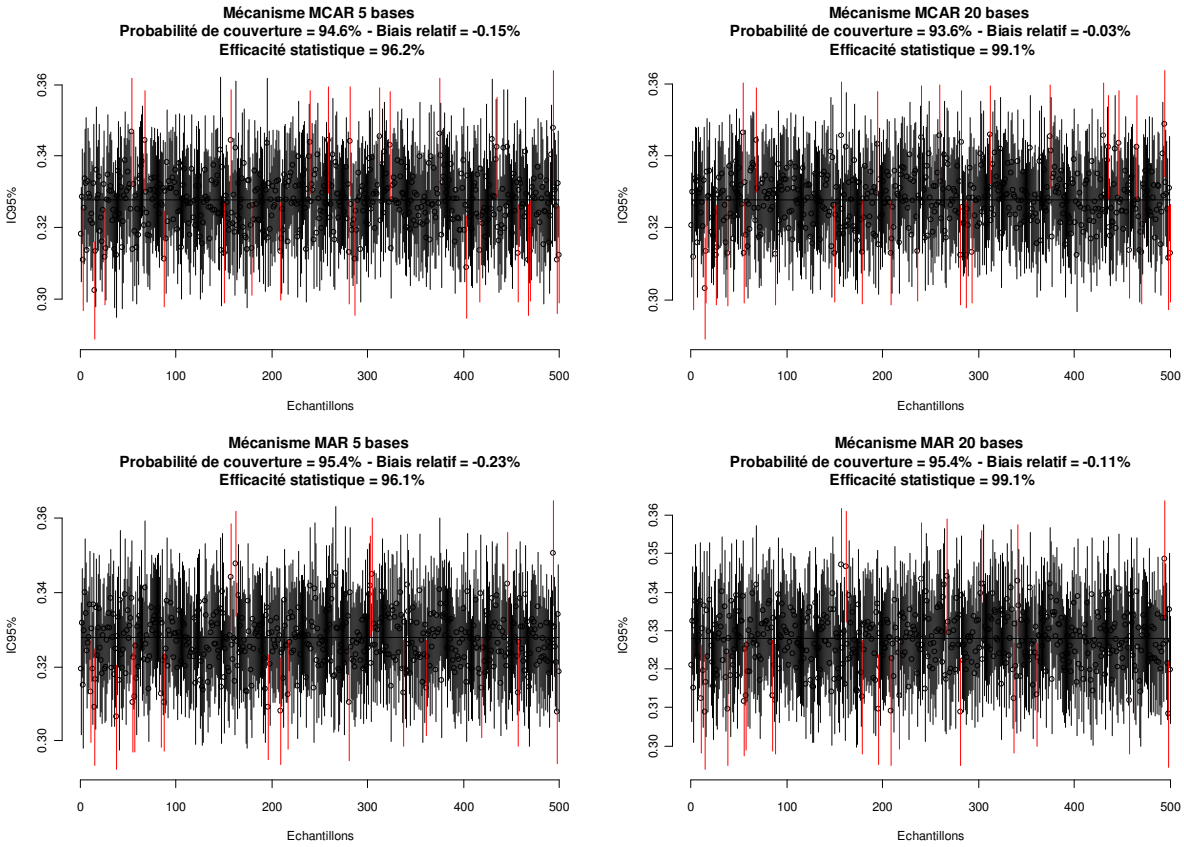


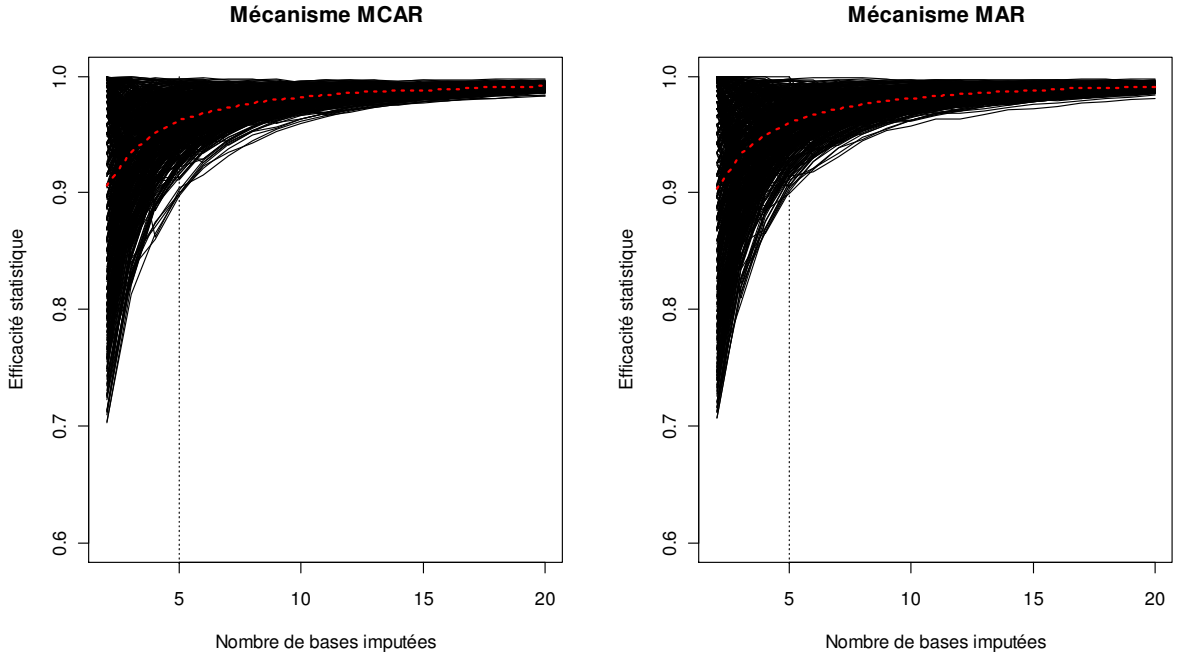
Tableau 4.10 – Synthèse des résultats de la Figure 4.8

	Biais relatif (%)	Probabilité de couverture de l'IC (%)	Efficacité statistique relative (%)
MCAR			
5 bases	-0.15	94.6	96.2
20 bases	-0.03	96.6	99.1
MAR			
5 bases	-0.23	95.4	96.1
20 bases	-0.11	95.4	99.1

Les résultats montrent que le biais relatif reste très modéré pour les 4 cas de figures. Cependant, on observe que ce biais relatif est plus limité respectivement pour un mécanisme MCAR et pour 20 bases. La probabilité de couverture reste proche de 95%, sans effet évident du mécanisme de données manquantes ou du nombre de bases. Par ailleurs, dans cet exemple, l'efficacité statistique est logiquement améliorée lorsque 20 bases sont générées, mais n'est pas sensible au mécanisme de données manquantes. Notons cependant que les mécanismes présentés restent de type aléatoire et que les estimations par imputation multiple doivent être non-biaisées, si le modèle d'imputation est correct.

Figure 4.9 – Représentation graphique de l'évolution de l'efficacité statistique relative selon le nombre de bases imputées, pour les 500 échantillons et pour les mécanismes MCAR et MAR

Est représentée en rouge la moyenne des efficacités statistiques des 500 échantillons



Pour chacun des 500 échantillons, l'efficacité statistique est calculée successivement pour un nombre global k de bases imputées, avec $k = 2,...20.$ De même, est représentée en ligne pointillée rouge la moyenne des efficacités statistiques des 500 échantillons, et ce pour chaque valeur de k.

Les graphes présentés en Figure 4.9 montrent que la variabilité de l'efficacité statistique par échantillon, importante pour $k = 2$ avec des valeurs variant entre 70 et 100%, est nettement réduite à partir de $k = 10$ et ne semble pas s'améliorer au-delà de $k = 15$. La moyenne de

l'efficacité statistique atteint la valeur de 95% dès $k = 4$. Si l'on considère le scénario comparable à celui retenu lors de l'imputation réelle, c'est-à-dire $k = 5$ et un mécanisme MAR, la perte moyenne d'efficacité statistique est inférieure à 4%, mais peut atteindre 10% pour certains échantillons.

En conclusion, cette étude de simulation montre que les valeurs du biais relatif et de la probabilité de couverture de l'intervalle de confiance se situent dans les valeurs attendues pour $k = 5$, et sont proches pour un mécanisme MCAR ou MAR. Concernant l'efficacité statistique, dont la valeur est plus dépendante du nombre de bases, les résultats montrent qu'il aurait été préférable d'imputer 20 bases. Cependant, il nous parait acceptable d'envisager une inflation de la variance des estimations de l'ordre de 4%, et pouvant aller jusqu'à 10%.

Cette analyse avait pour objectif de valider le modèle d'imputation de la première phase sur la base d'une variable catégorielle avec des modalités peu équilibrées et un mécanisme de données manquantes pouvant être partiellement MNAR. Les résultats de cette étude montrent que le modèle d'imputation de la première phase, bien que complexe, permet d'obtenir des estimations fiables sous l'hypothèse MAR.

8. Résultats comparés avec des données externes : essai de validation

Une étape de validation par des données externes avait été initialement prévue. L'objectif était de réaliser ce type de validation à partir des données d'une enquête spécifique pour laquelle un recueil différentiel aurait été effectué pour un sous-échantillon aléatoire de patients [141]. Une enquête hospitalière sur l'étude des co-infections VIH/hépatites avait été retenue pour cette étude de validation.

Celle-ci ayant été reportée, nous avons recherché des données d'enquêtes ou de cohortes recueillant des informations proches de celles de la déclaration obligatoire et avec un taux de complétude élevé. Nous avons finalement réalisé des comparaisons avec les données d'une des enquêtes Odyssée ainsi que de la cohorte ANRS Copana que nous allons maintenant décrire.

8.1. Sources de données

8.1.1. Enquête Odyssée

L'objectif de cette enquête nationale était d'estimer la prévalence de la résistance primaire aux antirétroviraux chez des patients chroniquement infectés et naïfs de traitement antirétroviral. Cette enquête, déjà réalisée en 2002, a inclus des patients entre novembre 2006 et mars 2007 (20 patients consécutifs par centre étaient sélectionnés). Au final 530 patients recrutés dans 31 centres ont été inclus, et 466 patients ont été retenus pour l'analyse [142].

Ces patients, infectés chroniques pour le VIH, avaient une durée entre le diagnostic de séropositivité et l'inclusion distribuée comme suit :

- > 6 mois : 26.2%
- 6 mois – 2 ans : 16.1%
- 2 – 5 ans : 19.7%
- > 5 ans : 8%

La population d'étude était donc a priori caractérisée par un recrutement plus tardif que pour la déclaration obligatoire.

8.1.2. Cohorte Copana

L'objectif de cette cohorte était d'étudier le pronostic à court, moyen et long terme des patients infectés par le VIH, récemment diagnostiqués et non traités par antirétroviraux [136]. Cette cohorte prospective multicentrique (36 centres) a inclus des patients depuis février 2004 et pendant une période de 4 ans. Ont été inclus des patients infectés par le VIH de type 1, dont la découverte de séropositivité était récente (datant de moins d'un an), âgés de 15 ans ou plus et non traités au moment de l'inclusion. La cohorte Copana a inclus au total 795 patients. Les patients dépistés au stade de primo-infection n'ont pas été inclus dans la cohorte Copana mais dans une cohorte dédiée, la cohorte Primo.

Les informations collectées ont été (i) des données cliniques et biologiques prévues dans le suivi des patients VIH, (ii) des données immuno-virologiques dont le taux de lymphocytes T4 et (iii) des données récoltées par des auto-questionnaires annuels portant sur les anomalies morphologiques, la consommation de tabac et d'alcool, l'alimentation et l'activité physique. Les données utilisées dans la comparaison avec les données de la DO sont issues du questionnaire administré à l'inclusion des patients, et ne tiennent pas compte des données recueillies annuellement.

8.2. Comparaison des sources de données

8.2.1. Méthodes

L'objectif initial de cette comparaison des données de la DO avec celles de deux autres sources de données était d'appliquer des tests statistiques à certaines estimations (proportions, moyennes, odds ratios). Cependant, deux problèmes méthodologiques se sont posés : (i) les échantillons comparés ne sont pas indépendants puisque la DO inclut en théorie tous les nouveaux diagnostics de VIH et donc les individus inclus dans Odyssée et Copana, et (ii) la taille des échantillons est très disproportionnée (DO : N=41049, Copana : N=795 et Odyssée : N=466). Des tests statistiques adaptés permettent de comparer des quantités calculées à partir d'échantillons qui ne sont pas indépendants, mais la taille de l'échantillon de la DO rend tout test statistique ininterprétable puisque systématiquement significatif.

Nous avons par ailleurs constitué une base de données unique à partir des trois bases de données (DO, Copana et Odyssée), restreinte aux variables communes aux trois bases et incluant une variable permettant d'identifier la base d'origine. Nous avons cherché à comparer les bases de données en construisant une régression logistique (2 bases) ou multinomiale (3 bases) expliquant la variable indicatrice de base de données par les covariables d'intérêt. Pour chaque covariable, un test de Wald ou de Fisher significatif mettrait en évidence une différence significative de la distribution en fonction des bases de données considérées. Les résultats obtenus n'ont pas été exploitables puisque les tests de Wald ou de Fisher étaient tous significatifs quand la régression multivariée impliquait la source DO, toujours en raison des effectifs.

Nous avons donc appliqué au final une méthode simple, consistant à comparer les proportions ou les moyennes des variables communes aux trois bases de données, en utilisant les intervalles de

confiance obtenus à partir des données imputées. On peut conclure à une différence non-significative pour une variable donnée lorsque le paramètre observé (Odyssée, Copana) appartient à l'intervalle de confiance du paramètre estimé (DO).

Notons que la proportion de données manquantes dans la base Odyssée est très faible, de l'ordre de 1% pour les variables d'intérêt, ainsi que pour Copana à l'exception de la variable mode de contamination qui contient 7% de données manquantes. Les résultats présentés ne tiennent pas compte des données manquantes.

8.2.2. Résultats

- ***Nouveaux diagnostics***

Les patients inclus dans la cohorte Copana sont récemment diagnostiqués, avec un délai d'inclusion variable et non-documenté, alors que pour l'enquête Odyssée, les patients inclus sont définis comme des malades chroniques du VIH, avec un diagnostic a priori plus ancien. Parmi les individus de la base de DO, certains ont déjà eu un test VIH positif, mais les analyses sont effectuées en restreignant la base de données aux individus correspondant à des découvertes de séropositivité, c'est-à-dire n'ayant pas eu de test antérieur positif ou un test antérieur positif datant de plus de 11 mois, soit environ 80% des effectifs.

Dans le cas d'Odyssée, si l'on reprend la définition de nouveau diagnostic appliquée aux patients de la DO, les données sur les délais entre test VIH et inclusion permettent d'estimer une proportion de "nouveaux diagnostics" variant de 60 à 74%, soit inférieure à celle de la DO. Pour la cohorte Copana, l'information de délai avec un éventuel test antérieur positif n'est pas disponible, mais seuls des patients récemment diagnostiqués sont inclus. Par ailleurs, on dispose pour Copana de l'information sur un éventuel test antérieur négatif et 60% des patients sont dans ce cas. Les délais en mois entre le test VIH de l'inclusion et un test antérieur ont une distribution proche de celle observée parmi les nouveaux diagnostics de la DO (délai moyen de 36 mois pour Copana et 33 mois pour la DO). Il est donc probable que l'inclusion dans Copana soit un peu plus tardive que pour la DO.

Ainsi, le délai entre diagnostic VIH et inclusion serait plus court pour la DO que pour Copana, et plus court pour Copana que pour Odyssée. Les analyses suivantes portent sur les nouveaux diagnostics de la DO, et sur les bases complètes pour Copana et Odyssée.

- ***Distribution d'âge***

Les distributions d'âge sont proches pour les trois sources de données lorsque l'on compare des critères quantitatifs (quantiles, moyenne, médiane), avec cependant un âge moyen plus bas pour Copana (Tableau 4.11). En effectuant une comparaison plus fine par classe d'âge, on constate que la classe d'âge 0-24 est sous-représentée dans Copana et de façon encore plus marquée dans Odyssée par rapport à la DO. Cela est dû à une non-inclusion des enfants de moins de 15 ans.

Tableau 4.11– Distribution d'âge pour les nouveaux diagnostics de la DO et les sources Copana et Odyssée

AGE	**DO** (N=15x41049)			**Copana** (N=795)	**Odyssée** (N=466)
Forme continue			**Quantiles**		
	Observés	Imputés		Observés	Observés
1%	17	16		20	21
25%	29	29		29	31
50%	**36**	**37**		**34**	**37**
75%	45	45		43	44
99%	69	70		69	65
Moyenne	37.64	38.04		36.84	37.85
Forme catégorielle			**Proportions (%)**		
	Observées	Imputées	IC 95%*	Observées	Observées
0-24 ans	13.55	**13.28**	12.9-13.7	9.31	5.58
25-34 ans	34.33	**34.18**	33.7-34.7	42.26	32.62
35-44 ans	29.25	**29.60**	29.1-30.1	27.67	37.34
45-54 ans	14.54	**14.77**	14.4-15.2	13.46	18.03
55-64 ans	6.53	**6.32**	6.1-6.6	5.91	5.36
65-90 ans	1.81	**1.84**	1.7-2.0	1.38	1.07
Selon le sexe					
Femmes					
0-24 ans	19.63	**18.82**	18.1-19.5	11.06	9.09
25-34 ans	40.74	**40.46**	39.6-41.4	45.96	41.32
35-44 ans	22.65	**23.67**	22.9-24.4	24.26	26.45
45-54 ans	10.64	**10.84**	10.3-11.4	9.79	16.53
55-64 ans	4.95	**4.76**	4.4-5.2	7.66	6.61
65-90 ans	1.39	**1.45**	1.2-1.7	1.28	0.00
Hommes					
0-24 ans	10.17	**10.09**	9.7-10.5	8.57	4.35
25-34 ans	30.76	**30.56**	29.9-31.2	40.71	29.57
35-44 ans	32.91	**33.03**	32.4-33.7	29.11	41.16
45-54 ans	16.70	**17.04**	16.5-17.6	15.00	18.55
55-64 ans	7.40	**7.22**	6.9-7.6	5.18	4.93
65-90 ans	2.05	**2.06**	1.9-2.3	1.43	1.45

* Intervalle de confiance à 95% des proportions imputées de la DO

Par ailleurs, la classe d'âge 25-34 ans est surreprésentée dans Copana par rapport à la DO et Odyssée, plus particulièrement chez les hommes. Pour Odyssée, c'est la classe 35-44 qui est surreprésentée par rapport aux deux autres sources, là aussi très nettement chez les hommes. Une hypothèse possible serait que ces différences soient dues à un biais de recrutement chez les hommes, avec une proportion plus élevée de patients homosexuels inclus dans Copana et Odyssée que dans la DO. Cependant, cette hypothèse n'est pas confirmée par les données de la DO qui montrent que, globalement, la proportion de patients contaminés par voie homosexuelle est plus élevée en consultation de ville qu'en structure hospitalière.

- ***Pays de naissance***

Si l'on compare les données observées et imputées de la DO à celles des autres sources, on observe que le recrutement des patients diffère selon les sources pour le pays de naissance (Tableau 4.12). Ainsi, la proportion de patients nés en France est plus élevée dans Copana que dans la DO, et elle s'accroit dans Odyssée par rapport à Copana. On observe un classement inverse entre les sources pour la proportion des patients nés en Afrique sub-Saharienne (AFSS).

En détaillant par sexe, les différences entre les proportions sont similaires pour Odyssée, alors que pour Copana des variations marquées apparaissent selon le sexe. Ainsi, la proportion de femmes nées en France est proche entre Copana et la DO (différence non significative), alors que la proportion d'hommes nés en France est plus élevée dans Copana que dans la DO. La proportion de femmes nées en AFSS est plus élevée dans Copana que dans la DO (68% versus 61%), alors qu'une relation inverse est observée pour les hommes nés en AFSS (19% versus 22%). Il est donc probable que Copana recrute bien des femmes originaires d'AFSS, sans doute en relation avec les mesures de dépistages dans le cadre de la transmission mère enfant (TME), alors qu'un biais de sélection est observé pour les hommes.

La répartition différentielle des patients de Copana et d'Odyssée selon le pays de naissance, visible à partir des données imputées mais aussi des données observées de la DO, permet de conclure que les deux sources de données retenues ne permettent pas de proposer une validation externe pour l'imputation. Il est cependant informatif de comparer les distributions des autres variables d'intérêt, tenant compte de cette différence établie entre les sources de données.

Tableau 4.12– Distribution de la variable pays de naissance pour les nouveaux diagnostics de la DO et les sources Copana et Odyssée

	DO (N=15x41049)		Proportions (%)	Copana (N=795)	Odyssée (N=466)
	Observées	Imputées	IC 95%*	Observées	Observées
PAYS DE NAISSANCE					
France	49.58	**47.84**	47.3-48.4	54.30	61.82
AFSS†	35.82	**36.81**	36.3-37.4	33.50	27.98
Europe	3.1	**3.25**	3.0-3.5	3.59	4.34
Amérique/Haïti	6.94	**7.32**	7.0-7.7	3.85	1.52
Autres	4.56	**4.78**	4.5-5.0	4.75	4.34
Selon le sexe					
Femmes					
France	23.66	**23.11**	22.2-24.0	22.94	33.88
AFSS†	61.90	**62.16**	61.3-63.0	68.40	58.68
Europe	1.72	**1.82**	1.5-2.1	2.16	4.13
Amérique/Haïti	8.84	**8.90**	8.2-9.6	2.60	0.00
Autres	3.88	**4.02**	3.6-4.4	3.90	3.31
Hommes					
France	63.92	**62.13**	61.3-62.9	67.52	71.76
AFSS†	21.40	**22.17**	21.5-22.8	18.80	17.06
Europe	3.86	**4.07**	3.7-4.4	4.20	4.41
Amérique/Haïti	5.89	**6.41**	6.1-6.8	4.38	2.06
Autres	4.93	**5.22**	4.9-5.5	5.11	4.71

* Intervalle de confiance à 95% des proportions imputées de la DO
† AFSS : Afrique sub-Saharienne

- ***Mode de contamination***

La proportion de mode de contamination par voir homo/bisexuelle est beaucoup plus élevée dans les deux sources extérieures, et plus particulièrement dans la source Copana (Tableau 4.13). Inversement, la modalité transmission par voie hétérosexuelle est sous-représentée dans Odyssée et surtout dans Copana. Ces relations sont retrouvées et même amplifiées après ajustement sur le sexe pour ces deux principaux modes de transmission.

Cette différence marquée entre Copana et la DO peut être partiellement expliquée par la répartition différentielle du pays de naissance, puisque le mode de contamination homosexuel est plus fréquent pour les patients nés en France que pour ceux nés en AFSS. Cependant, cet argument ne permet pas d'expliquer les différences de transmission par voie homo/bisexuelle (et parallèlement par voie hétérosexuelle) entre Copana et Odyssée puisqu'Odyssée inclut plus de patients nés en France que Copana et nettement moins de patients homo/bisexuels. L'ajustement

sur le pays de naissance ne permet donc pas d'expliquer entièrement les variations observées. Notons que la variable mode de contamination contient 7% de données manquantes dans la source Copana, ce qui peut impacter les différences entre les proportions.

L'examen selon les sources de la variable mode de contamination, variable considérée comme traitant d'un sujet sensible socialement, montre bien la difficulté d'obtenir des données de validation issues de sources extérieures. La différence pour cette variable entre proportions observées et estimées dans la DO est en effet bien moindre que la différence entre ces proportions et celles observées dans les autres sources.

- ***Stade clinique***

Il existe deux classifications pour décrire la progression de l'infection à VIH, basées sur les manifestations cliniques et les anomalies biologiques : la classification en 3 stades cliniques des Centers for Disease Control and Prevention (CDC) d'Atlanta et la classification en 4 stades cliniques proposée par l'Organisation Mondiale de la Santé (OMS) [85].

Le codage en 4 catégories n'était disponible que pour les sources DO et Copana. Une comparaison de proportions de ces deux sources selon ce codage montre que Copana inclut plus de patients à un stade de primo-infection virale ou asymptomatique que la DO, et donc nettement moins de patients à un stade symptomatique non-sida ou sida (Tableau 4.13).

Les proportions ont été recalculées à partir des stades cliniques définis par le CDC (A : primo-infection asymptomatique ou symptomatique, B : infection symptomatique non-sida, C : infection au stade sida). Les modalités "primo-infection virale" et "asymptomatique" ont été regroupées pour reformer la modalité A. Notons que celle-ci est largement majoritaire en effectifs pour les trois sources de données. On retrouve avec ce nouveau codage une proportion de patients au stade A plus élevée pour Odyssée et surtout pour Copana que pour la DO. Ainsi, bien que les patients d'Odyssée soient considérés comme des malades chroniques, ils sont globalement à un stade clinique moins avancé que les patients de la DO.

Notons que la définition de découverte de séropositivité utilisée pour discriminer les nouveaux et anciens diagnostics ne tient compte que de l'historique des tests positifs et non de l'ancienneté réelle de l'infection. Ainsi, les sous-échantillons de la DO constitués par l'enquête Odyssée et la cohorte Copana sélectionnent certainement, par leur mode de recrutement, des patients à un stade

clinique moins avancé que dans l'ensemble de la DO. Cela peut être partiellement dû au fait que les enquêtes hospitalières n'ont pas accès à des populations marginalisées (par exemple des populations de migrants en situation irrégulière, ou des personnes non-bénéficiaires de la CMU). L'inclusion dans les enquêtes hospitalières nécessite en effet d'obtenir le consentement du patient, et les personnes socialement fragiles peuvent ne pas souhaiter être incluses.

Tableau 4.13– Distribution des variables mode de contamination et stade clinique pour les nouveaux diagnostics de la DO et les sources Copana et Odyssée

	DO (N=15x41049)			**Copana** (N=795)	**Odyssée** (N=466)
			Proportions (%)		
	Oservées	Imputées	IC 95%*	Observées	Observées
MODE DE CONTAMINATION					
Homo/BI†	35.17	**33.14**	32.6-33.7	49.53	45.06
UDI†	1.72	**1.80**	1.6-2.0	0.41	1.93
Hétéro†	62.09	**63.72**	63.1-64.3	47.09	46.35
Autres	1.02	**1.34**	1.2-1.5	2.98	6.65
Selon le sexe					
Femmes					
UDI†	1.05	**1.06**	0.8-1.3	0.47	0.83
Hétéro†	97.51	**97.10**	96.7-97.5	94.37	90.91
Autres	1.44	**1.85**	1.6-2.1	5.16	8.26
Hommes					
Homo/BI†	54.86	**52.27**	51.5-53.1	69.58	60.87
UDI†	2.10	**2.23**	2.0-2.5	0.38	2.32
Hétéro†	42.27	**44.44**	43.7-45.2	27.95	30.72
Autres	0.78	**1.05**	0.9-1.2	2.09	6.09
STADE CLINIQUE					
PIV‡	10.16	**9.46**	9.1-9.9	13.08	
ASY‡	61.85	**61.25**	60.7-61.8	75.09	
SNS‡	13.61	**13.24**	12.8-13.7	4.15	
SID‡	14.38	**16.06**	15.6-16.5	7.67	
Stade CDC					
A	72.01	**70.71**	7.01-71.3	88.18	81.47
B	13.61	**13.24**	12.8-13.7	4.15	7.11
C	14.38	**16.06**	15.6-16.5	7.67	11.42

* Intervalle de confiance à 95% des proportions imputées de la DO

† HOMO/BI : mode homosexuel/bisexuel ; UDI : usage de drogue intraveineuse ; HETERO : mode hétérosexuel ;

‡ PIV : primo-infection virale ; ASY : asymptomatique ; SNS : symptomatique non sida ; SID : sida

- ***Taux de CD4***

La comparaison des valeurs du taux de CD4 en continu montre que les distributions ont des valeurs globalement plus élevées pour Copana et Odyssée, et proches pour ces deux sources (Tableau 4.14). Les distributions par quantiles montrent que le taux de CD4 est plus élevé en début de distribution pour Copana et Odyssée, puis que les valeurs sont proches pour les trois sources au-delà de la médiane.

La valeur du taux de lymphocytes T4 est étroitement liée à l'histoire naturelle de la maladie, et donc très corrélée au stade clinique, ce qui explique les différences entre la DO et les autres sources en début de distribution, liées à la proportion plus élevée de stade clinique A. Après ajustement sur le stade clinique du taux de CD4 selon un codage catégoriel, les proportions observées pour Copana et Odyssée sont relativement proches de celles de la DO, même si elles n'appartiennent pas aux intervalles de confiance des proportions estimées. Notons que cette comparaison n'a de sens que pour la catégorie A qui contient des effectifs suffisants pour Copana et Odyssée.

Tableau 4.14 - Distribution du taux de CD4 pour les nouveaux diagnostics de la DO et les sources Odyssée et Copana

	DO			Copana	Odyssée
TAUX DE CD4	(N=15x41049)			(N=795)	(N=466)
Forme continue			**Quantiles**		
	Observés	Imputés		Observés	Observés
1%	4	4		6	6
25%	170	170		238	260
50%	357	359		383	380
75%	542	541		546	520
99%	1242	1172		1177	1016
moyenne	384.92	383.71		410.75	398.18
Forme catégorielle			**Proportions (%)**		
	Observées	Imputées	IC 95%*	Observées	Observées
0-199	28.65	28.61	27.7-29.5	19.37	19.52
200-349	20.50	20.31	19.5-21.1	23.9	26.03
350-499	20.84	21.01	20.1-21.9	26.29	25.81
500 et +	30.00	30.08	29.1-31.0	30.44	28.63
Selon le stade clinique CDC					
A					
0-199	12.67	14.52	13.1-15.4	12.55	12.03
200-349	23.03	21.80	20.9-22.7	24.82	27.81
350-499	25.69	25.24	24.1-26.3	29.1	28.34
500 et +	38.62	38.43	37.2-39.6	33.52	31.82
B					
0-199	49.25	1.53	49.7-30.5	45.45	18.18
200-349	21.44	23.30	20.6-26.0	21.21	27.27
350-499	14.57	16.21	13.9-18.5	15.15	27.27
500 et +	14.74	13.82	11.7-15.9	18.18	27.27
C					
0-199	87.89	83.84	81.6-86.0	83.61	75.00
200-349	7.67	10.22	8.5-11.9	14.75	11.54
350-499	2.56	3.82	2.7-4.9	0	7.69
500 et +	1.88	2.12	1.3-3.0	1.64	5.77

* Intervalle de confiance à 95% des proportions imputées de la DO

8.2.3. Discussion

La phase de diagnostic et de validation interne permet de conclure que les résultats obtenus à partir des modèles d'imputation retenus pour chacune des deux phases sont stables puisque les variations entre données observées et imputées peuvent être en grande partie expliquées par des mécanismes de données manquantes de type MAR plus ou moins complexes.

Une limite de ce travail est que l'objectif initial de validation des résultats de l'imputation par des données externes n'a pu être atteint. Des exemples d'études transversales ou longitudinales montrent que, dans des circonstances particulières, il est parfois possible de recueillir des données complémentaires alors que l'imputation a été réalisée et d'obtenir ainsi un échantillon de validation. Par exemple, pour une enquête multicentrique européenne sur l'efficacité vaccinale du vaccin contre la grippe, un recueil complémentaire a pu être mené pour les données françaises [143]. Pour le système de surveillance du VIH, un retour au clinicien a lieu pour 40% des notifications mais il permet de compléter les fiches de notification et est réalisé avant l'imputation. En règle générale, ce type de validation n'est possible pour des données d'enquêtes que lorsqu'un recueil de données intensifié est programmé dès le protocole pour un sous-échantillon d'individus.

Il est possible de proposer deux axes d'amélioration de ce processus d'imputation des données de surveillance du VIH. Tout d'abord, nous avons dû, pour des raisons pratiques, imputer un nombre de bases insuffisant au vu des examens pré et post-imputation, puisqu'il aurait été préférable de générer au moins 30 bases de données. S'agissant d'un processus d'imputation pérenne, il est envisageable d'accroître le nombre de bases imputées au cours du temps. Par ailleurs, les variables binaires rapportant l'historique des tests sérologiques, serpo et serneg, ont fait l'objet d'un problème de codage notable au stade du remplissage des fiches papier. Les résultats du diagnostic montrent que la prise en compte de ce problème pour la variable serneg, en générant des données manquantes pour des fiches identifiées par des modèle prédictifs, améliore la qualité de l'imputation. Un traitement similaire de la variable serpo pourrait être envisagé. Ces problèmes de codages sont cependant limités dans le temps car corrigés depuis 2007. Une imputation conditionnelle tenant compte de la période devrait permettre à terme de mieux prendre en compte cette particularité pour les variables serpo et serneg.

Si l'on tient compte des limites de ce processus d'imputation appliqué à une importante base de données de surveillance, la question de la légitimité du choix d'analyser une base de données imputée versus une base de données incomplète mérite discussion. En effet, bien que des critères diagnostiques montrent la fiabilité globale de ce processus d'imputation multiple, ils en soulignent également les limites. Ainsi, l'estimation des données manquantes de la base de surveillance du VIH permet d'effectuer des analyses à partir d'une base de données complète pour les variables clés, ce qui est essentiel pour le calcul de l'incidence, mais impacte forcément les estimations par rapport aux analyses effectuée à partir de la base de données incomplète. L'objectif est d'obtenir des variations entre les données initiales et les données complètes qui intègrent le plus d'informations possible, ce qui constitue une amélioration par rapport à un processus de réaffectation proportionnelle. Cependant, ce processus d'imputation reste complexe, bien qu'il ait été mis en place progressivement et amélioré au fil du temps, et l'imputation de certaines variables peut et doit être améliorée. Au final, le choix d'utiliser une base imputée pour l'analyse des données de surveillance du VIH nous apparaît justifié puisque, malgré ses limites, la méthode d'imputation multiple permet d'intégrer dans les estimations toutes les informations de la base de données, ce qui représente un atout indéniable par rapport à l'analyse de la base incomplète.

Conclusion

La mise en place d'un processus d'imputation multiple pérenne tel que celui-ci nécessite un investissement en temps important, pour réaliser l'imputation d'abord mais aussi pour l'analyse des données de surveillance. Ce projet s'est inscrit dans un travail de recherche et la question de son devenir peut donc se poser puisqu'un soutien statistique très spécifique sera nécessaire pour adapter les modèles d'imputation à l'évolution naturelle du système de surveillance du VIH. L'exigence méthodologique de l'équipe de surveillance ainsi que sa forte implication dans ce projet constitueront donc des atouts essentiels pour la poursuite de ce projet.

CHAPITRE 5

SYNTHESE ET PERSPECTIVES

Nous présentons dans ce chapitre une synthèse de notre travail de recherche, en proposant pour chaque thème abordé des perspectives de recherches potentielles.

L'impact des données manquantes sur les résultats d'enquêtes épidémiologiques est encore largement sous-évalué. Ainsi, l'approche la plus commune consiste à décrire les particularités des individus présentant des variables incomplètes afin d'établir s'ils diffèrent du reste de l'échantillon, puis à restreindre l'analyse aux individus pour lesquels toutes les variables sont renseignées. Cette méthode d'analyse dite analyse cas-complet est appliquée automatiquement par les logiciels d'analyse statistique standards et, si l'impact des données manquantes sur les effectifs est modéré, ce choix est la plupart du temps considéré comme raisonnable. Cependant, l'examen des caractéristiques des non-répondants ne permet pas d'établir si l'analyse cas-complet risque ou non d'introduire des biais. En effet, si des différences sont mises en évidence entre les répondants et les non-répondants, cela signifie simplement que le mécanisme de données manquantes global est de type MAR, sans préciser s'il dépend de la variable à expliquer dans le cas d'une étude étiologique.

Il nous a donc paru important, sachant que ce sujet est peu abordé dans la littérature, de présenter les mécanismes pour lesquels une analyse cas-complet donne des estimations valides. Pour cela, nous avons proposé dans le chapitre 1 une procédure standardisée permettant de réaliser l'examen d'une base de données incomplète. Cet examen permet de préciser le schéma de données manquantes pour l'ensemble de la base de données, ainsi que le mécanisme de données manquantes pour chaque variable incomplète. De cet examen peuvent être déduit la perte de puissance attendue en analyse cas-complet et l'impact sur la sélection des variables lors d'une analyse multivariée. Ainsi, dans le cas d'une analyse étiologique, il est possible que certains facteurs de risques et/ou de confusion soient omis du modèle final du fait de la perte d'effectifs due aux données manquantes. Il faut d'ailleurs noter que ce processus de sélection des variables

peut également impacter des variables originellement complètes, du fait de l'effet cumulé sur les effectifs des données manquantes des variables incomplètes. D'autre part, il est possible de déduire de l'examen de la base de données incomplète le risque de biais des estimations en analyse multivariée. Ainsi, seul un mécanisme MAR ou MNAR dépendant conjointement d'une variable d'exposition et de la variable à expliquer peut induire des biais en analyse cas-complet, alors qu'un mécanisme de type MNAR indépendant de la variale à expliquer permet d'obtenir des estimations non biaisées. Une analyse par simulation reprenant un schéma d'étude étiologique simple avec trois variables (maladie, exposition, confusion) nous a permis de montrer que les situations théoriques présentées donnaient les résultats attendus à partir de données simulées. Nous avons également présenté la méthode d'imputation multiple retenue pour l'ensemble de ce travail et précisé les raisons de ce choix. La méthode d'imputation multiple par équations chaînées se révèle être une méthode fiable tout en étant d'une utilisation abordable, même si ses bases théoriques peuvent être discutées. Par ailleurs, il est nécessaire d'appliquer avec prudence ce type de méthode d'estimation des données manquantes, sous peine d'obtenir des résultats biaisés ou imprécis. Ainsi, nous avons détaillé les étapes d'élaboration d'un modèle d'imputation, aussi bien en termes de sélection des variables prédictrices que de choix du nombre de bases à imputer. Nous avons également présenté des critères diagnostiques simples permettant de juger de la validité du modèle d'imputation. De plus, l'examen du mécanisme de données manquantes est informatif puisque l'imputation multiple produit des estimations biaisées si les données sont de type MNAR, sachant que, pour des analyses étiologiques, l'amplitude des biais peut être plus marquée si le mécanisme dépend de la variable à expliquer.

Dans le chapitre 2, nous avons décliné les règles d'application de l'imputation multiple présentées dans le chapitre 1 à partir de données d'enquêtes et de surveillance. Nous avons traité les données manquantes dans trois études de schéma différent : une étude évaluant le risque de transmission du VIH par don de sang, une étude cas-témoins sur l'infection à Campylobacter et une étude capture-recapture estimant le nombre de nouveaux diagnostics VIH chez les enfants. Pour chacune de ces trois études, une proportion de données manquantes non-négligeable était considérée (30%).

Dans l'étude estimant le risque de transmission du VIH par don de sang ainsi que dans l'étude capture-recapture évaluant le nombre de nouveaux diagnostics VIH chez les enfants, une seule variable était incomplète, et une méthode de réaffectation proportionnelle avait été appliquée

dans un premier temps. Cependant, cette méthode ne prenait pas en compte toute l'information de la base de données, et fournissait une variance sous-estimée pour les estimations. Pour ces deux études, l'élaboration d'un modèle d'imputation a permis d'inclure toutes les variables prédictrices identifiées de la variable incomplète. Puisque les variables incluses dans les modèles d'analyse sont utilisées pour estimer les variables incomplètes, cela permet de contrôler, au moins en partie, les biais attendus avec une méthode de remplacement simple. De plus, alors que le mécanisme de données manquantes identifié était de type MAR pour la variable pays de naissance dans l'étude capture-recapture, un mécanisme de type MNAR a dû être envisagé pour la variable mode de contamination dans l'étude d'analyse de risque de transmission du VIH par transfusion. Il nous a donc paru important de modéliser cette variable mode de contamination en incluant le plus de variables prédictrices possible afin de rendre l'hypothèse MAR plus plausible. Pour ces deux études, les résultats de l'imputation peuvent impacter les résultats finaux des analyses. Ainsi, la variable pays de naissance est une variable d'hétérogénéité de capture potentielle. Elle est incluse dans les modèles log-linéaires permettant de sélectionner les interactions pertinentes entre les sources et les variables d'hétérogénéité de capture, et les estimations finales sont stratifiées selon les modalités de cette variable. De même, la proportion de HSH estimée lors de l'imputation de la variable mode de contamination est appliquée à l'estimation de l'incidence parmi les HSH ainsi qu'à l'estimation du risque résiduel de transmission du VIH attribuable à cette population à risque. D'un point de vue méthodologique, l'élaboration du modèle d'imputation s'est révélée simple dans les deux études. Pour l'étude d'analyse de risque de transmission du VIH par don de sang, l'étape de diagnostic était déterminante puisqu'un mécanisme de type MNAR était suspecté. Des comparaisons simples des données observées et imputées ont permis de valider les estimations.

Pour l'étude capture-recapture, l'intérêt méthodologique a résidé dans l'analyse combinée des bases imputées par des modèles log-linéaires. Une restructuration des bases de données imputées a été nécessaire pour obtenir les tables de contingence (c'est-à-dire les effectifs communs selon les sources et les variables d'hétérogénéité), puis des modèles log-linéaires ont été ajustés à l'ensemble des bases imputées. Les estimations finales et leurs variances ont été obtenues directement par des commandes spécifiques à l'imputation. Cependant, la sélection du modèle final a soulevé un problème méthodologique puisque celle-ci repose habituellement sur des statistiques calculées à partir de la vraisemblance de chaque modèle. Dans une étude capture-recapture, le test d'adéquation aux données habituellement réalisé est le test de la déviance, qui

correspond à un test de rapport de vraisemblances entre le modèle testé et le modèle saturé. Un mode de calcul spécifique aux données imputées, proposé par Meng et Rubin, a donc été appliqué. L'imputation de données manquantes dans des applications de la méthode capture-recapture a été rapportée dans quelques études seulement, dans lesquelles les auteurs avaient appliqué une méthode d'estimation par maximisation de la vraisemblance en utilisant un algorithme EM. A notre connaissance, une seule étude a présenté une approche par imputation multiple dans une étude capture-recapture, et les critères de post-estimation (AIC, BIC, déviance) avaient été estimés à partir d'une seule base de données imputée. Une perspective de recherche méthodologique consisterait donc à proposer un mode de calcul adapté à la présence de bases multiples des critères AIC et BIC classiquement utilisés pour la sélection du modèle. Par ailleurs, il est recommandé d'incorporer dans le calcul de la variance des estimations une composante rendant compte de l'incertitude liée au choix du modèle. Cela nécessite de simuler des échantillons par bootstrap non-paramétrique, ce qui parait difficile à partir d'un ensemble de bases imputées. Une perspective de recherche serait donc d'élaborer un mode de calcul de la variance à partir de bases imputées afin de tenir compte de cette incertitude.

Dans l'étude cas-témoins appariée sur les facteurs de risque d'infection à Campylobacter, la problématique des données manquantes était cruciale puisque la plupart des 28 variables étaient incomplètes selon un schéma arbitraire. La perte d'effectifs en analyse multivariée conditionnelle cas-complet était donc très importante et une stratégie d'analyse par des sous-modèles a été nécessaire pour obtenir des estimations en analyse cas-complet. Le recodage de la catégorie données manquantes en catégorie additionnelle avait été appliqué initialement mais il est établi que cette méthode donne toujours des résultats biaisés. L'intérêt méthodologique de cette étude a résidé dans la stratégie de sélection des variables lors de l'élaboration des équations de prédiction. Cette stratégie n'était pas alors clairement définie dans la littérature et nous avons proposé un processus de sélection des variables explorant les liens entre les indicatrices de données manquantes et les variables utilisables dans la base de données. Sur la base de critères proposés dans la littérature, nous avons retenu un modèle incluant des variables dites principales et auxiliaires. Cependant, ces critères de sélection des variables ne sont pas clairement établis. Les variables principales ont été retenues sur la base d'une analyse univariée cas-complet, puisque ces variables constituent le modèle d'analyse multivarié en analyse cas-complet et après imputation multiple. Les variables auxiliaires ont été sélectionnées en fonction de leurs capacités prédictrices pour les variables principales. La stratégie appliquée dans cette étude a consisté à

examiner en analyse univariée les liens entre les indicatrices de données manquantes des variables principales et les variables disponibles dans la base de données.

Cet examen gagnerait à être effectué en analyse multivariée puisque, lors du processus d'imputation, chaque variable est modélisée à partir de toutes les autres variables du modèle d'imputation. Il faudrait donc construire pour chaque variable incomplète une régression logistique expliquant son indicatrice de données manquantes par toutes les covariables retenues. Cette stratégie n'est pas applicable dans cette étude du fait de la proportion élevée de données manquantes. Dans la littérature récente, il a été proposé d'ajuster ces régressions logistiques multivariées à partir de données imputées. Cette stratégie consiste à identifier des variables prédictrices à partir de données qui ont été imputées à partir de ces mêmes prédicteurs et peut de ce fait paraître circulaire, mais elle repose sur le postulat qu'une imputation multiple valide préserve les relations entre les variables telles qu'elles existent dans la base de départ incomplète. Une perspective de recherche dans ce domaine pourrait consister à identifier une procédure de sélection des variables prédictrices à partir de données imputées. Cette procédure inclurait le mode de construction du modèle pour l'imputation de départ, le choix du nombre de bases à utiliser et les critères statistiques retenus pour la sélection des variables prédictrices. Il faudrait donc imputer sous un modèle le plus général possible une base ou un ensemble de bases de données permettant d'effectuer cette sélection. La stratégie la plus appropriée reste donc à définir.

Nous avons abordé dans le chapitre 3 le traitement des données manquantes dans un système de surveillance, le réseau des pôles de référence de l'hépatite C. Le processus d'imputation présenté ne concernait pas l'ensemble des données du système de surveillance, mais seulement un échantillon d'usagers de drogues positifs pour le VHC. L'objectif de l'étude était d'identifier les facteurs de risque de complications hépatiques graves pour la période 2001-2007. Les systèmes de surveillance des maladies infectieuses se caractérisent par le grand nombre de variables recueillies et, en relation avec le mode de recueil de données, par une proportion importante de données manquantes. Dans le cas de cette étude étiologique, l'échantillon a été réduit aux patients pour lesquels la variable à expliquer (complications hépatiques graves) était entièrement renseignée. Par ailleurs, les variables retenues pour l'analyse étaient majoritairement incomplètes, avec une proportion de données manquantes élevée. Malgré la perte d'effectifs, une analyse cas-complet avait pu être réalisée. Lors de la construction du modèle d'imputation, seules des

variables principales ont pu été incluses, puisqu'il n'a pas été possible d'identifier des variables auxiliaires permettant d'enrichir le modèle d'imputation.

Ce cas de figure est problématique lorsque l'hypothèse d'un mécanisme de données manquantes de type MNAR doit être envisagée. Or, c'est le cas dans cette étude puisque, pour les variables "historique de consommation excessive d'alcool" et "co-infection par le VIH", il est possible que la non-réponse dépende des valeurs non-observées de ces variables. Nous avons ainsi posé l'hypothèse que la probabilité d'avoir des données observées est moindre pour la variable consommation excessive d'alcool chez les gros consommateurs d'alcool, et pour la variable co-infection par le VIH chez les patients négatifs pour le VIH. Nous avons donc fait le choix d'appliquer une analyse de sensibilité par pondération proposée par Carpenter et al. Cette méthode consiste à pondérer les estimations obtenues par imputation multiple (OR et sa variance) de façon à simuler l'impact d'un mécanisme MNAR modéré et à mesurer l'étendue du biais induit par ce mécanisme MNAR sur les estimations. Les poids utilisés sont calculés en intégrant un paramètre de sensibilité à la somme des valeurs imputées de chaque base de données. Notre objectif était d'appliquer cette analyse de sensibilité aux variables consommation excessive d'alcool et co-infection par le VIH en lien avec les hypothèses épidémiologiques posées, ainsi qu'à la variable infection par un virus de l'hépatite C de génotype 3 en raison de sa proportion élevée de données manquantes.

L'intérêt méthodologique de cette étude a résidé dans le processus de sélection du paramètre de sensibilité delta. Nous avons donc proposé de décomposer ce processus de sélection en 4 étapes de détermination graphique et de diagnostic de ce paramètre de sensibilité, et identifié une valeur de delta pour chacune des trois variables. Ce processus de sélection inclut le choix du signe de delta selon la relation supposée entre la probabilité d'avoir des données manquantes pour une variable et les valeurs de cette variable. Puis, l'estimation de l'OR sous un mécanisme MNAR a été obtenue et comparée avec l'estimation MAR pour chacune des variables testées. L'interprétation des résultats est conditionnée par le mécanisme de données manquantes identifié par régression logistique de chaque variable indicatrice de réponse à partir des autres covariables. En effet, sous un mécanisme de données manquantes de type MNAR, on s'attend à observer un biais plus important lorsque ce mécanisme dépend de la variable à expliquer. C'est le cas pour les variables consommation d'alcool et co-infection par le VIH. Les résultats montrent qu'un mécanisme MNAR impacte probablement les résultats de l'analyse réalisée sous l'hypothèse

MAR pour la variable co-infection par le VIH, et que l'association entre cette variable et la variable à expliquer est en réalité plus forte que ce qui est observé dans l'analyse sous l'hypothèse MAR. Cette méthode d'analyse de sensibilité permet donc de tester à partir de données imputées l'effet d'un mécanisme MNAR potentiel sur les estimations obtenues sous l'hypothèse MAR.

Des perspectives de développement de ce travail peuvent être proposées. L'analyse de sensibilité proposée par Carpenter et al. est appliquée à une seule variable binaire. Nous avons testé dans ce travail trois variables binaires, en déterminant alternativement pour chacune de ces variables le paramètre de sensibilité delta et l'estimateur MNAR. Une perspective intéressante consisterait à tester plusieurs variables simultanément, sachant que cela nécessiterait une adaptation de la méthode. Par ailleurs, cette méthode d'analyse de sensibilité a été appliquée à l'ensemble de l'échantillon, ce qui revient à poser une hypothèse d'indépendance entre les variables, en particulier entre chacune des variables et la variable à expliquer. En effet, pour modéliser le mécanisme de données manquantes, nous supposons un modèle paramétrique (une régression logistique) liant la probabilité d'observer une variable incomplète à la valeur de cette variable, ajustée sur un vecteur de covariables. Il serait donc justifié d'appliquer cette analyse de sensibilité aux deux sous-échantillons correspondant aux deux modalités de la variable à expliquer (complications hépatiques graves), et de calculer ainsi deux estimateurs MNAR.

Le chapitre 4 présente une application de l'imputation multiple à des données de surveillance du VIH, mais avec cette fois une approche beaucoup plus globale. Jusqu'en 2008, les données manquantes des variables incomplètes étaient présentées comme un catégorie spécifique, et une réaffectation proportionnelle avait été appliquée pour des analyses ponctuelles. La communication globale sur le VIH était en pratique réalisée à partir de données brutes incomplètes. Les objectifs de notre travail de recherche dans ce domaine étaient de proposer et d'évaluer un processus d'imputation multiple permettant de produire les estimations annuelles nécessaires à la surveillance du VIH, c'est-à-dire les analyses descriptives des personnes nouvellement diagnostiquées pour le VIH. De plus, un travail avait été initié afin d'estimer l'incidence du VIH, indicateur majeur de la surveillance du VIH, à partir des données de la déclaration obligatoire. Ce travail a été mené à bien par Stéphane Le Vu dans le cadre de sa thèse de doctorat [71]. La méthode retenue pour cette estimation était celle appliquée par les CDC américains, reposant sur la caractérisation, par un test biologique, d'individus récemment infectés

parmi une population de personnes séropositives. Notre objectif a donc été de compléter les variables utilisées dans les analyses descriptives, mais aussi d'estimer les valeurs manquantes des variables supplémentaires liées à l'estimation de l'incidence du VIH, tels que les biomarqueurs du test d'infection récente.

Un objectif méthodologique a été d'élaborer un processus d'imputation en deux phases, tenant compte de la présence de variables filtres pour la description de l'historique des sérologies positives et négatives. Nous avons donc construit deux modèles d'imputation, et un processus de sélection des variables principales et auxiliaires a été réalisé pour chacun des modèles à partir de données imputées. De ce fait, un modèle très général est spécifié pour le processus d'imputation de la première phase, et les résultats de cette imputation sont exploités pour la sélection des variables. Pour la construction du modèle de la deuxième phase, le processus de sélection des variables principales et auxiliaires est plus direct puisque les résultats de l'imputation de la première phase ont été utilisés. Par ailleurs, lors de l'étape de validation interne, un retour aux données observées est recommandé afin de vérifier la cohérence des données imputées.

Il est important de noter que l'objectif de l'imputation n'est pas de produire des données calquées sur les données observées, mais qu'elle vise à prendre en compte toute l'information contenue dans les variables prédictrices retenues. Ainsi, une différence entre données observées et imputées est attendue, mais elle doit être explicable par le biais des mécanismes de données manquantes identifiés lors de l'examen préalable de la base de données incomplète. Nous avons cherché à comparer, par des tests statistiques, les proportions observées et estimées pour les variables discrètes, ainsi que les distributions de données observées et imputées pour les variables continues. Cependant, et ce problème méthodologique a déjà été soulevé dans ce type d'étude, l'importance des effectifs fausse les résultats de ces tests puisqu'ils sont systématiquement significatifs, et ce même si les valeurs observées et imputées sont très proches. Le diagnostic a donc été réalisé par des comparaisons simples tenant compte des valeurs centrales des estimations (proportion, moyenne) et de leur variance. Notons que cette phase essentielle de diagnostic n'est que rarement présentée dans la littérature sur l'imputation multiple, qu'elle constitue un domaine de recherche actuelle, et qu'elle n'est à ce jour pas implémentée dans les logiciels statistiques.

Cette étape de diagnostic a été complétée par une étape de validation croisée explorant, pour une modalité d'une variable catégorielle, la validité du modèle d'imputation. La modalité retenue est le mode de contamination par voie homosexuelle, car nous avions fait l'hypothèse que le

mécanisme de données manquantes pour cette modalité pouvait être de type MNAR. Cette étude de simulation permet de valider le modèle d'imputation pour la variable retenue, mais pas d'évaluer la pertinence de l'hypothèse MAR. Afin d'estimer l'impact d'un mécanisme MNAR pour la variable mode de contamination, nous avions prévu d'effectuer un recueil de données complémentaire fournissant un échantillon de validation. Nous avions envisagé de sélectionner, lors d'une enquête hospitalière, un sous-échantillon aléatoire de patients pour lequel le recueil de données aurait été intensifié afin d'obtenir des données complètes. Cet échantillon aurait permis de valider l'imputation effectuée sur l'ensemble de la base de données. Cette enquête a été reportée et nous avons fait le choix de comparer les données imputées du système de déclaration obligatoire du VIH à d'autres sources de données, sélectionnant des patients ayant des caractéristiques proches de ceux de la DO : une enquête Odyssée et la cohorte COPANA. Cependant, la comparaison des données observées de la DO et de ces deux sources de données extérieures a montré que les caractéristiques des patients différaient nettement, et que ces sources de données ne pouvaient être considérées pour valider le processus d'imputation des données de la DO. Les comparaisons entre les données de la DO et celles d'Odyssée et COPANA nous ont cependant paru intéressantes pour explorer la 'représentativité' des deux études plutôt que pour valider le processus d'imputation de la DO. Nous avons donc retenu comme perspective de recherche l'application d'une analyse de sensibilité abordable aux variables pour lesquelles un mécanisme MNAR est suspecté. L'analyse de sensibilité détaillée dans le chapitre 3 ne peut être appliquée aux variables de la DO sans adaptation de la méthode car (i) elle nécessite un grand nombre de réplications de la base de données initiale (500 à 1000) et (ii) elle est adaptée à des données d'étude de type étiologique et non descriptive. Notre objectif initial était de l'appliquer à la modalité homosexuel/bisexuel de la variable mode de contamination. Nous retenons donc comme perspective de recherche l'application d'une analyse de sensibilité après imputation aux données de surveillance, tenant compte des questions méthodologiques soulevées.

L'étape de diagnostic montre que le processus d'imputation multiple pourrait être significativement amélioré. Les variables sérologie antérieure négative et positive sont impactées par un problème de codage ancien. La stratégie de recodage a concerné seulement la variable sérologie négative antérieure, car c'est une variable clé dans l'estimation de l'incidence du VIH, et qu'elle a été plus affectée par les problèmes de codage que la variable sérologie positive antérieure. Cependant, cette variable sérologie positive antérieure est essentielle pour identifier les personnes nouvellement diagnostiquées pour le VIH, sachant que les analyses descriptives

portent uniquement sur cette population. Il parait donc important d'améliorer la qualité de l'imputation de cette variable, en envisageant une stratégie de recodage similaire à celle appliquée à la variable sérologie antérieure négative. Une autre perspective d'amélioration de l'imputation consisterait à accroitre le nombre de bases de données imputées, surtout pour la première phase d'imputation, afin d'améliorer les estimations en termes de précision (diminution de la variance inter-bases) et de puissance statistique (amélioration de la couverture des intervalles de confiance). Il apparait ainsi pertinent d'imputer 10 bases au cours de la première phase, tout en générant 3 bases au cours de la deuxième phase. Cependant, la faisabilité en terme de taille de fichier finale doit être explorée.

Même si l'apport méthodologique de cette étape d'estimation des données manquantes est incontestable, et que le processus apparaît globalement valide, cette étape d'imputation multiple impacte les résultats de toutes les analyses réalisées sur les données de surveillance du VIH. La complexité de la mise en œuvre de cette méthode, aussi bien en termes d'implémentation que d'analyse pour l'équipe de surveillance, soulève la question de la légitimité mais aussi de la viabilité d'un tel processus. Dans un scénario où, tel que l'envisageait Rubin aux débuts de l'imputation multiple, les équipes d'imputation et d'analyse représentent des entités distinctes, une étroite collaboration apparaît nécessaire pour garantir la qualité des données produites et la pérennisation du processus d'imputation.

Conclusion

Ce travail de thèse a permis d'aborder un sujet qui suscite un vif intérêt et de nombreuses interrogations, le traitement des données manquantes, tout en n'étant pas ou mal pris en compte dans la grande majorité des enquêtes et des systèmes de surveillance en épidémiologie. Si la mise en application d'une méthode performante d'estimation des données manquantes, l'imputation multiple, est devenue accessible à des utilisateurs non-experts grâce à une évolution rapide des logiciels statistiques dans ce domaine, cette simplicité apparente ne doit pas faire oublier toutes les questions méthodologiques qui restent encore en suspens.

Les études présentées nous ont permis de balayer le champ des applications pratiques de l'imputation multiple à des données de type transversal et de synthétiser une stratégie globale de traitement des données manquantes. Ce processus de traitement des données manquantes s'échelonne des étapes d'examen de la base de données à la construction du modèle d'imputation puis aux nécessaires étapes de validation des modèles et de vérification des hypothèses. De

nombreuses questions méthodologiques se sont posées au fil de ces études et nous avons cherché à proposer des solutions abordables et valides pour les traiter. Certaines problématiques méritent certainement d'être approfondies, tout particulièrement dans le domaine de la post-estimation et de la validation des résultats.

Enfin, deux recommandations nous apparaissent importantes.

Un objectif crucial serait de pouvoir planifier le traitement des données manquantes dès l'élaboration du protocole d'une enquête et de se procurer ainsi un sous-échantillon aléatoire de validation du processus d'imputation, composé de données complètes. Pour les données de surveillance, l'obtention de données de validation demeure à ce jour irréaliste, en particulier pour les maladies pour lesquelles un processus d'anonymisation n'autorise pas un retour efficace aux données.

Dans le domaine de l'épidémiologie d'observation, même si les études incluant une étape d'imputation se multiplient, les publications qui en découlent ne détaillent que rarement les points clés du processus d'imputation. Notons que, dans le domaine des essais cliniques, l'imputation multiple a été appliquée pour traiter les données manquantes de façon précoce et que des règles d'application et de valorisation claires ont été édictées. Il apparaît donc important que, dans le champ de l'épidémiologie d'observation, de telles recommandations soient développées de manière collégiale et plus largement diffusées, afin que les équipes de recherche en tiennent compte, tant dans la mise en œuvre de l'imputation que lors de la valorisation des résultats.

BIBLIOGRAPHIE

1. Vach W, Blettner M. Missing data in epidemiological studies. In: Armitrage P, Colton T. Encyclopedia of Biostatistics.Chichester: Wiley, 1998. p. 2641-54.

2. Van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Stat Methods Med Res 2007;16(3):219-42.

3. Little RJA, Rubin DB. Statistical analysis with missing data. Wiley series in Probability and Statistics. 2nd ed. New York: Wiley, 2002.

4. Vach W, Blettner M. 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. Am J Epidemiol 1991;134(8):895-907.

5. White JE. A two stage design for the study of the relationship between a rare exposure and a rare disease. Am J Epidemiol 1982;115(1):119-28.

6. Janssen KJ, Donders AR, Harrell FE, Jr., Vergouwe Y, Chen Q, Grobbee DE, et al. Missing covariate data in medical research: to impute is better than to ignore. J Clin Epidemiol 2010;63(7):721-7.

7. Miettinen OS. Theoritical epidemiology. New York: Wiley, 1995.

8. Molenberghs G, Kenward M. Missing data in clinical studies. Wiley series in Probability ans Statistics. Chichester: 2007.

9. Vergouw D, Heymans MW, Peat GM, Kuijpers T, Croft PR, de Vet HC, et al. The search for stable prognostic models in multiple imputed data sets. BMC Med Res Methodol 2010;10:81.

10. Raghunathan TE. What do we do with missing data? Some options for analysis of incomplete data. Annu Rev Public Health 2004;25:99-117.

11. Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol 1995 ;142(12):1255-64.

12. Allison PD. Missing data. Iowa City: Sage Publication, 2002.

13. Commenges D, Gagnon M, Letenneur L, Dartigues JF, Barberger-Gateau P, Salamon R. Improving screening for dementia in the elderly using, Mini-Mental State Examination subscores, Benton's Visual Retention Test, and Isaacs' Set Test. Epidemiology 1992;3(2):185-8.

14. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychol Methods 2001;6(4):330-51.

15. Jacqmin-Gadda H. Analyse de données incomplètes: typologie et méthodes d'analyse lorsque les données manquantes sont ignorables. Modélisation des données incomplètes: Analyses de sensibilité. La Londe-les-Maures 2007, p.1-12. La Londe-les-Maures 2007 p. 1-12.

16. Zhou XH, Eckert GJ, Tierney WM. Multiple imputation in public health research. Stat Med 2001;20(9-10):1541-9.

17. Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods 2002;7(2):147-77.

18. Rubin DB, Schenker N. Multiple imputation in health-care databases: an overview and some applications. Stat Med 1991;10(4):585-98.

19. Sinharay S, Stern HS, Russell D. The use of multiple imputation for the analysis of missing data. Psychol Methods 2001;6(4):317-29.

20. Kenward MG, Carpenter J. Multiple imputation: current perspectives. Stat Methods Med Res 2007;16(3):199-218.

21. Arnold AM, Kronmal RA. Multiple imputation of baseline data in the cardiovascular health study. Am J Epidemiol 2003;157(1):74-84.

22. Barnard J, Meng XL. Applications of multiple imputation in medical studies: from AIDS to NHANES. Stat Methods Med Res 1999;8(1):17-36.

23. Cattle BA, Baxter PD, Greenwood DC, Gale CP, West RM. Multiple imputation for completion of a national clinical audit dataset. Stat Med 2011;In Press.

24. Barzi F, Woodward M. Imputations of missing values in practice: results from imputations of serum cholesterol in 28 cohort studies. Am J Epidemiol 2004;160(1):34-45.

25. Wood AM, White IR, Hillsdon M, Carpenter J. Comparison of imputation and modelling methods in the analysis of a physical activity trial with missing outcomes. Int J Epidemiol 2005;34(1):89-99.

26. Klebanoff MA, Cole SR. Use of multiple imputation in the epidemiologic literature. Am J Epidemiol 2008;168(4):355-7.

27. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res 2011;20(1):40-9.

28. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med 2011;30(4):377-99.

29. Schafer JL. Analysis of incomplete multivariate data. Monographs on Statistics and Applied Probability 72. Chapman & Hall. London: 1997.

30. Rubin DB. Multiple imputation after 18+ years. Am Stat Assoc 1996;91(434):473-89.

31. Graham JW. Missing data analysis: making it work in the real world. Annu Rev Psychol 2009;60:549-76.

32. Wayman JC. Multiple imputation for missing data: what is it and how can I use it ? Annual Meeting of the American Educational Research. Chicago 2003.

33. Schafer JL. Multiple imputation: a primer. Stat Methods Med Res 1999 Mar;8(1):3-15.

34. Taylor JM, Cooper KL, Wei JT, Sarma AV, Raghunathan TE, Heeringa SG. Use of multiple imputation to correct for nonresponse bias in a survey of urologic symptoms among African-American men. Am J Epidemiol 2002;156(8):774-82.

35. Faris PD, Ghali WA, Brant R, Norris CM, Galbraith PD, Knudtson ML. Multiple imputation versus data enhancement for dealing with missing data in observational health care outcome analyses. J Clin Epidemiol 2002;55(2):184-91.

36. van der Heijden GJ, Donders AR, Stijnen T, Moons KG. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. J Clin Epidemiol 2006;59(10):1102-9.

37. Nur U, Longford NT, Cade JE, Greenwood DC. The impact of handling missing data on alcohol consumption estimates in the UK women cohort study. Eur J Epidemiol 2009;24(10):589-95.

38. Horton NJ, Lipsitz SR. Multiple imputation in practice: comparison of software packages for regression models with missing variables. American Statistician 2001;55(3):244-54.

39. Horton NJ, Kleinman KP. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. Am Stat 2007;61(1):79-90.

40. Royston P. Multiple imputation of missing values. Stata J 2004;4(3):227-41.

41. Cottrell G, Cot M, Mary JY. [Multiple imputation of missing at random data: General points and presentation of a Monte-Carlo method]. Rev Epidemiol Sante Publique 2009;57(5):361-72.

42. Van Buuren S., Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med 1999;18(6):681-94.

43. Van Buuren S, Groothuis-Oudshoorn K, Rubin DB. Fully conditional specification in multivariate imputation. J Stat Comput Simul 2006;76:1049-64.

44. Yu LM, Burton A, Rivero-Arias O. Evaluation of software for multiple imputation of semi-continuous data. Stat Methods Med Res 2007;16(3):243-58.

45. Lee KJ, Carlin JB. Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. Am J Epidemiol 2010;171(5):624-32.

46. Gelman AE, Smith AFM. Sampling-based approaches to calculating marginal densities. J Am Stat Assoc 1990;85:398-409.

47. Royston P. Multiple imputation of missing values: Update. Stata J 2005;5(2):188-201.

48. Royston P. Multiple imputation of missing values: Update of ice. Stata J 2005;5(4):527-36.

49. Meng XL. Multiple imputation inferences with uncongenial sources of input. Stat Sci 1994;9(4):538-73.

50. Moons KG, Donders RA, Stijnen T, Harrell FE, Jr. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol 2006;59(10):1092-101.

51. Bartlett JW, Frost C, Carpenter JR. Multiple imputation models should incorporate the outcome in the model of interest. Brain 2011 Jun 6.

52. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009;338:b2393.

53. Schafer JL. Multiple imputation in multivariate problems when the imputation and analysis model differ. Stat Neer 2003;57(1):19-35.

54. Wood AM, White IR, Royston P. How should variable selection be performed with multiply imputed data? Stat Med 2008;27(17):3227-46.

55. Horton NJ, Lipsitz SR, Parzen M. A potential for bias when rounding in multiple imputation. Am Stat 2003;57(4):229-32.

56. Royston P. Multiple imputation of missing values: furter update of ice, with an emphasis on categorical variables. Stata J 2009;9(3):466-77.

57. Bodner TE. What improves with increased missin data imputations? Struct Equ Modeling 2008;15:651-75.

58. Graham JW, Olchowski AE, Gilreath TD. How many imputations are really needed? Some practical clarifications of multiple imputation theory. Prev Sci 2007;8(3):206-13.

59. Yang X, Belin TR, Boscardin WJ. Imputation and variable selection in linear regression models with missing covariates. Biometrics 2005;61(2):498-506.

60. Potthoff RF, Tudor GE, Pieper KS, Hasselblad V. Can one assess whether missing data are missing at random in medical studies? Stat Methods Med Res 2006;15(3):213-34.

61. Abayomi K, Gelman A, Levy M. Diagnostics for multiple imputation. Appl Statist 2011;57(3):273-91.

62. Raghunathan TE, Bondarenko I. Diagnostics for multiple imputation.2007.[ressource électronique]. Disponible sur : http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1031750 [consulté le 22/09/2011].

63. Nevalainen J, Kenward MG, Virtanen SM. Missing values in longitudinal dietary data: a multiple imputation approach based on a fully conditional specification. Stat Med 2009;28(29):3657-69.

64. Royston P, Carlin JB, White IR. Multiple imputation of missing values: New features for mim. Stata J 2009;9(2):252-64.

65. Stata multiple-imputation reference manual - release 11.College Station, Texas:2011.

66. Meng X, Rubin D. Performing likelihood ratio tests with multiply-imputed data sets. Biometrika 1992;79(1):103-11.

67. Steyerberg EW, van VM. Imputation is beneficial for handling missing data in predictive models. J Clin Epidemiol 2007;60(9):979.

68. Carpenter J, Rucker G, Schwarzer G. Assessing the Sensitivity of Meta-analysis to Selection Bias: A Multiple Imputation Approach. Biometrics 2011;67(3):1066-72.

69. Wang C, Hall CB. Correction of bias from non-random missing longitudinal data using auxiliary information. Stat Med 2010;29(6):671-9.

70. Chavance M, Manfredi R. [Modeling incomplete observations]. Rev Epidemiol Sante Publique 2000;48(4):389-400.

71. Le Vu S, Le Strat Y, Barin F, Pillonel J, Cazein F, Bousquet V, et al. Population-based HIV-1 incidence in France, 2003-08: a modelling analysis. Lancet Infect Dis 2010;10(10):682-7.

72. Semaille C, Cazein F, Lot F, Pillonel J, Le VS, Le SY, et al. Recently acquired HIV infection in men who have sex with men (MSM) in France, 2003-2008. Euro Surveill 2009;14(48).

73. Pillonel J, Le MN, Girault A, David D, Laperche S. [Epidemiological surveillance of blood donors and residual risk of blood-borne infections in France, 2001 to 2003]. Transfus Clin Biol 2005;12(3):239-46.

74. Schreiber GB, Busch MP, Kleinman SH, Korelitz JJ. The risk of transfusion-transmitted viral infections. The Retrovirus Epidemiology Donor Study. N Engl J Med 1996;334(26):1685-90.

75. Pillonel J, Laperche S, Saura C, Desenclos JC, Courouce AM. Trends in residual risk of transfusion-transmitted viral infections in France between 1992 and 2000. Transfusion (Paris) 2002;42(8):980-8.

76. Enquête sur la sexualité en France: Pratiques, genre et santé. Editions La découverte. Paris.2008.

77. Pillonel J, Heraud-Bousquet V, Pelletier B, Semaille C, Velter A, Saura C, et al. Deferral from donating blood of men who have sex with men: impact on the risk of HIV transmission by transfusion in France. Vox Sang 2011; In Press.

78. Friedman CR, Hoekstra RM, Samuel M, Marcus R, Bender J, Shiferaw B, et al. Risk factors for sporadic Campylobacter infection in the United States: A case-control study in FoodNet sites. Clin Infect Dis 2004;38 Suppl 3:S285-S296.

79. Neimann J, Engberg J, Molbak K, Wegener HC. A case-control study of risk factors for sporadic campylobacter infections in Denmark. Epidemiol Infect 2003;130(3):353-66.

80. Studahl A, Andersson Y. Risk factors for indigenous campylobacter infection: a Swedish case-control study. Epidemiol Infect 2000;125(2):269-75.

81. Gallay A. Contribution à l'épidémiologie des infections à Campylobacter en France. Thèse en Epidémiologie. Université Paris XI. 2006.

82. Stuart EA, Azur M, Frangakis C, Leaf P. Multiple imputation with large data sets: a case study of the Children's Mental Health Initiative. Am J Epidemiol 2009;169(9):1133-9.

83. UNAIDS, WHO. AIDS epidemic update. 2009.[Ressource électronique]. Disponible sur: http://www.unaids.org/globalreport/Global_report.htm [consulté le 28 mars 2010].

84. Warszawski J, Tubiana R, Le Chenadec J, Blanche S, Teglas JP, Dollfus C, et al. Mother-to-child HIV transmission despite antiretroviral therapy in the ANRS French Perinatal Cohort. AIDS 2008;22(2):289-99.

85. Yeni P. Prise en charge médicale des personnes infectées par le VIH. Médecine-Sciences Flammarion, 2010.

86. Hook EB, Regal RR. Capture-recapture methods in epidemiology: methods and limitations. Epidemiol Rev 1995;17(2):243-64.

87. Lot F, Semaille C, Cazein F, Barin F, Pinget R, Pillonel J, et al. Preliminary results from the new HIV surveillance system in France. Euro Surveill 2004;9(10):34-7.

88. Cazein F, Le Vu S, Pillonel J, Le Strat Y, Couturier S, Basselier B, et al. Dépistage de l'infection par le VIH en France, 2003-2009. Bulletin Epidemiologique Hebdomadaire 2010;45-46:451-4.

89. Cazein F. Surveillance du VIH: notification obligatoire et surveillance virologique, Lutte contre le VIH/sida et les infections sexuellement transmissibles en France - 10 ans de surveillance, 1996-2005. 2007. [Ressource électonique]. Disponible sur : http://www.invs.sante.fr/pmb/invs/(id)/PMB_4047 [consulté le 3 mai 2011].

90. Stata module to perform capture-recapture analysis for three sources with goodness of fit based confidence intervals [computer program]. 2007.

91. Regal RR, Hook EB. Goodness-of-fit based confidence intervals for estimates of the size of a closed population. Stat Med 1984 Jul;3(3):287-91.

92. Draper D. Assessment and propagation of model uncertainty. J R Stat Soc [B] 1995;57:45-70.

93. Hook EB, Regal RR. Validity of methods for model selection, weighting for model uncertainty, and small sample adjustment in capture-recapture estimation. Am J Epidemiol 1997;145(12):1138-44.

94. Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. BMC Med Res Methodol 2009;9:57.

95. National Institute of Statistics and Economic Studies. National population census.2011.[Ressource électronique]. Disponible sur : http://www.insee.fr/fr/themes/detail.asp?reg_id=0&ref_id=ir-sd2008&page=irweb/sd2008/dd/sd2008_population.htm [consulté le 10 mai 2011].

96. Zwane EN, van der Heijden PG. Analysing capture-recapture data when some variables of heterogeneous catchability are not collected or asked in all registrations. Stat Med 2007;26(5):1069-89.

97. Robb ML, Bohning D. Imputing unobserved values with the EM algorithm under left and right-truncation, and interval censoring for estimating the size of hidden populations. Biom J 2011;53(1):75-87.

98. van der Heijden PG, Zwane E, Hessen E. Structurally missing data problems in multiple list capture-recapture data. AStA Adv Stat Anal 2009;93:5-21.

99. Zwane E, van der Heijden PG. Capture-recapture studies with incomplete mixed categorical and continuous covariates. J Data Sci 2008;6:557-72.

100. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med 2011;30(4):377-99.

101. Dubois F, Desenclos JC, Mariotte N, Goudeau A. Hepatitis C in a French population-based survey, 1994: seroprevalence, frequency of viremia, genotype distribution, and risk factors. The Collaborative Study Group. Hepatology 1997;25(6):1490-6.

102. Meffre C, LeStrat Y, Delarocque-Astagneau E, Dubois F, Antona D, Lemasson JM, et al. Prevalence of hepatitis B and hepatitis C virus infections in France in 2004: social factors are important predictors after adjusting for known risk factors. J Med Virol 2010;82(4):546-55.

103. Delarocque-Astagneau E, Roudot-Thoraval F, Campese C, Desenclos JC, The Hepatitis CSSS. Past excessive alcohol consumption: a major determinant of severe liver disease among newly referred hepatitis C virus infected patients in hepatology reference centers, France, 2001. Ann Epidemiol 2005 ;15(8):551-7.

104. Delarocque-Astagneau E, Campese C. Surveillance de l'hépatite C à l'échelon national à partir des pôles de référence volontaires, 2001-2001. Bull Epidemiol Hebd 2003;16-17:90-3.

105. Jauffret-Roustide M, Le Strat Y, Couturier E, Thierry D, Rondy M, Quaglia M, et al. A national cross-sectional study among drug-users in France: epidemiology of HCV and highlight on practical and statistical aspects of the design. BMC Infect Dis 2009;9:113.

106. Massard J, Ratziu V, Thabut D, Moussalli J, Lebray P, Benhamou Y, et al. Natural history and predictors of disease severity in chronic hepatitis C. J Hepatol 2006;44(1 Suppl):S19-S24.

107. Marcellin P, Pequignot F, Delarocque-Astagneau E, Zarski JP, Ganne N, Hillon P, et al. Mortality related to chronic hepatitis B and chronic hepatitis C in France: evidence for the role of HIV coinfection and alcohol consumption. J Hepatol 2008 ;48(2):200-7.

108. Barreiro P, Martin-Carbonero L, Nunez M, Rivas P, Morente A, Simarro N, et al. Predictors of liver fibrosis in HIV-infected patients with chronic hepatitis C virus (HCV) infection: assessment using transient elastometry and the role of HCV genotype 3. Clin Infect Dis 2006;42(7):1032-9.

109. Rubbia-Brandt L, Fabris P, Paganin S, Leandro G, Male PJ, Giostra E, et al. Steatosis affects chronic hepatitis C progression in a genotype specific way. Gut 2004;53(3):406-12.

110. Spratt M, Carpenter J, Sterne JA, Carlin JB, Heron J, Henderson J, et al. Strategies for multiple imputation in longitudinal studies. Am J Epidemiol 2010;172(4):478-87.

111. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Stat Med 2010;29(28):2920-31.

112. Diggle MG, Kenward MG. Informative drop-out in longitudinal data analysis. Appl Statist 1994;43(1):49-93.

113. Dodge HH, Shen C, Ganguli M. Application of the Pattern-Mixture Latent Trajectory Model in an Epidemiological Study with Non-Ignorable Missingness. J Data Sci 2008;6(2):247-59.

114. Curran D, Molenberghs G, Thijs H, Verbeke G. Sensitivity analysis for pattern mixture models. J Biopharm Stat 2004 Feb;14(1):125-43.

115. Molenberghs G, Verbeke G. Models for discrete longitudinal data. New York: Springer, 2005.

116. Carpenter JR, Kenward MG, White IR. Sensitivity analysis after multiple imputation under missing at random: a weighting approach. Stat Methods Med Res 2007;16(3):259-75.

117. Ma G, Troxel AB, Heitjan DF. An index of local sensitivity to nonignorable drop-out in longitudinal modelling. Stat Med 2005;24(14):2129-50.

118. Rubin DB. The calculation of posterior distributions by data augmentation (in discussion of Tanner MA, Wong WH). J Am Stat Assoc 1987;82(398):543-6.

119. Pernanen K. Validity of survey data on alcohol use. New York: Wiley, 1974.

120. Van Oers JA, Bongers IM, van de Goor LA, Garretsen HF. Alcohol consumption, alcohol-related problems, problem drinking, and socioeconomic status. Alcohol Alcohol 1999;34(1):78-88.

121. Knibbe R. Measuring drinking context. Alcohol Clin Exp Res 1998;22(2 Suppl):15S-20S.

122. Lemmens PH, Tan ES, Knibbe RA. Bias due to non-response in a Dutch survey on alcohol consumption. Br J Addict 1988;83(9):1069-77.

123. Lahaut VM, Jansen HA, van de MD, Garretsen HF. Non-response bias in a sample survey on alcohol consumption. Alcohol Alcohol 2002;37(3):256-60.

124. Dhumeaux D, Marcellin P, Lerebours E. Treatment of hepatitis C. The 2002 French consensus. Gut 2003;52(12):1784-7.

125. He Y, Zaslavsky A, Landrum M, Harrington D, Catalano P. Multiple imputation in a large-scale complex survey: a practical guide. Stat Methods Med Res 2009;In Press.

126. Marston L, Carpenter JR, Walters KR, Morris RW, Nazareth I, Petersen I. Issues in multiple imputation of missing data for large general practice clinical databases. Pharmacoepidemiol Drug Saf 2010;19(6):618-26.

127. Green TA. Using surveillance data to monitor trends in the AIDS epidemic. Stat Med 1998;17(2):143-54.

128. Hall HI, Song R, Rhodes P, Prejean J, An Q, Lee LM, et al. Estimation of HIV incidence in the United States. JAMA 2008;300(5):520-9.

129. Prejean J, Song R, Hernandez A, Ziebell R, Green T, Walker F, et al. Estimated HIV incidence in the United States, 2006-2009. PLoS One 2011;6(8):e17502.

130. Harrison KM, Kajese T, Hall HI, Song R. Risk factor redistribution of the national HIV/AIDS surveillance data: an alternative approach. Public Health Rep 2008;123(5):618-27.

131. Cazein F, Lot F, Pillonel J, Pinget R, Bousquet V, Le Strat Y, et al. Surveillance de l'infection à VIH-sida en France, 2009. Bull Epidemiol Hebd 2010;45-46:467-72.

132. Lot F, Semaille C, Cazein F, Barin F, Pinget R, Pillonel J, et al. Preliminary results from the new HIV surveillance system in France. Eurosurveillance 2004;9(10-12):34-7.

133. Semaille C, Barin F, Cazein F, Pillonel J, Lot F, Brand D, et al. Monitoring the dynamics of the HIV epidemic using assays for recent infection and serotyping among new HIV diagnoses: experience after 2 years in France. J Infect Dis 2007;196(3):377-83.

134. Barin F, Meyer L, Lancar R, Deveau C, Gharib M, Laporte A, et al. Development and validation of an immunoassay for identification of recent human immunodeficiency virus type 1 infections and its use on dried serum spots. J Clin Microbiol 2005;43(9):4441-7.

135. Barin F, Plantier JC, Brand D, Brunet S, Moreau A, Liandier B, et al. Human immunodeficiency virus serotyping on dried serum spots as a screening tool for the surveillance of the AIDS epidemic. J Med Virol 2006;78 Suppl 1:13-8.

136. ANRS CO9 COPANA. Cohorte de patients non traités par antirétroviraux à l'inclusion.[Ressource électronique]. Disponible sur : http://www.anrs.fr/index.php/content/download/783/5243/file/ANRSCO9 COPANA.pdf [consulté le 05 février 2011].

137. White IR, Daniel R, Royston P. Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. Comput Stat Data Anal 2010;54:2267-75.

138. Van Buuren S, Groothuis-Oudshoorn K. MICE: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software 2010;In press.

139. Royston P. Multiple imputation of missing values: further update of ice, with an emphasis on interval censoring. Stata J 2007;7(4):445-64.

140. Carlin JB, Galati JC, Royston P. A new framework for managing and analyzing multiply imputed data in Stata. Stata J 2008;8(1):49-67.

141. Fraser G, Yan R. Guided multiple imputation of missing data: using a subsample to strengthen the missing-at-random assumption. Epidemiology 2007;18(2):246-52.

142. Costagliola D, Descamps D, Assoumou L, Morand-Joubert L, Marcelin AG, Brodard V, et al. Prevalence of HIV-1 drug resistance in treated patients: a French nationwide study. J Acquir Immune Defic Syndr 2007;46(1):12-8.

143. Valenciano M, Kissling E, Cohen JM, Oroszi B, Barret AS, Rizzo C, et al. Estimates of Pandemic Influenza Vaccine Effectiveness in Europe, 2009-2010: Results of Influenza Monitoring Vaccine Effectiveness in Europe (I-MOVE) Multicentre Case-Control Study. PLoS Med 2011;8(1):e1000388.

Annexe 1

Deferral from donating blood of men who have sex with men: Impact on the risk of HIV transmission by transfusion in France.

VoxSanguinis
The International Journal of Transfusion Medicine

International Society of Blood Transfusion

Vox Sanguinis (2011)

ORIGINAL PAPER

DOI: 10.1111/j.1423-0410.2011.01509.x

Deferral from donating blood of men who have sex with men: impact on the risk of HIV transmission by transfusion in France

J. Pillonel[1], V. Heraud-Bousquet[1], B. Pelletier[2], C. Semaille[1], A. Velter[1], C. Saura[1], J.-C. Desenclos[1], B. Danic[3] & the blood donor epidemiological surveillance study group*

[1] *Institut de Veille Sanitaire, Saint-Maurice, France*
[2] *Etablissement Français du Sang, Paris, France*
[3] *Etablissement Français du Sang de Bretagne, Rennes, France*

Vox Sanguinis

Background In France, men who have sex with men (MSM) are permanently excluded from blood donation. This policy is felt to be discriminatory by MSM activists. Furthermore, the policy is not fully respected because some MSM do not report their sexual behaviour before donating.

Methods We estimated the fraction of the current risk of HIV attributed to MSM. We then constructed a model based on data obtained from behavioural and epidemiological surveys to assess the impact of a new strategy in which MSM would only be deferred if they report more than one sexual partner in the last 12 months.

Results Thirty-one HIV seroconversions occurred among repeat donors between 2006 and 2008, giving a risk of one in 2 440 000 donations. Fifteen of these seroconversions (48%) were MSM. If all MSM had abstained from donating blood, the risk would have been 1 in 4 700 000 donations, half the current risk. The new strategy would result in an overall HIV risk of between 1 in 3 000 000 (close to the current risk) to 1 in 650 000 donations (3·7 times higher than the current risk).

Conclusions Changing the current MSM deferral policy may increase the risk of transfusion–transmission of HIV. However, this does not take into account a possible better compliance with MSM with a less stringent policy that would be perceived as more equitable. Conversely, relaxing the policy could encourage some MSM to seek an HIV test in blood centres. Thus, further qualitative study is needed to assess possible changes in compliance linked to a new policy.

Key words: blood donor selection, HIV residual risk, men who have sex with men.

Received: 5 November 2010, revised 29 April 2011, accepted 1 May 2011

Introduction

In France, men who have sex with men (MSM) are permanently excluded from blood donation, mainly because they are at higher risk of human immunodeficiency virus (HIV) [1, 2] but also at increased risk of other transfusion-transmissible infections as hepatitis B virus and syphilis [3, 4]. Exclusion of MSM began in the early 1980s when male-to-male sex was documented as a major mode of transmission of HIV and virus detection techniques were not established [5]. The Health Authorities insisted that the deferral policies had to be maintained, even after HIV screening of blood donations became mandatory in France in August 1985, essentially because of the 'window period'.

Since 1985, there has been a remarkable improvement in the viral safety of the blood supply due to improvements in donor selection and continuous progress in screening assays, including nucleic acid amplification testing (NAT).

*See appendix.

Correspondence: Josiane Pillonel, Institut de Veille Sanitaire, Département des Maladies Infectieuses, 12 rue du Val d'Osne, 94415 Saint-Maurice Cedex, France
E-mail: j.pillonel@invs.sante.fr

Despite these measures, there is still a residual risk of transmitting HIV by transfusion of blood components. In France, this risk was estimated to be 1 in 2 950 000 donations in 2005–2007, equivalent to fewer than one potentially infected donation per year on the basis of 2·7 million collected donations [6]. This residual risk is mainly linked to the 'window period', which occurs shortly after the donor is infected and before the markers of infection can be detected. Consequently, donor selection remains an important element in ensuring the viral safety of the blood supply.

The permanent deferral policy for MSM has been the subject of debate, mainly because this stringent criterion is felt by MSM activists to be discriminatory and outdated. They request a change to deferral based, as for heterosexual donors, on sexual behaviour during months preceding donation. They would prefer to be asked more specific sexual behaviour questions such as having had multiple sexual partners or unprotected sex with a new partner, rather than be asked whether they had sex with another man or whether they are homosexual. In addition, the permanent deferral policy is not fully effective because some MSM do not report their sexual behaviour before donating. This seems to be increasing over time and raises the question of changing the permanent deferral policy because of its lack of efficiency. In this study, we chose to assess a new strategy involving a deferral policy close to that applied to heterosexuals in France. With this new strategy, MSM would donate blood and be only deferred if they report more than one male sexual partner in the last year. The only difference with heterosexuals is the duration of 12 months instead of 4 months after the end of the risky behaviour. Furthermore, this proposal seems to be more acceptable by the MSM community than those assessed in previous studies which are based on different periods of sexual abstinence before donation (1, 5 or 10 years) [7–10].

In the first part of this study, we estimated the risk of HIV transmission by transfusion associated with the lack of compliance with MSM with the current policy. We then assessed the impact on this risk of the proposed strategy.

Methods

Study population

Since 1992, all blood centres in France report quarterly the total numbers of donors and donations according to donor status (first time and repeat) and epidemiological characteristics (sex, age, probable mode of transmission, HIV-1 subtype, geographic origin and donor status) of donors confirmed HIV-positive to the national surveillance system for blood donors [11]. More than three-quarters of donors found to be HIV positive return for a post-donation medical interview during which information on the probable mode of transmission of HIV is investigated. As the data set concerning the risk factor for acquiring HIV, including MSM, was not complete (missing for 30% of HIV-positive donors for the 1992–2008 period), we applied multiple imputation to estimate missing values from observed values using the method of imputation by chained equations (ado ice, STATA®11.0; Stata Corporation, College Station, TX, USA) [12] with the generation of 100 data sets. The proportion of MSM among donors found to be HIV positive was then calculated according to Rubin's rules [13].

HIV residual risk in 2006–2008 in France

The residual risk of transfusion-transmitted HIV infection per million donations was calculated for periods of 3 years as the product of the HIV incidence among blood donors who had made at least two donations during the study period and the length of the HIV window period (in years) [14].

The HIV incidence is the number of repeat donors who underwent HIV seroconversion (S) during the 3-year study period divided by the number of donor-years (DY) calculated by summing time intervals between the first and the last donation by each donor during the study period. The HIV window period was estimated to be 12 days with the use of minipool NAT [15].

Assessment of the fraction of the current HIV residual risk that can be attributed to MSM

HIV-positive MSM are regularly found donating blood despite the deferral policy. To estimate the current impact of MSM donations on HIV residual risk, the HIV incidence among blood donors was divided into two parts: HIV incidence estimated for MSM and HIV incidence estimated for the other donors (women + men not having sex with men) according to the following equation:

$$I = \frac{\mathrm{DY_{MSM}}}{(\mathrm{DY_{MSM}} + \mathrm{DY_{others}})}\ I_{\mathrm{MSM}} + \frac{\mathrm{DY_{others}}}{(\mathrm{DY_{MSM}} + \mathrm{DY_{others}})}\ I_{\mathrm{others}},$$

where I is the overall HIV incidence among blood donors, I_{MSM} the HIV incidence among MSM donors, I_{others} the HIV incidence among the other donors, $\mathrm{DY_{MSM}}$ the number of Donor-Years for MSM donors, and $\mathrm{DY_{others}}$ the number of Donor-Years for the other donors.

HIV incidence was calculated as follows: $I_{\mathrm{MSM}} = \frac{S_{\mathrm{MSM}}}{\mathrm{DY_{MSM}}} 10^5$, where S_{MSM} is the number of HIV seroconversions observed among MSM donors, and $I_{\mathrm{others}} = \frac{S_{\mathrm{others}}}{\mathrm{DY_{others}}} 10^5$, where S_{others} is the number of HIV seroconversions observed among the other donors during the study period.

Unlike the numbers of seroconversions (S_{MSM} and S_{others}), the numbers of Donor-Years could not be obtained

separately for MSM and other donors (DY_{MSM} and DY_{others}), through the national surveillance system for blood donors. We, therefore, used external data to split the known total number of Donor-Years (DY) into the two groups. A survey (CSF2006 study) on sexual behaviour of 12 364 people randomly recruited from the general population aged 18–69 years was conducted in France in 2006 [16]. In this survey, the percentage of men aged 18–69 years having sex with men within the previous 12 months was estimated to be 1·5% and the percentage of men having sex with men at least once during their life to be 4·1%. After having separated the number of Donor-Years between men and women (50·2% of repeat donors between 2006 and 2008 were men), we applied these proportions of MSM to the number of Donor-Years among men to obtain two estimates for MSM HIV incidence and then derived those for the other donors:

Hypothesis 1: 1·5% of the male donor population is MSM

$$I_{\mathrm{MSM}.1} = \frac{S_{\mathrm{MSM}}}{\mathrm{DY} \times 0.502 \times 0.015} 10^5$$

and

$$I_{\mathrm{others}.1} = \frac{S_{\mathrm{others}}}{\mathrm{DY} \times (1 - (0.502 \times 0.015))} 10^5,$$

where DY is the total number of Donor-Years and 0·502 the proportion of men among repeat blood donors between 2006 and 2008.

Hypothesis 2: 4·1% of the male donor population is MSM

$$I_{\mathrm{MSM}.2} = \frac{S_{\mathrm{MSM}}}{\mathrm{DY} \times 0.502 \times 0.041} 10^5 \quad \text{and}$$

$$I_{\mathrm{others}.2} = \frac{S_{\mathrm{others}}}{\mathrm{DY} \times (1 - (0.502 \times 0.041))} 10^5.$$

Residual risks could then be derived for MSM donors and the other donors.

Thus, we used proportions of MSM observed in the general population without taking into account the effects of the deferral policy on the blood donor population. Consequently, for both these two hypotheses, and particularly for hypothesis 2, incidence and thus HIV residual risk for MSM donors should be considered to be minimum estimates.

Impact of a new strategy in which MSM would only be deferred if they have had more than one sexual partner in the previous 12 months

We constructed a model based on data obtained from behavioural and epidemiological surveys of MSM to assess the consequences of the new strategy on the HIV residual risk.

The starting point of the model was the estimate of the number of sexually active MSM (N_{MSM}) in France. This estimate was obtained by applying the percentage of men aged 18–69 having sex with men within the last 12 months in the CSF 2006 study (1·5%) [16], to the male population aged 18–65 obtained from the national census. As MSM with more than one sexual partner in the previous 12 months would be deferred in the new strategy, we multiplied the percentage of such MSM as estimated in the CFS 2006 study (47·3%) [16] (plus personal communication, CSF 2006 study, INSERM-INED) by the number of sexually active MSM (N_{MSM}) to calculate the maximum number of donors who would be excluded for that reason. The number of MSM eligible to donate blood ($E_{BD\ MSM}$) was then calculated as the difference between the two values. We assume that the proportion of eligible MSM giving blood is the same as that of men in the general population aged 18–65 giving blood (p_{BD}). As incidence and residual risk are calculated among repeat blood donors, we divided the observed number of repeat male blood donors by the number of men of the general population aged 18–65 years to obtain the proportion p_{BD} that was stable, around 2·8%, for the years 2006, 2007 and 2008. Then, the number of MSM expected to donate blood (BD_{MSM}) can be estimated by multiplying $E_{BD\ MSM}$ by $p_{BD.}$

From the number of MSM expected to donate blood (BD_{MSM}), we then estimated the number of newly HIV-infected MSM expected to donate blood (BD_{MSMhiv}) as follows:

$$BD_{\mathrm{MSMhiv}} = BD_{\mathrm{MSM}} \times I_{\mathrm{MSM}},$$

where I_{MSM} is the HIV incidence among MSM expected to donate blood.

We considered two scenarios for I_{MSM}: best case and worst case. For the best-case scenario, we used the HIV incidence estimated for MSM blood donors during the period 2006–2008 (hypothesis 1 of the incidence estimate, $I_{MSM.1}$). For the worst-case scenario, we used the HIV incidence estimated in 2008 for MSM in France: 1·0% [1].

For the worst-case scenario, using data from HIV case reporting in France [17], we assumed that 25% of these newly HIV-infected MSM expected to donate blood would not donate because they would have been diagnosed to be HIV positive before their next blood donation and therefore would not subsequently donate blood. In the best-case scenario, as the incidence used was directly derived from the blood donor population, we did not apply this proportion because the corresponding incident cases did not enter the incidence calculation.

Finally, for each scenario, the annual number of HIV incident cases among MSM blood donors was multiplied by three to obtain the number for a 3-year period and added to

the number of incident cases observed among the other donors (men not having sex with men + women) to obtain the numerator of the incidence for the new strategy. For the number of Donor-Years for MSM, we used the estimate obtained under hypothesis 1 for HIV incidence calculation used to assess current HIV residual risk attributed to MSM ($DY_{MSM} = DY \times 0{\cdot}502 \times 0{\cdot}015$).

The 95% confidence intervals (95% CI) of the incidence and the residual risk were obtained by the Fleiss Quadratic method, which is pertinent when proportions are near to zero [18].

Results

HIV residual risk in 2006–2008 in France

During the 3-year period 2006–2008, 31 HIV seroconversions were observed among blood donors who had made at least two donations, such that the incidence was 1·3 per 100 000 Donor-Years (Table 1). On the basis of this incidence, the current HIV residual risk was estimated to be 0·41 per million donations or 1 in 2 440 000 donations (95% CI: 0–1 in 700 000).

Assessment of the fraction of the current HIV residual risk that can be attributed to MSM

Among the 31 HIV incident cases observed between 2006 and 2008, 23 were men and eight were women. Eleven of these men (48%) reported having had sex with men during the post-donation medical interview; for four (17%), the probable mode of transmission was heterosexual intercourse, and for eight (35%), the probable mode of transmission was unknown. After performing multiple imputation, we estimated the number of MSM to be 15 (65%) and the number of men contaminated through heterosexual intercourse to be eight (35%). By adding these eight incident cases among men to the eight incident cases among women, the number of other donors (not MSM) was estimated to be 16.

With hypothesis 1, where 1·5% of male donors would be MSM, the HIV residual risk linked to MSM was estimated to be 26·5 per million donations or 1 in 38 000 donations (Table 2), which is about 125 times higher than that linked to other donors (1 in 4 700 000 donations).

With hypothesis 2, where 4·1% of male donors would be MSM, the HIV residual risk linked to MSM was estimated to be 9·7 per million donations or 1 in 100 000 donations or about 45 times higher than that linked to other donors (1 in 4 600 000 donations). If all MSM had abstained from donating blood between 2006 and 2008, the HIV residual risk would have been around 1 in 4 600 000–1 in 4 700 000 donations, which is half the current risk (1 in 2 440 000 donations).

Incidence among MSM and other blood donors by 3-year period since 1994 is shown in Fig. 1: between the 1997–1999 period and the 2006–2008 period, the incidence increased significantly ($P = 0{\cdot}02$) among MSM, although the global HIV incidence was stable.

Table 1 Residual risk of transfusion-transmitted HIV infection associated with the window period – France, 2006–2008

No. of HIV incident cases	No. of donor- years	Incidence per 100 000 donor years (CI 95%)	Length of window period in days (range)	Residual risk per million donations (CI 95%)
31	2 467 560	1·26 (0·87–1·81)	12 (0–28)	0·41 (0·0–1·39)

Table 2 HIV incidence and residual risk of transfusion-transmitted HIV infection among MSM and other donors – France, 2006–2008

	MSM donors	Other donors
Number of HIV incident cases (seroconversions)	$S_{MSM} = 15$	$S_{others} = 16$
Number of donor–years		
Hypothesis 1: 1·5% of male donors are MSM	18 590	2 448 970
Hypothesis 2: 4·1% of male donors are MSM	50 813	2 416 747
HIV Incidence per 100 000 donor-years		
Hypothesis 1: 1·5% of male donors are MSM	$I_{MSM.1} = 80{\cdot}7$	$I_{others.1} = 0{\cdot}65$
Hypothesis 2: 4·1% of male donors are MSM	$I_{MSM.2} = 29{\cdot}5$	$I_{others.2} = 0{\cdot}66$
Residual risk per 1 million donations		
Hypothesis 1: 1·5% of male donors are MSM	$RR_{MSM.1} = 26{\cdot}5$	$RR_{others.1} = 0{\cdot}21$
Hypothesis 2: 4·1% of male donors are MSM	$RR_{MSM.2} = 9{\cdot}7$	$RR_{others.2} = 0{\cdot}22$

MSM, men who have sex with men.

Impact of a new strategy in which MSM would only be deferred if they have had more than one sexual partner in the previous 12 months

The model used to assess the impact of the new strategy predicted that an average of 4355 sexually active MSM would be expected to donate blood each year (Table 3). Depending on the scenario, between 3 and 44 of them would be newly infected with HIV. After eliminating those who would not donate because they would have been diagnosed HIV positive before their subsequent blood donation (11 in the worst-case scenario), the number of HIV incident cases among MSM was estimated to be between 3 and 33

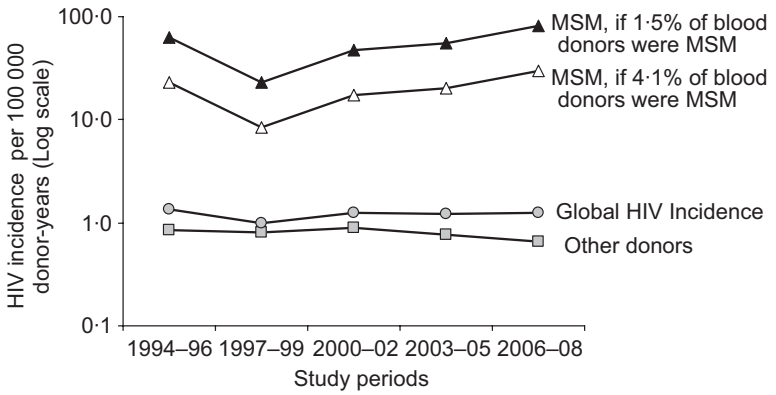


Fig. 1 Trend in HIV incidence among men who have sex with men and other blood donors – France, 1994–2008.

Table 3 Estimate of HIV seroconversions among MSM blood donors under a new strategy in which MSM would only be deferred if they have had more than one sexual partner in the last 12 months – France, 2006–2008

Number of men aged 18–65 years (Census, 01/01/2008)		19 676 810
Percentage of men having sex with men within the previous 12 months (CSF 2006)		1·5%
Number of sexually active MSM aged 18–65 years N_{MSM}		295 152
Percentage of MSM having more than one partner in the last 12 months (CSF 2006)		47·3%
Annual number of MSM deferred from blood donations		139 607
Annual number of MSM eligible to donate blood: $E_{BD\ MSM}$		155 545
Percentage of repeat male blood donors in the population aged 18–65 years: p_{BD}		2·8%
Annual number of MSM expected to donate blood: BD_{MSM}		4355
Scenarios	**Best case**	**Worst case**
HIV incidence rates	0·08%	1·00%
Annual number of newly HIV-infected MSM expected to donate blood: BD_{MSMhiv}	3	44
Percentage and annual number of donors diagnosed HIV positive before their next donation	–	25% (11)
Number of newly HIV-infected MSM expected to donate blood over 3 years	9	99

MSM, men who have sex with men.

for 1 year and between 9 and 99 for the 3-year period 2006–2008.

These estimates were added to the 16 incident cases observed for the other donors (men not having sex with men + women) during the same period to obtain the numerator of the incident rate for the new strategy (Table 4). This strategy would change the overall HIV residual risk from 0·33 per million donations or 1 in 3 000 000 to 1·53 per million or 1 in 650 000 donations. In the best-case scenario, this risk estimate is close to the current risk (1 in 2 440 000). In the worst-case scenario, the risk is 3·7 times higher corresponding, for France as a whole, to a mean of four donations potentially infected with HIV each year instead of one currently.

Discussion

This study is the first to assess the excess of risk of transmitting HIV by transfusion associated with the lack of compliance with MSM with the current policy for blood donor selection. Despite the permanent deferral of MSM from donating blood, we estimated that nearly half of the current risk of HIV transmission by transfusion in France can be attributed to MSM who do not comply with the policy. Although HIV incidence has been estimated to be much lower among MSM blood donors (0·08%) than among the MSM of the general population (1%) [1], the proportion of donors infected by male-to-male sex among donors newly infected with HIV was as high as the estimated proportion of MSM among people newly infected with HIV in the general population (48%): in 2008, an estimated 6940 persons became infected with HIV in France of whom 3320 were MSM [1]. The similarity of the proportion of MSM for blood donors and the general population would have been expected if the criteria for the donor selection were the same for all potential donors. However, the policy for heterosexuals is only temporary deferral based on risky behaviour (4 months after the end of the high-risk situation). Therefore, the percentage of blood donors who are MSM infected with HIV should be much lower than that of the general population. This result reveals the failure of a deferral policy based on a stringent criterion for MSM. There are at least two factors that may contribute to this: some MSM currently consider 'lifetime deferral' discriminatory while others may use blood transfusion centres as HIV screening centres. Both these groups may give blood and not reveal their sexual behaviour. Data from the Retrovirus Epidemiology Donor Study (REDS) showed that in a sample of

Table 4 Estimate of HIV residual risk (RR) if the current MSM deferral policy was replaced with a strategy in which MSM would only be deferred if they have had more than one sexual partner in the last 12 months – France, 2006–2008

Estimates	No of HIV incident cases	No. of Donor-Years (D-Y)	Incidence per 100 000 D-Y (CI 95%)	Residual risk per million donations (CI 95%)	% of current RR estimate
Current	31	2 467 560	1·26 (0·87–1·81)	0·41 (0·0–1·39)	100
Other donors only (current less 15 MSM)	16	2 448 970	0·65 (0·39–1·09)	0·21[a] (0·0–0·83)	51
Best-case scenario (adding 9 MSM)	25 (16 + 9)	2 467 560	1·01 (0·67–1·52)	0·33 (0·0–1·17)	81
Worst-case scenario (adding 99 MSM)	115 (16 + 99)	2 467 560	4·66 (3·75–5·62)	1·53 (0·0–4·31)	371

[a]This is the estimated HIV residual risk if the current deferral policy was effective, under the hypothesis 1 (see Table 2: $RR_{others\text{-}1}$).
MSM, men who have sex with men.

25 168 male blood donors interviewed anonymously, 2·4% reported male-to-male sex and, among them, 9·9% of those who reported sex with men within the last 12 months also reported recently seeking an HIV test [19]. It is likely that this proportion is much higher among HIV-positive MSM donors.

A worrying finding of our analysis is the significant increase in the HIV incidence among MSM between the 1997–1999 and 2006–2008 periods. It is unlikely that this increase reflects an increase in HIV incidence in MSM in the general population: a recent study reports that the incidence of HIV in MSM of the general population was stable between 2003 and 2008 [1].

This increase could be due to a 'magnet effect' linked to the implementation of NAT testing in July 2001 in France. Indeed, some MSM may seek a free RNA test by giving blood after risky behaviour. Of the 11 donors, who were found to be HIV RNA positive and antibody negative between July 2001 and December 2009 in France, nine were men, of whom four (45%) reported during the post-donation medical interview having had sex with men; for two (22%), the probable mode of contamination was heterosexual intercourse whereas for three (33%), it was unknown. However, these small numbers are difficult to interpret.

The increase may also be the consequence of the feeling of discrimination growing over time. If it were the case, more and more MSM would come to give blood and intentionally not report their sexual behaviour before donating. This point is very difficult to investigate as we do not know the sexual behaviour of HIV-negative blood donors. To calculate the denominator for HIV incidence among MSM blood donors in our study, we used estimated proportions of MSM in the general population (from 1·5% to 4·1% depending on the definition, last 12-months or lifetime). This probably underestimates the incidence as MSM are in theory excluded from blood donation. Furthermore, we used the same proportions of MSM over the entire period, such that the denominator remained stable. However, if this denominator increased between 1997 and 2008, the apparent increase in HIV incidence among blood donors over this period could be artificial. Despite these limitations, our data indicate that MSM make a large contribution to the current HIV residual risk of transfusion and therefore raise the question of changing the permanent deferral policy because of its lack of efficiency.

In the second part of our study, we used a model to estimate the impact on the HIV residual risk of a change in strategy in which MSM would be only deferred if they have had more than one sexual partner in the last 12 months. Depending on the scenario used, this new strategy would result in an overall HIV residual risk of between 1 in 3 000 000 donations, close to the current risk (1 in 2 440 000) and 1 in 650 000 donations. The worst-case scenario corresponds to a mean of three donations potentially infected with HIV each year in France in addition to the current estimate of one donation infected with HIV each year. Based on these numbers, the incremental risk of this change in policy concerning MSM would probably be low; nevertheless, considering the worst-case scenario, it could be judged to be unacceptable. However, these estimates do not take into account the possible better compliance with MSM with a less stringent policy. As the perpetuation of the lifetime deferral for MSM is discrimination in the opinion of many within the homosexual community, some avoid the ban by hiding their sexual behaviour during the pre-donation interview. Currently, with the lifetime deferral policy, more than half of HIV incident cases among blood donors are subsequently found to be MSM. It is therefore reasonable to assume that a less stringent policy would be perceived to be more equitable and as a result enhance responsibility. MSM who donate blood under a new policy based on risky behaviour may self-defer in a more appropriate way. Conversely, relaxing the policy could be perceived as indicating that the policy is less important, and as a result of that perception, self-deferral might be reduced. Thus, the new policy could encourage some MSM at risk to seek an HIV test in the

blood centres. Unfortunately, these potential changes in the attitude of MSM are difficult to predict, showing the complexity of estimating the true impact of a less stringent deferral policy for MSM.

Several other factors regarding the assumptions or parameters used in the model may have affected the results. The number of sexually active MSM and the annual number of MSM that would be deferred from blood donations were based on a sexual behaviour survey of more than 12 000 people, randomly recruited from the French general population in 2006 (CSF 2006) [16]. Socially stigmatised behaviours, including sex between men, can be underreported in questionnaire surveys [20]. Nevertheless, the reported proportion of men having sex with men within the previous 12 months (1·5%) was the same as that estimated 14 years earlier with the same study design (ACSF 1992) [21]. Furthermore, others countries have reported similar proportions of sexually active MSM in their population [22, 23]. We assumed in our model that all MSM having more than one partner in the last 12 months would be excluded from donating blood, and this is probably too optimistic even in a more confident climate. This assumption might have led to underestimating the impact of the proposed strategy on the HIV residual risk. On the other hand, in the worst-case scenario, we used the HIV incidence estimated in the overall MSM population in France (1%) to calculate the HIV incidence among MSM not having more than one partner in the last 12 months: this may have led to an overestimation of the risk associated with the proposed strategy. Another limitation is that our results are derived from repeat donors only. However, these donors gave more than 80% of all donations, and furthermore, the HIV incidence is comparable between repeat and first-time donors in France [24]. Thus, our results can be extended to the entire blood donor population. Finally, our model did not take into account the risk due to the release of false negative donations, which is nevertheless much lower than the risk linked to the window period [25].

Several studies have examined different deferral criteria for MSM. Germain *et al.* [7] calculated the incremental risk of accepting 12-month abstinent MSM in the United States and Canada. They estimated one additional HIV-contaminated unit for every 136 000 new MSM donations, representing an overall increase of 8% in HIV risk. Soldan and Sinka [8] estimated the increased risk of accepting 12-month abstinent MSM to be approximately 60% in the United Kingdom. Using REDS data, Sanchez *et al.* [19] estimated that the prevalence of reactive infectious screening tests was significantly higher among donors who reported male-to-male sex within the past 5 years than among non-MSM in the United States. Nevertheless, as donors who reported male-to-male sex within 5 years constituted less than 0·6% of all male donors, the impact of these donors on the overall risk would be low. More recently, Anderson *et al.* [9] predicted annual increases in risk of HIV-infected blood of 0·5% if accepting donations from 5-year abstinent MSM or 3·0% when accepting donations from 12-month abstinent MSM. In Australia, where the deferral policy had been modified more than 10 years ago, a recent study showed that neither the global rate of HIV-positive donations nor the proportion of HIV positive with male-to-male sex as risk factor had been changed by the implementation of a 12-month deferral for male-to-male sex [26].

Lacking data on HIV incidence in MSM stratified by time of abstinence from MSM behaviour, we were unable to produce equivalent estimates for France. In our model, we chose to assess a new strategy involving a deferral policy close to that applied to heterosexuals in France. This new strategy is less stringent than those assessed in previous studies [7–9]. Thus, in our worst-case scenario, our estimate of the impact on the risk of HIV transmission by transfusion was higher. Leiss *et al.* [10] suggested that an increase in risk is a clear violation of ethical principles and therefore not acceptable. Therefore, a strategy of accepting 12-month abstinent MSM would probably be the most acceptable change because the current incremental risk appears to be extremely low as estimated in some recent studies [9], compared with currently tolerated transfusion risks [27]. If such a deferral policy was implemented in France, communication efforts to publicise donor selection policy to encourage appropriate self-deferral would be warranted because compliance is a crucial parameter.

Conclusion

Despite the lifetime deferral policy of MSM, nearly half of HIV transmission by transfusion in France was attributed to MSM blood donors. Furthermore, the proportion of MSM among blood donors was as high as among people newly infected with HIV in the general population. These findings raise the question of modifying the lifetime deferral policy. A strategy in which MSM would be only deferred if they have had more than one sexual partner in the last 12 months may increase the risk of transfusion–transmission of HIV. However, this does not take into account a possible change in compliance with MSM linked to a modification of the deferral policy. As some MSM currently consider 'lifetime deferral' discriminatory, they give blood while hiding their sexual behaviour. A less stringent policy may be perceived to be more equitable and should enhance responsibility. Conversely, relaxing the policy could be perceived as indicating that the policy is less important and encourage some MSM at risk to seek an HIV test in the blood centres. Further qualitative study is thus needed in addition to our quantitative analysis to assess possible

changes in compliance with the MSM population linked to a less stringent blood donor selection policy.

Acknowledgements

We thank all colleagues from the Etablissement Français du Sang and the Centre de Transfusion des Armées who participated in the national blood donor surveillance. We thank Marlène Leclerc for her technical assistance.

Conflicts of interest

None.

Funding

No external funding.

References

1 Le Vu S, Le Strat Y, Barin F, *et al.*: Population-based HIV-1 incidence in France, 2003–08: a modelling analysis. *Lancet Infect Dis* 2010; **10**:682–687
2 Semaille C, Cazein F, Lot F, *et al.*: Recently acquired HIV infection in men who have sex with men (MSM) in France, 2003–2008. *Euro Surveill* 2009; **14**: 1–4
3 Bouyssou-Michel A, Gallay A, Janier M, *et al.*: Surveillance de la syphilis en France, 2000–2006 : recrudescence des diagnostics en 2006. *Bull Epidemiol Heb* 2008; **6**:39–42
4 Meffre C, Le Strat Y, Delarocque-Astagneau E, *et al.*: Prevalence of hepatitis B and hepatitis C virus infections in France in 2004: social factors are important predictors after adjusting for known risk factors. *J Med Virol* 2010; **82**:546–555
5 Leads from the MMWR. Prevention of acquired immune deficiency syndrome (AIDS): Report of inter-agency recommendations. *JAMA* 1983; **249**:1544–1545
6 Pillonel J, Brouard C, Laperche S, *et al.*: Quantitative estimate of the risk of blood donation contamination by infectious agents. *Transfus Clin Biol* 2009; **16**:138–145
7 Germain M, Remis RS, Delage G: The risks and benefits of accepting men who have had sex with men as blood donors. *Transfusion* 2003; **43**:25–33
8 Soldan K, Sinka K: Evaluation of the deselection of men who have had sex with men from blood donation in England. *Vox Sang* 2003; **84**:265–273
9 Anderson SA, Yang H, Gallagher LM, *et al.*: Quantitative estimate of the risks and benefits of possible alternative blood donor deferral strategies for men who have had sex with men. *Transfusion* 2009; **49**:1102–1114
10 Leiss W, Tyshenko M, Krewski D: Men having sex with men donor deferral risk assessment: an analysis using risk management principles. *Transfus Med Rev* 2008; **22**:35–57
11 Pillonel J, Le Marrec N, Girault A, *et al.*: Epidemiological surveillance of blood donors and residual risk of blood-borne infections in France, 2001 to 2003. *Transfus Clin Biol* 2005; **12**:239–246
12 Royston P: Multiple imputation of missing values: further update of ice, with an emphasis on interval censoring. *Stata J* 2007; **7**:445–464
13 Little RJ, Rubin D: Statistical Analysis with Missing Data, 2nd edn. Hoboken, NJ, John Wiley & Sons Inc., 2002
14 Schreiber GB, Busch MP, Kleinman SH, *et al.*: The risk of transfusion-transmitted viral infections. The Retrovirus Epidemiology Donor Study. *N Engl J Med* 1996; **334**:1685–1690
15 Pillonel J, Laperche S: Trends in risk of transfusion-transmitted viral infections (HIV, HCV, HBV) in France between 1992 and 2003 and impact of nucleic acid testing (NAT). *Euro Surveill* 2005; **10**:5–8
16 Bajos N, Bozon M: Enquête sur la sexualité en France: Pratiques, genre et santé. Paris, Editions La Découverte, 2008
17 Cazein F, Pillonel J, Bousquet V, *et al.*: Caractéristiques des personnes diagnostiquées avec une infection à VIH ou un sida, France, 2008, BEHweb 2009; 1–5
18 Fleiss J: Statistical Methods for Rates and Proportions, 2nd edn. New York, John Wiley, 1981. 2010
19 Sanchez AM, Schreiber GB, Nass CC, *et al.*: The impact of male-to-male sexual experience on risk profiles of blood donors. *Transfusion* 2005; **45**:404–413
20 Sandfort T: Sampling male homosexuality; in Bancrofs J (ed.): Researching Sexual Behavior: Methodological Issues. Bloomington, Indiana University Press, 1997: 261–275
21 Groupe ACSF (coordonné par Bajos N, Giami A, Laurent R, Leridon H, Spira A): Comportements sexuels et sida en France. Données de l'enquête 'Analyse des comportements sexuels en France'. Paris, Editions INSERM, 1998
22 Mercer CH, Fenton KA, Copas AJ, *et al.*: Increasing prevalence of male homosexual partnerships and practices in Britain 1990–2000: evidence from national probability surveys. *AIDS* 2004; **18**:1453–1458
23 Xu F, Sternberg MR, Markowitz LE: Men who have sex with men in the United States: demographic and behavioral characteristics and prevalence of HIV and HSV-2 infection: results from National Health and Nutrition Examination Survey 2001–2006. *Sex Transm Dis* 2010; **37**:399–405
24 Pillonel J, Barin F, Laperche S, *et al.*: Human immunodeficiency virus type 1 incidence among blood donors in France, 1992 through 2006: use of an immunoassay to identify recent infections. *Transfusion* 2008; **48**:1567–1575
25 Pillonel J, Laperche S, Saura C, *et al.*: Trends in residual risk of transfusion-transmitted viral infections in France between 1992 and 2000. *Transfusion* 2002; **42**:980–988
26 Seed CR, Kiely P, Law M, *et al.*: No evidence of a significantly increased risk of transfusion-transmitted human immunodeficiency virus infection in Australia subsequent to implementing a 12-month deferral for men who have had sex with men. *Transfusion* 2010; **50**:2722–2730
27 Vamvakas EC: Relative risk of reducing the lifetime blood donation deferral for men who have had sex with men versus currently tolerated transfusion risks. *Transfus Med Rev* 2011; **25**:47–60

Appendix

The workgroup 'The blood donor epidemiological surveillance study group' consists of (in alphabetic order):

Agence Française de Sécurité Sanitaire des Produits de Santé: C. Caldani, JF. Legras, E. Pouchol, MP. Vo-Mai.

Agence Régionale de Santé Languedoc Roussillon (Coordonateur Régional d'Hémovigilance): G. Daurat.

Centre de Transfusion Sanguine des Armées: A. Kerleguer.

Direction Générale de la Santé: B. Worms.

Etablissement Français du Sang: A. Assal, B. Danic, B. David, MC. Dupuy-Montbrun, MH. Elghouzzi, P. Gallian, MF. Lecomte des Floris, P. Morel, B. Pelletier, D. Rebibo, C. Waller.

Institut National de la Transfusion Sanguine: G. Andreu, S. Laperche.

Institut de Veille Sanitaire: JC. Desenclos, J. Pillonel, C. Saura.

Annexe 2

Risk factors for acquiring sporadic Campylobacter infection in France: Results from a national case-control study.

Risk Factors for Acquiring Sporadic *Campylobacter* Infection in France: Results from a National Case-Control Study

Anne Gallay,[1] Vanina Bousquet,[1] Virginie Siret,[1,2] Valérie Prouzet-Mauléon,[3] Henriette de Valk,[1] Véronique Vaillant,[1] Fernando Simon,[1] Yann Le Strat,[1] Francis Mégraud,[3] and Jean-Claude Desenclos[1]

[1]Département Maladies Infectieuses and [2]Field Epidemiology Training (Programme de Formation d'Épidémiologie de Terrain), Institut de Veille Sanitaire, Saint-Maurice, and [3]Centre National de Référence des *Campylobacter* et *Helicobacter,* Bordeaux, France

Background. To better document the risk factors for sporadic *Campylobacter* infection in France, we conducted a national case-control study from September 2002 to June 2004.

Methods. Cases with confirmed *Campylobacter* infection were sampled through the national surveillance laboratory network. Cases and controls who were matched for age, as well as attending physicians, were interviewed about foods consumed, food preparation practices, travel history, contact with cases and animals during the 8 days before the onset of infection, and any antibiotic use occurring during the 30 days before onset. Matched odds ratios [ORs] were calculated using conditional logistic regression and multiple imputation methods.

Results. A total of 285 pairs of cases and matched controls were enrolled. "Ate undercooked beef" (OR, 2.86; 95% confidence interval [CI], 1.65–4.95), "ate at restaurant" (OR, 2.20; 95% CI, 1.23–3.93), and "poor utensils hygiene in the kitchen" (OR, 2.12; 95% CI, 1.33–3.37) were the main independent risk factors for infection. Cases infected with a ciprofloxacin-resistant *Campylobacter jejuni* strain were more likely than controls to have used antibiotics in the month before onset.

Conclusion. Good hygiene practices in the kitchen remain a strong recommendation to avoid cross-contamination. However, studies are needed to explore the mechanism of contamination throughout the food chain. The use of antibiotics in humans may favor the development of a resistant infection.

In most industrialized countries, *Campylobacter* organisms are, along with *Salmonella* organisms, the most common cause of foodborne bacterial gastroenteritis [1, 2]. Previous studies have shown that *Campylobacter* infection is frequently acquired through consuming undercooked poultry and other meats, drinking untreated water, having contact with pets or farm animals, and traveling abroad [3–10]. Even if there is evidence that cross-contamination between raw meat and other foods during meal preparation is responsible for a consequent number of cases, this factor has not been identified in epidemiologic studies [5–7, 9, 11–13]. In addition, what were identified as risk factors in some studies were found to be protective factors in others, even in studies performed in the same country [14]. The specific immunity conferred by *Campylobacter* infection suggests that susceptibility should vary by age and exposure [15]. Because of their inability to multiply in food products, *Campylobacter* organisms have a low outbreak potential; however, they have been responsible for several large outbreaks, most of them waterborne [16–19].

Campylobacter infection is usually self-limited, but extraintestinal infection or septicemia may occur and may require treatment with appropriate antibiotics. Since the beginning of the 1990s, the resistance of *Campylobacter* organisms to antibiotics has increased. Epidemiologic studies, modeling, and experimental studies have documented a strong association between the increase in resistance to quinolones/fluoroquinolones in human isolates and the use of these antibiotics in animals [20–22].

Received 26 October 2007; accepted 12 December 2007; electronically published 1 April 2008.

Potential conflicts of interest: none reported.

Presented in part: 11th European Program for Intervention in Epidemiology Training Scientific Seminar, Mahon, Menorca, Spain, 12–14 October 2006 (abstract 20060090).

Reprints or correspondence: Dr. Anne Gallay, Institut de Veille Sanitaire, Département des Maladies Infectieuses, 12, rue du Val d'Osne, 94 415 Saint-Maurice, France (a.gallay@invs.sante.fr).

The Journal of Infectious Diseases 2008; 197:1477–84

0022-1899/2008/19710-0018$15.00
DOI: 10.1086/587644

Surveillance of *Campylobacter* infection was expanded in 2002 in France to monitor the incidence of and trends associated with bacterial resistance of *Campylobacter* infection [23, 24]. So far, very little information has been available on risk factors in France. We therefore performed a prospective case-control study of sporadic *Campylobacter* infection among French residents.

MATERIAL AND METHODS

Study design. A "case" was defined as a resident of metropolitan France who had clinical symptoms of campylobacteriosis (i.e., gastroenteritis, systemic infection, or extradigestive infections) and a culture-confirmed *Campylobacter* isolate identified, from 15 September 2002 to 30 June 2004, either in stool or in a biological liquid that is normally sterile. When samples from >1 household member yielded *Campylobacter* species, or when the case was either part of a household in which there was a cluster of instances of gastroenteritis or part of a recognized outbreak, only the first identified case was enrolled.

The present study was a collaboration between the National Reference Center (NRC) for *Campylobacter* and *Helicobacter* (Centre National de Référence des *Campylobacter* and *Helicobacter*) and the French Institute for Public Health Surveillance (Institut de Veille Sanitaire) in France, conducted as part of their routine activity. Potential cases were recruited through the national *Campylobacter* surveillance network, which is based on a voluntary network of private ($n = 342$) and public ($n = 92$) laboratories that send their isolates to the NRC [24]. Laboratories notified the cases to the French Institute for Public Health Surveillance by fax, to reduce the delay until an interview could be conducted. Each working day, a maximum of 4 cases with the shortest delay between the date of isolation and the date of notification were sampled. A letter was faxed to the case's physician to request from the case or the case's parents informed consent to participate in the study.

Each case was matched, by age, sex, and geographic area, with one control selected from the patient registry of the case's physician. The age groups of the matched pairs were defined as <6 months, 6 months to 3 years, >3 to <15 years [±3 years], and ⩾15 years [±10 years]. Cases and controls were not matched by sex before 15 years of age, under the assumption that food habits were similar by sex. Potential controls were excluded from the study if they had experienced diarrhea within 1 month before or 1 week after the onset of infection in the matched case or if they were identified >15 days after the date that a *Campylobacter* isolate was identified in the matched case. If a potential control refused to participate or could not be reached after 5 attempts, a search for another control was conducted.

The study received ethics approval from the Commission Nationale de l'Informatique et des Libertés.

Interviews and questionnaires. Cases and matched controls were contacted and interviewed by telephone by the same interviewer, who was not blinded to the case/control status. For children ⩽12 years of age, a parent or a person who was familiar with the usual environment and food habits of the children was interviewed. A structured questionnaire was used to collect information on demographic characteristics; medical history, including use of any antibiotic within the month before onset of infection in the matched case; clinical symptoms; occupational activity; and specific exposures. Information on food consumption, the level of cooking of meat, food handling practices, travel history, animal and water (i.e., drinking or swimming) exposures, and contact with a person with diarrhea was recorded for cases and controls during the 8 days before onset in the matched case.

Susceptibility testing of* Campylobacter *isolates. *Campylobacter* isolates were tested at the NRC for susceptibility to ciprofloxacin, erythromycin, amoxicillin, gentamicin, and tetracycline, by use of the agar diffusion method, on Mueller-Hinton agar enriched with 5% sheep blood, with use of antibiotic disks according to the Antibiogram Committee of the French Society for Microbiology [25].

Analysis. The level of cooking was classified as "undercooked" when the meat or poultry consumed was raw, rare, or pink in color and as "cooked" when the meat or poultry consumed was well done or overcooked. Food handling hygiene practices were classified as "poor utensils hygiene in the kitchen" when hands or utensils used in the kitchen were not cleaned or were dried only with a dish towel in between the handling of meat or poultry and other foods and as "good utensils hygiene in the kitchen" when hands or utensils were cleaned with water only or a detergent or when utensils changed.

Variables that were associated with the outcome and for which $P < .2$ in matched univariate analysis were considered for multiple conditional logistic regression analysis. Matched odds ratios (ORs) for which $P \leqslant .05$ were considered to be significant in the multivariate analysis. We repeated the analysis after imputing for missing responses [26, 27]. All analyses were performed using Stata software (version 9; Stata Corporation).

The risk associated with the use of antibiotics within the month before onset of illness was analyzed by comparing cases with an antibiotic-resistant strain with (1) their matched controls and (2) cases with an antibiotic-susceptible strain.

RESULTS

Study population. During the 21-month study, 2743 patients with culture-confirmed *Campylobacter* infection were notified. Of the 954 cases (34.8%) sampled, 285 (29.9%) were included in the study. Cases that were excluded ($n = 669$) included 15 cases with an undetermined date of onset, 285 with a notification delay of >10 days from the date of onset, 174 who could not be

Table 1. Clinical symptoms of cases, by age group, in a case-control study of *Campylobacter* infection in France, September 2002–June 2004.

Symptom	Age ⩽15 years (N = 169)	Age >15 years, (N = 116)
Diarrhea	161/168 (95.8)	113/116 (97.4)
Abdominal pain	134/149 (89.9)	104/116 (89.7)
Fever[a]	116/163 (71.2)	76/105 (72.4)
Bloody diarrhea	91/166 (54.8)[b]	39/111 (35.1)
Vomiting	66/168 (39.3)[c]	28/116 (24.1)
Weight loss	81/139 (58.3)[c]	78/104 (75.0)

NOTE. Data are the proportion (%) of cases; in each fraction, the numerator denotes the no. of cases who reported having the symptom, and the denominator denotes the no. of cases who answered the question.

[a] Temperature, ⩾38°C.

[b] $P = .001$.

[c] $P = .007$.

reached after 5 attempts, 117 who refused to participate, 42 for whom a matched control was not found, and 36 who were excluded for other reasons. The 285 enrolled cases were younger (mean age, 19.5 years) than the cases who were excluded from the study (mean age, 28.5 years) and were comparable with respect to sex (percentage of cases that were male, 60.1% vs. 56.3%, respectively) and region of residence (data not shown). The mean age of the cases and controls was 19.5 and 20.0 years, respectively. The median delay between the time of onset of disease and the time when the interview was conducted was 15 days (range, 5–44 days), and the median delay between the time that interviews were conducted for cases and controls was 4 days (range, 0–64 days).

Clinical description. Of the 235 isolates in which the *Campylobacter* species was fully identified, 192 (81.7%) were *C. jejuni,* 36 (15.3%) were *C. coli,* 3 (1.3%) were *C. fetus,* and 4 (1.7%) were *C. lari.* The main symptoms were diarrhea (in 96.5% of patients), abdominal pain (in 90.0%), and fever (in 71.6%). Bloody diarrhea and vomiting were more common in children <15 years of age (table 1). Diarrhea lasted longer (1) in cases >15 years of age (median duration, 5 days) than in younger cases (median duration, 4 days) ($P = .002$) and (2) in cases with *C. fetus* infection (median duration, 13 days) than in cases with *C. jejuni* or *C. coli* infection (median duration, 5 days) ($P < .001$). Forty-one (14.4%) of the 285 cases were hospitalized (median duration, 2 days; range, 1–10 days).

Risk factors. Sixteen cases (5.6%) reported traveling abroad in the 8 days before onset, compared with 7 controls (2.4%) (OR, 2.5; 95% confidence interval [CI], 0.9–6.4). Countries visited included North Africa ($n = 8$), Europe ($n = 8$), Burkina Faso and Ethiopia ($n = 2$), and Tahiti and Thailand ($n = 5$). Because travelers were likely to have exposures that were remarkably different from those of nontravelers, and because exposures occurring during travels have not been investigated, travelers were excluded from the univariate and multivariate matched analysis, regardless of whether they were cases or controls.

In matched univariate analysis, consumption of undercooked beef was associated with an increased risk of campylobacteriosis (OR, 2.0; 95% CI, 1.2–3.4) (table 2). Cases were not more likely than controls to report consumption of any poultry or other meat, even that which was undercooked (data not shown). Eating in a restaurant increased the risk significantly (OR, 1.9; 95% CI, 1.1–1.3). Cases were less likely than controls to have eaten raw vegetables, fruits or berries, and fish or seafood.

Tasting or handling raw meat in the kitchen was not a risk factor, whereas poor hygiene of hands and utensils in the kitchen was; 30.0% and 37.4% of cases had poor hands hygiene or poor utensils hygiene, respectively, between handling raw meat and handling other food, compared with 21.0% and 26.1% of controls, respectively (table 2).

Contact with any pet or farm animals was not a significant risk factor (table 2). Having contact with a person who had diarrhea in the 8 days before onset increased the risk of *Campylobacter* infection (OR, 2.5; 95% CI, 1.4–4.4). Of the 51 cases (19.0%) who reported contact with a person(s) with diarrhea, 28 reported contact with individuals in the same household, 12 reported contact with children outside the household who had diarrhea, 7 reported contact with friends or colleagues at work, and 4 reported contact with children, friends, or colleagues.

Multivariate analysis. Fourteen variables have been included in the multivariate model, both for the complete-case (CC) and the multiple imputation (MI) analyses. No significant first-order interaction was detected. In both models, having contact with a person with diarrhea (OR_{CC}, 3.32 [95% CI, 1.72–6.4]; OR_{MI}, 2.27 [95% CI, 1.24–4.14]), eating undercooked beef (OR_{CC}, 2.26 [95% CI, 1. 2–4.24]; OR_{MI}, 2.86 [95% CI, 1.65–4.95]), and eating at restaurant (OR_{CC}, 2.46 [95% CI, 1.26–4.74]; OR_{MI}, 2.20 [95% CI, 1.23–3.93]) were independent risk factors for *Campylobacter* infection while eating beef bought from a butcher shop, farm, or market, and eating raw vegetables was associated with a decreased risk of infection (table 3). Poor utensils hygiene when preparing foods in the kitchen (OR_{MI}, 2.12; 95% CI, 1.33–3.37), which was found to be a significant factor in univariate analysis, remained in multiple imputation analysis.

Antibiotic treatment and resistance. Of the 285 strains, 233 (81.8%) were tested for antibiotic susceptibility. Results of testing susceptibility to ampicillin, tetracycline, and gentamicin were missing for 15 strains, and results of testing susceptibility to ciprofloxacin were missing for 10 strains. Eighty-nine (48.0%) of 218 strains were resistant to ampicillin; 67 (30.7%) of 218, to tetracycline; 60 (26.9%) of 223, to ciprofloxacin; 4 (1.7%) of 233, to erythromycin; and 2 (0.9%) of 218 to gentamicin. The duration of illness was longer for cases infected with a strain resistant to any antibiotic (median, 8.0 days; range, 1–38 days) than for cases infected with a susceptible strain (median, 6.0 days; range, 2–30 days) ($P = .02$).

Table 2. Matched univariate analysis of selected dichotomous risk factors for sporadic indigenous campylobacteriosis in France, September 2002–June 2004.

Potential risk factor	Persons reporting/ total respondents,[a] no.		Matched OR (95% CI)	*P*
	Cases	Controls		
Poultry-related exposure[b]				
Ate chicken				
Any	105/191	106/177	0.6 (0.4–1.1)	.1
Outside	14/106	8/110	1.0 (0.3–3.4)	1.0
Bought raw	71/169	90/173	0.6 (0.3–1.0)	.05
Bought at retail vs. packed or frozen	32/187	49/202	0.6 (0.3–1.0)	.06
That was undercooked	22/134	20/164	0.8 (0.4–1.9)	.7
Ate any poultry	143/201	135/185	0.7 (0.4–1.3)	.2
Beef-related exposure[b]				
Ate beef				
Any	123/228	135/233	0.8 (0.5–1.2)	.2
Outside	13/260	12/260	1.0 (0.4–2.4)	1.0
Bought from a butcher shop, farm, or market	58/257	75/251	0.6 (0.4–1.0)	.02
Bought raw	102/211	117/217	0.7 (0.5–1.1)	.1
Bought at retail	67/214	90/222	0.6 (0.4–0.9)	.02
That was undercooked	74/234	46/243	2.0 (1.2–3.4)	.007
That was undercooked or minced	97/250	66/257	2.1 (1.2–3.5)	.007
Other food exposure[b]				
Ate at a restaurant	58/260	34/247	1.9 (1.1–3.3)	.01
Ate any meat at a barbecue	31/263	25/263	1.3 (0.7–2.5)	.3
Ate fish or seafood	159/237	178/230	0.6 (0.3–0.9)	.01
Ate raw vegetables	161/252	184/244	0.4 (0.3–0.9)	.01
Ate fruits or berries	77/234	104/243	0.5 (0.4–0.8)	.004
Ate any milk product	223/255	228/255	0.7 (0.4–1.4)	.3
Ate cheese	185/263	199/263	0.7 (0.5–1.1)	.1
Drank tap water	102/260	100/261	1.1 (0.7–1.6)	.7
Behavior in the kitchen[c]				
Prepared meat or poultry	50/258	48/253	0.9 (0.5–1.7)	.7
Tasted meat or poultry	13/268	10/277	1.8 (0.7–4.9)	.2
Practiced poor hygiene				
Of hands	75/246	52/252	1.6 (1.0–2.3)	.03
Of utensils	94/248	67/254	1.7 (1.1–2.6)	.007
Of kitchen cutting boards	54/244	41/249	1.3 (0.8–2.0)	.3
Contact with a person with diarrhea[b]	51/249	26/254	2.5 (1.4–4.4)	.001
Animal exposure[b]				
Contact with any pet or farm animal	201/255	178/247	1.5 (1.0–2.4)	.06
Occupational	60/263	49/263	1.3 (0.8–2.0)	.2

NOTE. A total of 269 cases and controls were included in the analysis. CI, confidence interval; OR, odds ratio.

[a] For specific exposures assessed during the 8 days before onset in the matched case, all cases and controls could not always remember precisely and were therefore considered to be nonrespondents.

[b] During the 8 days before onset in the matched case.

[c] Not during the 8 days before onset in the matched case but as a general habit.

Table 3. Multivariate analysis (conditional logistical regression) of risk factors for sporadic indigenous campylobacteriosis, by use of complete-case (CC) and multiple imputation (MI) analyses in France, September 2002–June 2004.

	CC analysis (n = 372)[a]		MI analysis (n = 526)[a]	
Variable	OR (95% CI)	P	OR (95% CI)	P
Ate beef				
Bought from a butcher shop, farm, or market	. . .	. . .	0.56 (0.35–0.91)	.02
That was undercooked	2.26 (1.20–4.24)	.01	2.86 (1.65–4.95)	<.001
Ate at a restaurant	2.46 (1.26–4.79)	.01	2.20 (1.23–3.93)	.01
Ate raw vegetables	. . .	. . .	0.42 (0.23–0.74)	.003
Poor utensils hygiene in the kitchen	. . .	. . .	2.12 (1.33–3.37)	.002
Contact with a person with diarrhea	3.32 (1.72–6.40)	<.001	2.27 (1.24–4.14)	.01

NOTE. CI, confidence interval; OR, matched odds ratio.

[a] The n value denotes the no. of cases or controls in the final model.

Fifteen (12.9%) of 116 cases ⩾15 years of age, compared with 2 (1.7%) of 116 controls, reported receiving treatment with antibiotics during the month before the onset of illness (OR, 7.5; 95% CI, 1.7–32.8). No difference in the receipt of antibiotics was noted between cases and controls <15 years of age. Cases with ciprofloxacin-resistant *Campylobacter* infection (due to any *Campylobacter* species or due to *C. jejuni*) were more likely to have received any antibiotic before onset than were the controls (21.4% for those with *C. jejuni* infection vs. 0% for controls; OR, undefined), regardless of whether they had traveled abroad (table 4). Antibiotic use was a risk factor for indigenous ciprofloxacin-resistant *C. jejuni* infection, compared with ciprofloxacin-susceptible *C. jejuni* infection, and it was not a risk factor for ciprofloxacin-susceptible infection when compared with the matched control (table 4).

Nine (40.9%) of the 22 indigenous cases who reported receiving any antibiotic treatment in the month before onset were infected with a ciprofloxacin-resistant *C. jejuni* strain, compared with 27 (17.9%) of the 151 indigenous cases who did not report receiving antibiotic treatment during that time (OR, 3.2; 95% CI, 1.2–8.3). None of the 6 cases with travel-associated ciprofloxacin-resistant infection due to a *C. jejuni* strain had received any antibiotics.

DISCUSSION

In the present study, we found that eating undercooked beef, using poor utensil hygiene practices in the kitchen, eating at a restaurant, and having contact with a person who had diarrhea were independent risk factors for the acquisition of *Campy-*

Table 4. Association between use of antibiotics in the month before disease onset and ciprofloxacin-resistant *Campylobacter* infection in a case-control study of *Campylobacter* infection in France, September 2002–June 2004.

Infection	Cases with resistant strains, n/N (%)	Controls matched with cases with resistant strains, n/N (%)	OR[a] (95% CI)	Cases with susceptible strains, n/N (%)	OR[b] (95% CI)	Controls matched with cases with susceptible strains, n/N (%)	OR[c] (95%) CI
All *Campylobacter*							
All species	10/60 (16.7)	1/60 (1.7)	10.0 (1.2–78.1)	19/164 (11.6)	1.5 (0.7–3.5)	24/164 (14.6)	0.75 (0.4–1.5)
C. jejuni	9/42 (21.4)	0/42	UD[d]	15/142 (10.6)	2.3 (0.9–5.8)	19/142 (13.4)	0.75 (0.4–1.6)
Indigenous *Campylobacter*							
All species	9/53 (17.0)	1/53 (1.9)	9 (1.5– 199)	17/159 (10.7)	1.9 (0.8–4.6)	25/159 (15.7)	0.65 (0.3–1.3)
C. jejuni	8/35 (22.9)	0/35	UD[e]	13/137 (9.5)	3.2 (1.2–8.4)	3/137 (2.2)	0.62 (0.3–1.4)

NOTE. CI, confidence interval; n, no. of cases (with either resistant or susceptible strains) or matched control subjects who received antibiotics during the month before onset of disease in the case; N, no. of cases (with either resistant or susceptible strains) or matched control subjects who answered the question. OR, odds ratio; UD, undefined.

[a] OR for the comparison of cases with ciprofloxacin-resistant strains with their matched controls.

[b] OR for the comparison of cases with ciprofloxacin-resistant strains with cases with ciprofloxacin-susceptible strains (unmatched analysis).

[c] OR for the comparison of cases with ciprofloxacin-susceptible strains with their matched controls.

[d] P = .003.

[e] P = .004.

lobacter infection. Receipt of antibiotic treatment within the month before onset of illness increased the risk of developing a ciprofloxacin-resistant *C. jejuni* infection.

The most important food-specific risk factor was consumption of undercooked (raw, rare, or "pink") beef. Few studies have shown that beef was a food vehicle for *Campylobacter* infection. In the United States, the fact that riding in a shopping cart next to meat or poultry has been identified as risk factor in children suggests that contamination may occur through direct contact. Investigators discussed the risk of cross-contamination through indirect exposure—that is, via the hands of caretakers as a result of the external contamination of retail meat packages [28]. In other studies, "ate red meat at barbecue/open fire," "ate other raw or undercooked meat or fish," and "ate nonpoultry meat prepared at restaurant" were identified as independent risk factors, without specification as to whether the meat could be pork, veal, or beef [5, 7, 9]. Our finding may reflect a different food habit in France or in southern European countries. Most case-control studies have been conducted in northern European countries, and results may not reflect the food habits of individuals in southern Europe. The lack of association between consumption of beef and development of *Campylobacter* infection in studies performed in the northern European countries could be due to the less frequent consumption of undercooked beef in those countries. Beef ranks third among the animal products consumed in France, following pork and poultry in popularity [29], and it is often eaten undercooked (rare or pink) or even raw. In the present study, 59.3% of cases and 36.5% of controls ate undercooked beef. *Campylobacter* organisms are known to contaminate the gut and carcass of cattle [30, 31]. However, the interpretation of this finding is complex, because *Campylobacter* organisms are present only on the outside of beef. Eating undercooked beef may not reflect a direct mode of transmission but, rather, may be a marker of cross-contamination when foods are prepared in the kitchen. Eating at a restaurant, which has also been identified as a risk factor in other studies, may be related to consumption of undercooked, contaminated meat or to poor hygiene practices used in the kitchen.

In this study, several food exposures identified as risk factors in other studies were not found to increase the risk of *Campylobacter* infection. Although many studies identified consumption of chicken or poultry—and, more specifically, consumption of undercooked chicken [3–10]—as a risk factor [5, 7, 9, 11, 32], this was not the case in the present study. However, the lack of an increased risk associated with poultry or chicken consumption has been previously reported in several studies [13, 33–35], and one study even found that eating chicken prepared at home was protective [14]. Several reasons may explain these divergent results. In France, 80% of all broilers are colonized by *Campylobacter* species [36, 37]. Because chicken is widely consumed in France, it is very likely that many people have immunity to *Campylobacter* species, and those who are the most often exposed may have the highest immunity. Therefore, identifying chicken as a risk factor for *Campylobacter* infection is often difficult. Furthermore, eating undercooked chicken may play a major role through cross-contamination from the juice of raw chicken rather than through the direct risk associated with consumption of undercooked chicken. A social desirability bias could have also hidden the association, with cases being less likely than controls to report consumption of undercooked poultry. However, this is unlikely, because cases were more likely than controls to report poor hygiene in the kitchen, an exposure for which social desirability bias may play a greater role.

Eating raw vegetables was associated with a decreased risk of *Campylobacter* infection. Several hypotheses have been proposed to explain this association, including artifacts, bias, or confounding associated with lifestyle and food habits not explored in the study, as well as a causal effect [12, 32, 38].

The use of poor utensils hygiene in the kitchen when handling raw meat and other foods was an independent risk factor for infection. Although previous studies explored hygiene practices in the kitchen during food preparation, to our knowledge, no study has been able to show an independent increased risk for *Campylobacter* infection [5–7, 9, 11–13]. Hygiene practices perceived to be acceptable are not sufficient to prevent cross-contamination between a contaminated source (i.e., chicken) and a vehicle (i.e., other food), which is the direct route of infection [12, 39]. Vehicles with low-level contamination are sufficient to cause *Campylobacter* infection [40]. Drying utensils or hands with a dish towel in between handling meat or poultry and other foods may be wrongly perceived as an acceptable measure, although it does not prevent cross-contamination.

Person-to-person transmission of *Campylobacter* organisms is thought to be rare. Other pathogens, like *Salmonella* organisms, that are mainly transmitted through food consumption can be transmitted from person to person, particularly among children [41]. For more susceptible hosts, this mode of transmission may occur more frequently in association with poor hygiene conditions, and for *Campylobacter* organisms, it may be facilitated by its low infective dose. In the present study, contact with a person who had diarrhea in the 8 days before onset was an independent risk factor. Most contacts with a person with diarrhea occurred in the family household or with children who had diarrhea outside the home. Our design did not allow differentiation of whether more frequent diarrhea among contacts before the onset of infection was related to a common source of infection or truly revealed person-to-person transmission. However, *Campylobacter* organisms may account for a large number of household foodborne outbreaks that are unreported [42].

Like other studies, our study shows an association between travel abroad and acquisition of fluoroquinolone-resistant *Campylobacter* infection. Africa and Asia are the more common destinations for French travelers, whereas, in northern European countries (Sweden, Finland, and Norway), France and

other European Mediterranean countries are at risk for foreign travellers [43, 44]. However, traveling abroad accounted for a low proportion of cases of *Campylobacter* infection.

An increased risk for acquiring ciprofloxacin-resistant *C. jejuni* infection and *Salmonella* infection after receipt of any antibiotic treatment has been reported in few studies [45, 46]. This increased risk was observed when cases infected with a resistant strain were compared with healthy controls, as well when they were compared with cases infected with a susceptible strain, thereby suggesting that antibiotic use may facilitate ciprofloxacin-resistant *C. jejuni* infection. Such an effect is likely related to disruption of the normal flora in a person taking antibiotics for another reason, and it may be also related to the selection of a resistant strain related to the antibiotic used. If so, in addition to their use in veterinary medicine, antibiotics used in human medicine may contribute to the occurrence of ciprofloxacin-resistant *C. jejuni* infections in humans. However, we were not able to document precisely the antibiotics used, and we cannot assess whether this increased risk is more related to the use of a particular class of antibiotics [45]. Also, the limited number of cases for which the analysis was done according to *Campylobacter* species and antibiotic resistance profile reduced our ability to explore this association in detail.

Our results need to be interpreted with caution, however. The fact that this exploratory study assessed many exposures could have led to several random findings. Two factors reduced the statistical power of the study: (1) the planned sample size was not attained, and (2) data on specific exposures were missing. However, in comparison with complete-case analysis, the multiple imputation method increased the statistical power, allowing 2 additional risks factors to be identified and more-precise estimates of the odds ratios to be obtained.

Although we made important efforts to reduce the delay between the time of onset of the disease and the time when cases and controls were interviewed, memory effects cannot be excluded, particularly when many factors, including food consumption and cooking behaviors, were explored retrospectively. These memory effects could lead to recall bias. In the present study, such bias may more likely be nondifferential than differential, which tends to bias toward the nil rather than create a biased association. Missing data may be generated by the same memory effects, and a nonresponse bias could occur when the proportion of missing data differs between cases and controls for each exposure variable [47, 48]. In the present study, the proportion of data that were missing was similar for cases and controls. Because the results obtained using CC and MI analyses were similar, and because the assumptions made during the use of the MI analysis were met, the nonresponse bias was probably negligible. When many factors contribute to the disease, as for *Campylobacter* infection, we may also miss interaction or confounding variables for which the analysis cannot account. We also need to be cautious regarding the external validity of our results, because cases enrolled in the case group were not completely representative of those identified by the surveillance system (i.e., they were younger and represented a smaller proportion of cases that occur in summer). Cases were also recruited from laboratories and could, therefore, differ from cases with *Campylobacter* infection not diagnosed microbiologically.

Despite these limitations, the present study identifies important risk factors that are modifiable through changes in behavior, particularly when they relates to cooking habits and hygiene practices that favor cross-contamination during the preparation of food in the kitchen. However, prevention of transmission of *Campylobacter* organisms should be considered at each stage of the food chain transformation, from animal feed to food that is ready to eat. Efforts are needed to decrease contamination; in particular, poultry and meat (including beef) have to be considered as a potential source of contamination. The present study also strongly suggests that antibiotics used in humans have an influence on the risk of *Campylobacter* infection and may contribute to select infection with antibiotic-resistant strains.

Acknowledgments

We thank the private community and the public hospital laboratories that participated in the national surveillance of *Campylobacter* infections in human by sending their isolates to the National Reference Center for *Campylobacter* and *Helicobacter* and by notifying cases to the French Institute for Public Health Surveillance. We also thank the general practitioners who agreed to contact cases and controls.

References

1. Frost JA. Current epidemiological issues in human campylobacteriosis. Symp Ser Soc Appl Microbiol **2001**; 90:85S–95S.
2. Allos BM, Blaser MJ. *Campylobacter jejuni* and the expanding spectrum of related infections. Clin Infect Dis **1995**; 20:1092–9.
3. Saeed AM, Harris NV, DiGiacomo RF. The role of exposure to animals in the etiology of *Campylobacter jejuni/coli* enteritis. Am J Epidemiol **1993**; 137:108–14.
4. Ikram R, Chambers S, Mitchell P, Brieseman MA, Ikam OH. A case control study to determine risk factors for campylobacter infection in Christchurch in the summer of 1992–3. N Z Med J **1994**; 107:430–2.
5. Eberhart-Phillips J, Walker N, Garrett N, et al. Campylobacteriosis in New Zealand: results of a case-control study. J Epidemiol Community Health **1997**; 51:686–91.
6. Studahl A, Andersson Y. Risk factors for indigenous *Campylobacter* infection: a Swedish case-control study. Epidemiol Infect **2000**; 125:269–75.
7. Neimann J, Engberg J, Mølbak K, Wegener HC. A case-control study of risk factors for sporadic *Campylobacter* infections in Denmark. Epidemiol Infect **2003**; 130:353–66.
8. Schönberg-Norio D, Takkinen J, Hänninen ML, et al. Swimming and *Campylobacter* infections. Emerg Infect Dis **2004**; 10:1474–7.
9. Friedman CR, Hoekstra RM, Samuel M, et al. Risk factors for sporadic *Campylobacter* infection in the United States: a case-control study in FoodNet sites. Clin Infect Dis **2004**; 38(Suppl 3):S285–96.
10. Michaud S, Ménard S, Arbeit RD. Campylobacteriosis, Eastern Townships, Quebec. Emerg Infect Dis **2004**; 10:1844–7.
11. Kapperud G, Skjerve E, Bean NH, Ostroff SM, Lassen J. Risk factors for sporadic *Campylobacter* infections: results of a case-control study in southeastern Norway. J Clin Microbiol **1992**; 30:3117–21.

12. Rodrigues LC, Cowden JM, Wheeler JG, et al. The study of infectious intestinal disease in England: risk factors for cases of infectious intestinal disease with *Campylobacter jejuni* infection. Epidemiol Infect **2001**; 127: 185–93.
13. Tenkate TD, Stafford RJ. Risk factors for campylobacter infection in infants and young children: a matched case-control study. Epidemiol Infect **2001**; 127:399–404.
14. Adak GK, Cowden JM, Nicholas S, Evans HS. The Public Health Laboratory Service national case-control study of primary indigenous sporadic cases of campylobacter infection. Epidemiol Infect **1995**; 115:15–22.
15. Scott DA, Tribble DR. Protection against *Campylobacter* infection and vaccine development. In: Nachamkin I, Blaser MJ, eds. *Campylobacter.* 2nd ed. ASM Press, **2000**:303–48.
16. Frost JA, Gillespie IA, O'Brien SJ. Public health implications of *Campylobacter* outbreaks in England and Wales, 1995–9: epidemiological and microbiological investigations. Epidemiol Infect **2002**; 128:111–8.
17. Kirk M, Waddell R, Dalton C, Creaser A, Rose N. A prolonged outbreak of *Campylobacter* infection at a training facility. Commun Dis Intell **1997**; 21:57–61.
18. Mazick A, Ethelberg S, Nielsen EM, Mølbak K, Lisby M. An outbreak of *Campylobacter jejuni* associated with consumption of chicken, Copenhagen, 2005. Euro Surveill **2006**; 11:137–9.
19. Gallay A, de Valk H, Cournot M, Ladeuil B, et al. A large multi-pathogen waterborne community outbreak linked to faecal contamination of a groundwater system, France, 2000. Clin Microbiol Infect **2006**; 12:561–70.
20. Engberg J, Aarestrup FM, Taylor DE, Gerner-Smidt P, Nachamkin I. Quinolone and macrolide resistance in *Campylobacter jejuni* and *C. coli:* resistance mechanisms and trends in human isolates. Emerg Infect Dis **2001**; 7:24–34.
21. Luber P, Wagner J, Hahn H, Bartelt E. Antimicrobial resistance in *Campylobacter jejuni* and *Campylobacter coli* strains isolated in 1991 and 2001–2002 from poultry and humans in Berlin, Germany. Antimicrob Agents Chemother **2003**; 47:3825–30.
22. Gupta A, Nelson JM, Barrett TJ, et al. Antimicrobial resistance among *Campylobacter* strains, United States, 1997–2001. Emerg Infect Dis **2004**; 10:1102–9.
23. Vaillant V, de Valk H, Baron E, et al. Foodborne infections in France. Foodborne Pathog Dis **2005**; 2:221–32.
24. Gallay A, Simon F, Mégraud F. Surveillance of human Campylobacter infections in France—part 2—implementation of national surveillance. Euro Surveill **2003**; 8:218.
25. Société Française de Microbiologie (SFMG). Recommendations du Comite de l'Antibiogramme de la Société Française de Microbiologie. Available at: http://www.sfm.asso.fr/nouv/general.php?pa=2. Accessed January 2005.
26. Rubin DB. Multiple imputation after 18+ years. Am Stat Assoc **1996**; 91:473–89.
27. Royston P. Multiple imputation of missing values: update. Stata Journal **2005**; 5:188–201.
28. Fullerton KE, Ingram LA, Jones TF, et al. Sporadic campylobacter infection in infants a population-based surveillance case-control study. Pediatr Infect Dis J **2007**; 26:19–24.
29. Office National Interprofessionnel des Viandes. Consommation des produits carnés en 2003. Available at: http://www.office-elevage.fr/publications/cahier/conso03/conso03.htm. Accessed 14 March 2008.
30. Hérau V. Risque sanitaires microbiologiques lié aux effluents d'abattoirs: comparaison d'une synthèse bibliographique avec une étude de terrain. Toulouse, France: Thèse-Université Paul Sabatier, **2003**.
31. Agence Française de Sécurité Sanitaire des Aliments. Appréciation des risques alimentaires liés aux campylobacters: application au couple poulet/Campylobacter jejuni. Available at: http://www.afssa.fr/Documents/MIC-Ra-campylobacter.pdf. Accessed 14 March 2008.
32. Kapperud G, Espeland G, Wahl E, et al. Factors associated with increased and decreased risk of *Campylobacter* infection: a prospective case-control study in Norway. Am J Epidemiol **2003**; 158:234–42.
33. Cameron S, Ried K, Worsley A, Topping D. Consumption of foods by young children with diagnosed *Campylobacter* infection—a pilot case-control study. Public Health Nutr **2004**; 7:85–9.
34. Potter RC, Kaneene JB, Hall WN. Risk factors for sporadic *Campylobacter jejuni* infections in rural Michigan: a prospective case-control study. Am J Public Health **2003**; 93:2118–23.
35. Carrique-Mas J, Andersson Y, Hjertqvist M, Svensson A, Torner A, Giesecke J. Risk factors for domestic sporadic campylobacteriosis among young children in Sweden. Scand J Infect Dis **2005**; 37:101–10.
36. Agence Française de Sécurité Sanitaire des Aliments. Farm 2003-2004 French antibiotic resistance monitoring in bacteria of animal origin. Available at: http://www.afssa.fr/Documents/SANT-Ra-FARM.pdf. Accessed 14 March 2008.
37. Moore JE, Barton MD, Blair IS, et al. The epidemiology of antibiotic resistance in *Campylobacter.* Microbes Infect **2006**; 8:1955–66.
38. Swift L, Hunter PR. What do negative associations between potential risk factors and illness in analytical epidemiological studies of infectious disease really mean? Eur J Epidemiol **2004**; 19:219–23.
39. Cowden J. *Campylobacter*: epidemiological paradoxes. BMJ **1992**; 305: 132–3.
40. Black RE, Levine MM, Clements ML, Hughes TP, Blaser MJ. Experimental *Campylobacter jejuni* infection in humans. J Infect Dis **1988**; 157: 472–9.
41. Delarocque-Astagneau E, Desenclos JC, Bouvet P, Grimont PA. Risk factors for the occurrence of sporadic *Salmonella* enterica serotype enteritidis infections in children in France: a national case-control study. Epidemiol Infect **1998**; 121:561–7.
42. Ethelberg S, Olsen KE, Gerner-Smidt P, Mølbak K. Household outbreaks among culture-confirmed cases of bacterial gastrointestinal disease. Am J Epidemiol **2004**; 159:406–12.
43. Hakanen A, Jousimies-Somer H, Siitonen A, Huovinen P, Kotilainen P. Fluoroquinolone resistance in *Campylobacter jejuni* isolates in travelers returning to Finland: association of ciprofloxacin resistance to travel destination. Emerg Infect Dis **2003**; 9:267–70.
44. Ekdahl K, Giesecke J. Travellers returning to Sweden as sentinels for comparative disease incidence in other European countries, campylobacter and giardia infection as examples. Euro Surveill **2004**; 9:6–9.
45. Smith KE, Besser JM, Hedberg CW, et al. Quinolone-resistant *Campylobacter jejuni* infections in Minnesota, 1992–1998. Investigation Team. N Engl J Med **1999**; 340:1525–32.
46. Effler P, Ieong MC, Kimura A, et al. Sporadic *Campylobacter jejuni* infections in Hawaii: associations with prior antibiotic use and commercially prepared chicken. J Infect Dis **2001**; 183:1152–5.
47. Vach W, Blettner M. 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. Am J Epidemiol **1991**; 134:895–907.
48. Chavance M, Manfredi R. Modeling incomplete observations [in French]. Rev Epidemiol Sante Publique **2000**; 48:389–400.

Annexe 3

A three-source capture-recapture estimate of the number of new HIV diagnoses in children in France during 2003-2006 with multiple imputation of a variable of heterogeneous catchability.

A three-source capture-recapture estimate of the number of new HIV diagnoses in children in France during 2003-2006 with multiple imputation of a variable of heterogeneous catchability.

Corresponding author:

Vanina Héraud-Bousquet

email: v.bousquet@invs.sante.fr

Tel: 00 33 1 41 79 66 76

Fax: 00 33 1 41 79 68 02

Authors:

Vanina Héraud-Bousquet : Institut de Veille Sanitaire, Département des maladies infectieuses, St Maurice.

Florence Lot : Institut de Veille Sanitaire, Département des Maladies Infectieuses, St Maurice.

Maxime Esvan : Institut de Veille Sanitaire, Rennes.

Françoise Cazein : Institut de Veille Sanitaire, Département des Maladies Infectieuses, St Maurice.

Josiane Warszawski : Inserm CESP U1018, Le Kremlin-Bicêtre. Université Paris-Sud, Le Kremlin Bicêtre. AP-HP, Public Health Department.

Corinne Laurent : Inserm CESP U1018, Le Kremlin-Bicêtre. AP-HP, Public Health Department.

Pascale Bernillon : Institut de Veille Sanitaire, Département des Maladies Infectieuses, St Maurice.

Anne Gallay : Institut de Veille Sanitaire, Département de Coordinations des Alertes et des Régions, St Maurice.

ABSTRACT

Background: Nearly all HIV infections in children worldwide are acquired through mother-to-child transmission (MTCT) during pregnancy, labour, delivery or breastfeeding. The objective of our study was to estimate the number of new HIV diagnoses in children under 13 years of age in mainland France during 2003-2006.

Methods: We performed a capture-recapture analysis based on three sources of information: the mandatory HIV case reporting (DOVIH), the French Perinatal Cohort (ANRS-EPF) and a laboratory based surveillance of HIV (LaboVIH). Missing values of a variable of heterogenous catchability were estimated through multiple imputation. Log-linear modelling provided estimates of the number of new HIV infections in children, taking into account dependencies between sources and variables of heterogeneous catchability.

Results: The three sources reported 216 new HIV diagnoses. The number of new HIV diagnoses in children was estimated at 387 (95%CI 271-503) during 2003-2006, among whom 60% were born abroad. The estimated rate of new HIV diagnoses in children in mainland France was 9.1 per million in 2006 and was 38 times higher in children born abroad than those born in France. The estimated completeness of the three sources combined was 55.8% (CI 95% [42.9 – 79.7]) and varied according to the source : completeness of DOVIH (28.4%) and EPF (26.1%) were lower than LaboVIH (33.3%).

Conclusion: Our study provided for the first time an estimated annual rate of new HIV diagnoses in children under 13 years old in mainland France. A more systematic HIV screening of pregnant women, repeated during pregnancy among women likely to engage in risky behavior, is needed to optimize prevention of MTCT. The high prevalence of HIV infection in some regions of the world justifies the proposal of a screening test to children who migrate to France. Thus, children diagnosed as HIV infected would benefit from an early and appropriate treatment.

INTRODUCTION

Nearly all HIV infections in children worldwide are acquired through mother-to-child transmission (MTCT) during pregnancy, labour, delivery or breastfeeding. Worldwide, it has been estimated that there were 430 000 new pediatric infections in 2008 [1]. Nearly all such infections can be avoided by MTCT prevention programmes.

In France, the risk of HIV transmission from mother to child has been dramatically reduced since the end of the eighties by the prophylactic use of antiretroviral therapy (ART) during pregnancy and the administration of these drugs to the baby during the first weeks of life and is now only around 1% [2]. Early diagnosis of HIV infection during pregnancy and early treatment of the mother allow an effective prevention of mother-to-child HIV transmission. In France, the national policy since 1993 is to offer universal voluntary HIV testing in the first trimester of pregnancy.

It is estimated that about 1,500 children are living with HIV in France among a total of 150,000 people HIV positive and that 10 to 15 new HIV infections are diagnosed in newborns each year [3]. In addition, new cases of HIV-infection are regularly diagnosed in children born abroad (in high endemic HIV countries), after arrival in France, for whom no estimates are currently available.

In this paper we estimate the total number of new HIV diagnoses in children under 13 years old in mainland France during the 2003-2006 period using capture-recapture methods. We used three data sources: the mandatory HIV case reporting (DOVIH), the French perinatal cohort (ANRS-EPF) and the HIV laboratory surveillance system (LaboVIH). We also assessed the completeness of the mandatory reporting of HIV in children and the French perinatal cohort.

METHODS

The capture-recapture method allows to estimate the total number of cases of a disease by matching cases reported in at least two sources [4].

Case definition

Cases were defined as new HIV diagnoses in children under 13 years old, according to biological criteria [3], in mainland France during the 2003-2006 period.

Description of the three data sources

The mandatory HIV case reporting (DOVIH)

The mandatory HIV case reporting was implemented in 2003 by the French Institute for Public Health Surveillance (InVS) to follow the epidemic trends and to describe characteristics of HIV infections newly diagnosed [5]. Any HIV positive diagnosis confirmed for the first time by a microbiologist has to be notified with a unique anonymous code for each person allowing the detection of duplicates. Clinicians complete a notification form with epidemiological and clinical data. For children under 13 years old, the case reporting is based on pediatricians. Notifications are sent to the district health authorities and then transmitted to the InVS. All information are entered into a national database. To take into account reporting delays, cases were selected in the data set including notifications up to the 31st March 2010.

The ANRS French Perinatal Cohort (ANRS-EPF CO1/CO10/CO11)

The French perinatal cohort, supported by the French National Agency for AIDS Research (ANRS) has prospectively collected data on HIV-infected pregnant women and their children in around one hundred centers throughout France since 1984 [2]. Informed consent is obtained from all mothers. Inclusion criteria were extended since 2005 to all children <13 years old recently diagnosed for HIV in the pediatric sites borne to mothers not included in the EPF, with parental consent. Semestrial clinical and biological examinations are collected since birth to 18

years for infected children and to 24 months for non infected children. Duplicates are deleted. The objectives of this cohort are to identify factors associated with HIV MTCT, to evaluate tolerance to ART prophylaxis among pregnant women and their babies, and to assess the prognosis of pediatric HIV infection. Cases meeting the inclusion criteria were selected in the database updated in April 2008.

The laboratory surveillance (LaboVIH)

Since 2001, the InVS has implemented a national surveillance of the HIV testing activity in France. The number of HIV tests performed and the number of HIV positive confirmed diagnoses for the first time are collected from the 4,200 French microbiological laboratories each year [6].

Among the participating laboratories (85% response rate), those who reported at least one new HIV diagnosis in children under 15 years old from 2003 to 2006 were included in our survey. Those 137 laboratories were asked to complete a questionnaire in order to collect individual information for each pediatric diagnosis. Among these 137 laboratories, 113 reported pediatric HIV cases. Duplicate notifications were identified and deleted.

Identification of common cases between sources

Because no common identification code was available among the three sources, algorithms were set up using variables common to all three sources in order to detect common cases. Date of birth, sex, reference hospital (or district number) and date of diagnosis (or date of the first medical care) were available in all three sources. The algorithm detecting common cases between the sources DOVIH and EPF included also the maternity of birth for children born in France or the country of birth for others, the mother's country of origin and the vital status of the children.

The identification of common cases between sources was performed with the « SQL » procedure in SAS© 9.1 version and was completed by a manual verification of matched records.

Imputation of the variable "country of birth" in the source LaboVIH

We wanted to estimate the total number of new HIV diagnoses according to the place of birth: "born in France" or "born abroad". This binary variable was not collected in the source LaboVIH. However it was partly available for cases of LaboVIH that matched the two others sources of information (DOVIH and the EPF) in which this information had been collected. The overall proportion of missing values for this variable was around 30% (66/216) globally and varied according to sources: 59.6% (64/129) in LaboVIH, 1.8% (2/110) in DOVIH and 0.0 % in EPF. We estimated missing values through a multiple imputation (MI) method, which consists in using the distribution of the observed data to estimate a set of plausible values for the missing observations [7]. Multiple data sets were created and an estimate was calculated for each imputed data set. The estimates were then combined to obtain overall estimates, variances and confidence intervals.

The MI method applied was multiple imputation by chained equations using Stata's user written program *ice* (STATA ® 11.0, Stata Corporation, College Station, Texas, USA) {Royston, 2009 169 /id;Van Buuren S., 1999 20 /id}. The variables "age" (continuous), "region of diagnosis" (categorical) and "year of diagnosis" (categorical) were completely observed and used as predictors in the imputation model. As the variable "place of birth" contained 30% of missing values, we chose to generate 100 imputed data sets.

Capture-recapture estimates

The reliability of the estimates depends on the respect of conditions including (1) the absence of false positive cases, (2) a closed population, (3) identification of all and only true common

cases, (4) independence between sources and (5) capture homogeneity [4]. Two sources are independent when the probability for a case to be reported in a source does not depend on the probability to be reported in the other source. For three or more sources analyses, the independence assumption is not any more required as interaction terms can be incorporated into regression models to adjust for source dependencies. Homogeneity of capture is fulfilled when the probability for a case to be reported in a source is the same for all cases, that is to say the probability of notification does not depend on its characteristics (i.e., age, sex, place of birth etc.). This probability may vary from one source to another or be constant overall [4].

Dependence between sources was first assessed by comparing the estimates provided by each pair of sources [10, 11] and by calculating the odds ratio between two sources among cases notified to the other one (and its 95 %CI) as proposed by Wittes [4].

A preliminary three-source analysis was performed by fitting 8 log-linear models to the data arranged in a 2^3 contingency table according to case presence or absence in each source. The dependent variable for each model is the logarithm of the number of cases in each of the 7 non-empty cells of the contingency table. These preliminary analyses assumed homogeneity of capture within each source and were performed using the Stata's user written program "recap" [12]. The latter is a STATA module providing standard three-source capture-recapture analyses without covariates. Confidence interval estimates for the population size are computed according to a goodness-of-fit based method proposed by Regal et Hook [13].
Then, three variables of potential heterogeneous catchability were considered: place of birth (born in France; born abroad), region of diagnosis (Paris area; other regions), and year of diagnosis (2003 to 2006). Data were then arranged in a 2^3x2x2x4 contingency table. Log-linear models were fitted via the STATA 'glm' command specifying a logarithmic link and a Poisson distribution. A manual backward stepwise variable selection procedure was performed starting with the model including all two-way interactions between sources, between sources and

variables of catchability, and between variables of catchability. Log-linear modelling was jointly performed for the 100 imputed data sets using the Stata 11.0 analysis module "mi estimate" applying Rubin's rules.

Population size estimates, calculated as a sum of exponentiated regression coefficients, were obtained through commands specific to MI. Their respective variances were estimated using the delta method. Confidence intervals were computed using Student's t statistics with degrees of freedom specific to each coefficient depending both on the number of imputations and on the proportion of missing values.

Classically, in capture-recapture studies, the choice of the final model is based on the likelihood ratio test statistic (G^2), the Akaike Information Criterion (AIC) and the Bayesian Information Criterion adapted by Draper (DIC) which are functions of the likelihood ratio statistic [14, 15]. AIC and DIC criteria were derived for each imputed data set according to the following formulas: $AIC = G^2 - 2(df)$ and $DIC = G^2 - \left(\ln\left(N_{obs}/2\pi\right)\right)\cdot\left(df\right)$ where df is the number of degrees of freedom associated with any model.

The naïve approach averaging the likelihood ratio statistic over the imputed data sets does not provide accurate p-values [16]. The pooled likelihood ratio test statistic and its corresponding p-value were calculated using the Meng and Rubin approach [17], recently illustrated by Marshall *et al.* [18]. Each log-linear model was constrained to the regression coefficients obtained from the joint analysis (i.e. the average over the 100 imputed data sets according to Rubin's rules). The AIC and DIC estimates were the average of the 100 AICs and DICs. We selected the most parcimonious model among models with a goodness-of-fit p-value >0.05, and with the lowest AIC and DIC values.

This method, by selecting a single model, ignores the issue of "model uncertainty" [15]. We calculated a composite estimate derived from all the *K* models involved in the stepwise procedure as suggested by Draper [14]. We retained the DIC weighted estimate ("Weighted DIC") which is expressed as $\hat{N}_{WDIC} = \sum_{i=1}^{K} \hat{N}_i \cdot e^{-(DICi/2)} \Big/ \sum_{i=1}^{K} e^{-(DICi/2)}$, where $\hat{N}_i$ is the estimate associated with the model *i,* and DIC_i is the DIC associated with the model *i (i=1,…,K)*.

Source completenesses were estimated by dividing the number of new HIV diagnoses reported in each source by the total number estimated by the final log-linear model. The completeness was also calculated for each stratum of "place of birth", "year of diagnosis" and "region of diagnosis".

The annual rate of new HIV diagnoses was the estimated number of new HIV diagnoses divided by the size of the population of children under 13 years old living in mainland France up to December 2007 [19]. The rate was also calculated according to the place of birth, knowing the number of children less than 13 born in France or abroad.

RESULTS

Cross-matches

The three sources reported 216 new HIV diagnoses in children under 13 years old, in mainland France, between January the 1st, 2003 and December the 31st, 2006 (Figure 1). After imputation of the variable "place of birth", 60% of the children newly diagnosed were born abroad (Table 1). The distribution of new HIV diagnoses according to the place of birth and the age group varied between sources. The proportion of children born abroad in children aged over 1 year was greater in the source DOVIH and in EPF cohort than in LaboVIH.

Capture-recapture estimates

When performing two-source capture-recapture analysis, the estimate of the number of new HIV diagnoses provided by matching sources DOVIH and EPF (N_{est} = 188; 95%CI [171 – 206]) was lower than the estimate provided by matching LaboVIH and EPF (N_{est} = 261; 95%CI [224-297]) and LaboVIH and DOVIH (N_{est} = 330; 95%CI 272-389), suggesting a positive dependence between the sources DOVIH and EPF. The Wittes odds ratio confirmed the dependence between the sources DOVIH and EPF (OR = 5,4 ; 95%CI [2,5-12,1]) and suggested a positive dependence between LaboVIH and EPF (OR = 2,2 ; 95%CI [1.0-4,8]).

Preliminary log-linear modelling using the three sources and including the dependencies between sources provided an estimate of 369 (95%CI [294-521] new HIV diagnoses during the 2003-2006 period (Table 2). This model took into account two dependencies between sources (DOVIH*EPF and EPF*LaboVIH).

When considering the dependencies with variables of catchability, the model with the lowest AIC and a likelihood ratio test with p>0.05 provided an estimate of 387 (95%CI [271-503]) new HIV diagnoses during the same period (Table 3). This model (model 8) included two interactions between sources, interactions between sources and variable of catchability (DO*place of birth, EPF*place of birth, DO*region of diagnosis, EPF*region of diagnosis, LaboVIH* region of diagnosis, and EPF*year of diagnosis). The corresponding weighted estimate was 384 new HIV diagnoses. The estimated annual number of new HIV diagnoses decreased over time from 108 in 2003 to 89 in 2006 (Table 4).

The estimated completeness of the three sources combined was 55.8% (CI 95% [42.9 – 79.7]) but varied according to the source (Table 4). Completeness of DOVIH (28.4%) and EPF(26.1%) were lower than LaboVIH (33.3%). Completeness slightly decreased in both DOVIH and EPF since 2004, and particularly during the last year (2006). Completeness was greater in the Paris

area than in other regions in the three sources, and was greater for children born in France rather than abroad in the sources EPF and LaboVIH.

Based on the estimated number of new diagnosis obtained in table 4, the rate of new HIV diagnoses in children under 13 years old in mainland France was 9.1 per million (CI 95% [5.7 – 12.5]) in 2006. This annual rate was 38 times higher in children born abroad (161.1 per million) than in children born in France (4.2 per million).

DISCUSSION

Our study provided for the first time an estimate of the total number of new HIV diagnoses in children under 13 years old in mainland France, during the 2003-2006 period (n = 387). The three sources combined reported more than two-thirds of cases. Nonetheless, the completeness of the mandatory notification system (DOVIH) and the French perinatal cohort (EPF) were low (<30%).

Estimating missing values

The variable 'place of birth' was not recorded in the source LaboVIH but was nearly complete for the other two sources. Usually, the standard approach in capture-recapture is to ignore variables not available in every source, which often leads to biased estimates of the population size [20]. One approach commonly used to analyze incomplete data sets is to fill in each missing data with an imputed value and analyze the data set as if it were complete. Such methods of single imputation are not statistically valid, may yield biased estimates, and lead to underestimated variances [21]. Two methods are currently recommended to handle missing values adequately, maximum likelihood estimation and multiple imputation (MI). These methods are asymptotically equivalent and require the same assumption that the data are missing at random (MAR), meaning that the missing data mechanism depends on observed values only [7,

22]. In our study, the variable “place of birth” was missing by design in the source LaboVIH. As a result, the missingness provides no information about the underlying process, implying that the MAR assumption is met.

Only few studies reported imputation of unobserved values in capture-recapture applications [20, 23, 24], and maximum likelihood estimation using an Expectation Maximisation (EM) algorithm was applied. Van der Heidjen *et al*. [24] estimated missing values for variables of heterogeneous catchability that were not collected in all sources such as gender and region of residence. The authors stressed that the Expectation Maximization (EM) algorithm often involves complex numerical integration (step E) and that multiple imputation has the advantage of being computationally much simpler for most practical situations (specially for incomplete continuous variables) [25].

Model selection and estimation

The final model selection in the stepwise selection analysis including variables of catchability was based on the AIC and DIC, provided the goodness of fit of this model is correct according to the likelihood ratio test. The approach proposed by Meng et Rubin was applied to carry out the likelihood ratio test, and provided p-values slightly lower than the naïve approach (data not shown). A limitation of our study is that AIC/DIC criteria have been obtained by averaging their values over the imputed data sets and should be interpreted with caution [16]. Differences between models according to these criteria may be overestimated, and may lead to select a model too complex. Nevertheless, in our application, the model selected according to the AIC/DIC (model 9) was simpler than the one selected according to the likelihood ratio statistic (model 8), and the overall estimates provided by these two models were identical (387.0). We retained the less parcimonious model (model 8) to ensure that the adequacy of the model was correct according to the likelihood ratio statistic (p=0.07). Moreover, model 8 included an interaction between EPF and year of diagnosis, which appeared sensible.

An advantage of MI is that standard errors and CI of the estimates are directly available as part of the model estimation. A parametric bootstrap approach has been recommended to calculate CI for the final estimate [26, 27]. This method yields asymetrical CI, and allows one to take model uncertainty into account. Future research should adress the possibility to combine this approach with multiple imputation.

Capture-recapture method – conditions of application

Our estimates should be considered with caution, since all the conditions of application of the capture-recapture method have not been met [4]. The algorithms of identification of the common cases between sources combined several criteria. Thus, false positive or false negative matches might have been induced, leading respectively to an under- or overestimated number of new HIV diagnoses. However, because of the small number of common cases, we were able to validate all identified matches by checking all other available variables, minimising the number of false positive. Cross-match with the source LaboVIH was performed according to a set of few variables which could generate false-negative matches. The complementary investigation for the source LaboVIH was performed only among laboratories participating to the LaboVIH surveillance and may have missed HIV diagnoses in non participating laboratories (15%). Moreover, only the participating laboratories which reported HIV diagnoses in children were required to give details on each case, assuming the other laboratories had not done any diagnosis in children. Finally, 20% of the laboratories did not answer the complementary investigation.

The study period and the geographic area were the same for all the sources. However, it was estimated that the EPF cohort covered 70% of the HIV-positive pregnant women which could have introduced a bias in the estimation. However, if such a bias did exist, it would be difficult to surmise its direction (towards an over or under-estimation).

The positive dependence between the sources DOVIH and EPF may be explained by the heightened paediatricians' awareness that participate in the EPF cohort for several years, to report to the mandatory notification implemented in 2003. Two large laboratories participated to both the EPF cohort and the laboratory survey which could result in a positive dependence between the sources EPF and LaboVIH.

Estimates of number of new HIV diagnoses

Among the 89 estimated new HIV diagnoses in children under 13 years old in 2006, about 40 occurred in children born in France. This estimate is more than the double of the number of about 15 commonly reported [3]. The latter estimate is based on the following reasoning: 1,500 HIV-positive pregnant women deliver each year with a MTCT rate of 1% if the pregnant woman is known as HIV-positive. However, this calculation does not take into account women who are not tested for HIV during pregnancy, and women who seroconvert during pregnancy following a first negative test, with a much higher risk of transmission. In absence of any prevention strategy, the HIV MTCT rate was around 20% in France before 1994 [28]. Such situations were identified in a retrospective analysis of children diagnosed with HIV infection at Necker Hospital in Paris [29].

Our capture-recapture findings allowed us to estimate the rate of new HIV diagnoses in children in mainland France at 9.1 per million in 2006. This rate was 38 times greater for children born abroad than those born in France. This ratio is higher than that observed in adults (6,0 per million in adults born abroad versus 0,6 in those born in France) [6] and could be explained by lower access to HIV screening and to prevention of MTCT during pregnancy in HIV-endemic countries.

Our results can be compared to data of the United Kingdom as both countries have similar sized populations (around 60 millions of whom 10 millions of children), similar concentrated HIV epidemics and similar sized foreign born populations (around 8% of the total population and around 0.5 million from sub-Saharan Africa). The rate of new HIV diagnoses in children under 15 in UK in 2006 was close but greater (10.1 per million) than our estimate in France. This is likely to be due to different HIV prevalence rates within countries of origin (Eastern or Southern Africa versus Western or Central African countries where HIV prevalence is lower). The number of new diagnoses has decreased in UK from 2003 to 2006 (from 148 to 117) as in France and has continued to decline since [30]. As in France, two-thirds of children diagnosed as HIV-infected in the UK were born abroad [31].

Completenesses

The completeness of the mandatory notification of new HIV diagnoses in children was low (28%) compared to the estimated completeness of the overall system (62% in 2004) [6]. This difference might be due to the participation of biologists to the mandatory notification of HIV diagnoses adults only, allowing the recall of clinicians who have not reported the cases. The notification system of HIV infection in children was modified in 2007: biologists have also to report new HIV diagnoses in children. However, low completeness and modification of the surveillance system make it difficult to assess trends in new HIV diagnoses since 2007.

Several hypotheses may explain the low completeness for HIV diagnoses in children in EPF (26%) as well as the capture heterogeneity according to the place of birth (children born abroad less reported than children born in France). As mentioned above, around 70% of HIV pregnant women are included in the EPF cohort. The EPF cohort has prospectively collected data on HIV-infected children until 2004 only if they were born from mothers known as HIV-infected at delivery in EPF sites. Data on infected children born from mothers not included in EPF, that is who delivered abroad or who have not been diagnosed as HIV-infected during pregnancy, have

been collected retrospectively for 2003 and 2004, and prospectively since 2005. Thus, cases may have been missed. Moreover, obtaining parental consent after HIV diagnosis in children requires more time and motivation from paediatricians than asking consent to pregnant women.

CONCLUSION

Our study provided for the first time an estimated annual rate of new HIV diagnoses in children under 13 years old in mainland France. A more systematic HIV screening of pregnant women, repeated during pregnancy among women likely to engage in risky behavior, is needed to optimize prevention of MTCT. The high prevalence of HIV infection in some regions of the world, especially in sub-Saharan Africa, justifies the proposal of a screening test to children who migrate to France. Thus, children diagnosed as HIV infected would benefit from an early and appropriate treatment. Notification of new HIV diagnoses in children should also be improved in order to better describe the evolving epidemiology of HIV infection in children.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

VHB contributed to study conception, carried out the statistical analysis and prepared the final draft of the manuscript. AG conceived and supervised the study, and helped to draft the manuscript. FL participated in the design of the study and coordination, and helped to draft the manuscript. FC, JW, and CL participated to the data gathering and contributed to interpretation of the study. ME participated in the design of the study, contributed to the acquisition of data, and helped performing the statistical analysis. PB supervised the statistical analysis and helped to draft the manuscript. All authors critically revised and approved the final manuscript.

Acknowledgements

We thank all the laboratories which participate in LaboVIH surveillance, all the paediatricians who report notifications of HIV diagnoses in children, and all the investigators and participating centers of the ANRS French Perinatal Cohort.

The French Institute for Public Health Surveillance conducted this study as part of his surveillance activities and is funded by the French Ministry of Health. The French perinatal cohort is funded by the French National Agency for AIDS Research (ANRS). Vanina Héraud-Bousquet is funded by a doctoral grant by the French National Agency for AIDS Research [n° NM/DF/1754].

Reference List

1. UNAIDS, WHO. **AIDS epidemic update.** December 2009.

2. Warszawski J, Tubiana R, Le Chenadec J, Blanche S, Teglas JP, Dollfus C, Faye A, Burgard M, Rouzioux C, Mandelbrot L: **Mother-to-child HIV transmission despite antiretroviral therapy in the ANRS French Perinatal Cohort**. *AIDS* 2008, **22:**289-299.

3. Yeni P: **Prise en charge médicale des personnes infectées par le VIH**: Médecine-Sciences Flammarion; 2010.

4. Hook EB, Regal RR: **Capture-recapture methods in epidemiology: methods and limitations**. *Epidemiol Rev* 1995, **17:**243-264.

5. Lot F, Semaille C, Cazein F, Barin F, Pinget R, Pillonel J, Desenclos JC: **Preliminary results from the new HIV surveillance system in France**. *Euro Surveill* 2004, **9:**34-37.

6. Cazein F, Le Vu S, Pillonel J, Le Strat Y, Couturier S, Basselier B, Lot F, Semaille C: **Dépistage de l'infection par le VIH en France, 2003-2009.** *Bulletin Epidemiologique Hebdomadaire* 2010, **45-46:**451-454.

7. Little RJA, Rubin DB: *Statistical analysis with missing data.* 2nd ed. New York: Wiley; 2002.

8. Royston P: **Multiple imputation of missing values: further update of ice, with an emphasis on categorical variables.** *Stata J* 2009, **9:**466-477.

9. Van Buuren S., Boshuizen HC, Knook DL: **Multiple imputation of missing blood pressure covariates in survival analysis**. *Stat Med* 1999, **18:**681-694.

10. Chapman DG: **Some properties of the hypergeometric distribution**. *University of California* 1951, **1:**131-160.

11. Seber GA: **The effects of trap response on tag recapture estimates**. *Biometrics* 1970**:**13-22.

12. van der Heiden M: **Stata module to perform capture-recapture analysis for three sources with goodness-of-fit based confidence intervals. Available from http://ideas.repec.org/c/boc/bocode/s456859.html.**; 2007.

13. Regal RR, Hook EB: **Goodness-of-fit based confidence intervals for estimates of the size of a closed population**. *Stat Med* 1984, **3:**287-291.

14. Draper D: **Assessment and propagation of model uncertainty**. *J R Stat Soc [B]* 1995, **57:**45-70.

15. Hook EB, Regal RR: **Validity of methods for model selection, weighting for model uncertainty, and small sample adjustment in capture-recapture estimation**. *Am J Epidemiol* 1997, **145:**1138-1144.

16. White IR, Royston P, Wood AM: **Multiple imputation using chained equations: Issues and guidance for practice**. *Stat Med* 2011, **30:**377-399.

17. Meng X, Rubin D: **Performing likelihood ratio tests with multiply-imputed data sets.** *Biometrika* 1992, **79:**103-111.

18. Marshall A, Altman DG, Holder RL, Royston P: **Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines**. *BMC Med Res Methodol* 2009, **9:**57.

19. **National Institute of Statistics and Economic Studies. National population census. http://www.insee.fr/fr/themes/detail.asp?reg_id=0&ref_id=ir-sd2008&page=irweb/sd2008/dd/sd2008_population.htm (accessed Mai 10, 2011).**; 2011.

20. Zwane EN, van der Heijden PG: **Analysing capture--recapture data when some variables of heterogeneous catchability are not collected or asked in all registrations**. *Stat Med* 2007, **26:**1069-1089.

21. Allison PD: *Missing data*. Iowa City: Sage Publication; 2002.

22. Schafer JL, Graham JW: **Missing data: our view of the state of the art**. *Psychol Methods* 2002, **7:**147-177.

23. Robb ML, Bohning D: **Imputing unobserved values with the EM algorithm under left and right-truncation, and interval censoring for estimating the size of hidden populations**. *Biom J* 2011, **53:**75-87.

24. van der Heijden PG, Zwane E, Hessen E: **Structurally missing data problems in multiple list capture-recapture data.** *A StA Adv Stat Anal* 2009, **93:**5-21.

25. Zwane E, van der Heijden PG: **Capture-recapture studies with incomplete mixed categorical and continuous covariates.** *J Data Sci* 2008, **6:**557-572.

26. Buckland ST, Burnham KP, Augustin NH: **Model selection: an integral part of inference**. *Biometrics* 2009, **53:**603-618.

27. Sutherland JM, Schwarz CJ, Rivest LP: **Multilist population estimation with incomplete and partial stratification**. *Biometrics* 2007, **63:**910-916.

28. Mayaux MJ, Blanche S, Rouzioux C, Le CJ, Chambrin V, Firtion G, Allemon MC, Vilmer E, Vigneron NC, Tricoire J, .: **Maternal factors associated with perinatal HIV-1 transmission: the French Cohort Study: 7 years of follow-up observation. The French Pediatric HIV Infection Study Group**. *J Acquir Immune Defic Syndr Hum Retrovirol* 1995, **8:**188-194.

29. Macassa E, Burgard M, Veber F, Picard C, Neven B, Malhaoui N, Rouzioux C, Blanche S: **Characteristics of HIV-infected children recently diagnosed in Paris, France**. *Eur J Pediatr* 2006, **165:**684-687.

30. Health Public Agency. United Kingdom. **New HIV diagnoses data to end December 2010.** Tables N°2; 2010.

31. Health Public Agency. **HIV in the United Kingdom:2010. Health protection report.** 4(47) edition; 2010.

Figures legend

Figure 1. Distribution of new HIV diagnoses in children identified by the three sources (N=216).

Table 1. Distribution of new HIV diagnoses after multiple imputation according to place of birth and age.

	Total	DOVIH		EPF		LaboVIH	
		< 1 year	≥ 1 year	< 1 year	≥ 1 year	< 1 year	≥ 1 year
Born abroad	130 (60.2%)	1 (12.5%)	24 (85.8%)	1 (12.5%)	9 (75%)	4 (33.3%)	37 (71.2%)
Born in France	86 (39.8%)	7 (87.5%)	4 (14.3%)	7 (87.5%)	3 (25%)	8 (66.7%)	15 (28.8%)
	216	8	28	8	12	12	52

DOVIH : Mandatory notification system for HIV; EPF: French peri-natal survey; LaboVIH : Survey evaluating the screening activity for HIV in laboratories

Table 2. Preliminary log-linear analyses assuming homogeneity of capture within each source.

Log-linear models	$\hat{n}$	$\hat{N}$	95% CI	df	G^2	p	AIC	DIC
Dependencies between sources								
LaboVIH*DOVIH, LaboVIH*EPF, DOVIH*EPF	126	342	259,573	0	0	1	0	0
LaboVIH*DOVIH, LaboVIH*EPF	23	239	225,263	1	18.83	$<10^{-4}$	16.83	16.89
LaboVIH*DOVIH, DOVIH*EPF	58	274	243,331	1	3.78	0.05	1.78	1.84
LaboVIH*EPF, DOVIH*EPF	153	369	294,521	1	0.24	0.63	-1.76	-1.71
LaboVIH*DOVIH, EPF	29	249	234,272	2	18.49	$<10^{-4}$	14.49	14.6
LaboVIH*EPF, DOVIH	51	267	245,300	2	30.12	$<10^{-4}$	26.12	26.23
DOVIH*EPF, LaboVIH	85	301	268,349	2	5.96	0.05	1.96	2.07
LaboVIH, DOVIH, EPF	49	265	246,292	3	30.2	$<10^{-4}$	24.20	24.36

DOVIH : Mandatory HIV case reporting; EPF: French perinatal cohort; LaboVIH : Survey evaluating the screening activity for HIV in laboratories

$\hat{n}$: Estimate of the number of diagnoses not reported to any source; $\hat{N}$: Estimate of the number of diagnoses; 95% CI : 95% confidence interval for $\hat{N}$; df: number of degrees of freedom; G^2 : deviance statistic; p: p-value of the deviance goodness-of-fit test; AIC : Akaike Information Criterion; DIC : Draper Information Criterion

Table 3. Log-linear analyses incorporating variables of potential heterogeneous catchability.

Log-linear models	$\hat{n}$	$\hat{N}$	95% CI ($\hat{N}$)	df	G^2	p	AIC	DIC
Model 1: DO*LABO, DO*EPF, LABO*EPF,DO*place, Labo*place, EPF*place, DO*region, Labo*region, EPF*region, DO*year, Labo*year, EPF*year, place*region, place*year, region*year.	108.3	324.3	210.9,437.6	78	93.84	0.07	-62.16	-130.85
Model 2: DO*LABO, DO*EPF, LABO*EPF,DO*place, EPF*place, DO*region, Labo*region, EPF*region, DO*year, Labo*year, EPF*year, place*region, place*year, region*year.	106.9	322.9	211.2,434.5	79	93.95	0.08	-64.05	-133.62
Model 3: DO*LABO, DO*EPF, LABO*EPF,DO*place, EPF*place, DO*region, Labo*region, EPF*region, DO*year, Labo*year, EPF*year, place*region, place*year.	108.6	324.6	211.7,437.5	82	96.17	0.10	-67.83	-140.05
Model 4: DO*EPF, LABO*EPF,DO*place, EPF*place, DO*region, Labo*region, EPF*region, DO*year, Labo*year, EPF*year, place*region, place*year.	148.0	364.0	257.9,470.1	83	96.82	0.10	-69.18	-142.27
Model 5: DO*EPF, LABO*EPF,DO*place, EPF*place, DO*region, Labo*region, EPF*region, DO*year, Labo*year, EPF*year, place*year.	152.3	368.3	259.9,476.7	84	98.31	0.10	-69.69	-143.66
Model 6: DO*EPF, LABO*EPF,DO*place, EPF*place, DO*region, Labo*region, EPF*region, Labo*year, EPF*year, place*year.	143.3	359.3	259.1,459.6	87	102.28	0.09	-71.72	-148.34
Model 7: DO*EPF, LABO*EPF,DO*place, EPF*place, DO*region, Labo*region, EPF*region, EPF*year, place*year.	140.8	356.8	258.0,455.6	90	105.54	0.09	-72.46	-153.71
Model 8: DO*EPF, LABO*EPF,DO*place, EPF*place, DO*region, Labo*region, EPF*region, EPF*year.	171.0	387.0	271.0,502.9	93	112.72	0.07	-73.28	-155.17
Model 9: DO*EPF, LABO*EPF,DO*place, EPF*place, DO*region, Labo*region, EPF*region.	171.0	387.0	271.0,502.9	96	118.48	0.05	-73.52	-158.06
Weighted DIC		384.26						

DOVIH : Mandatory HIV case reporting; EPF: French perinatal cohort; LaboVIH : Survey evaluating the screening activity for HIV in laboratories

$\hat{n}$: Estimate of the number of diagnoses not reported to any source; $\hat{N}$: Estimate of the number of diagnoses; 95% CI : 95% confidence interval for $\hat{N}$; df: number of degrees of freedom; G^2 : deviance statistic; p: p-value of the deviance goodness-of-fit test; AIC : Akaike Information Criterion; DIC : Draper Information Criterion

Weighted DIC estimate: weighted average of all estimates ($\hat{N}$) provided by each model (p>0.05)

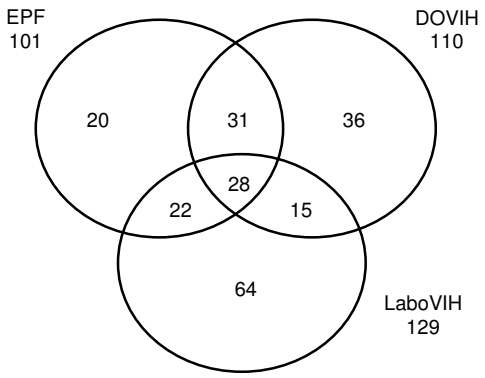
Table 4. Estimates of completeness of each source (model 8).

Strata	Total					DOVIH			EPF			LaboVIH		
	$\hat{N}$	(95% CI)	N_{obs}	Compl(%)	(95% CI)	N_{obs}	Compl(%)	(95% CI)	N_{obs}	Compl(%)	(95% CI)	N_{obs}	Compl(%)	(95% CI)
Year of diagnosis														
2003	107.6	(72.4;142.7)	60	55.8	(42.0:82.9)	30	27.9	(21.0;41.4)	28	26.0	(19.6:38.7)	30	27.9	(21.0;41.4)
2004	99.1	(68.9;129.4)	59	59.5	(45.6;85.7)	35	35.3	(27.0;50.8)	32	32.3	(24.7;46.5)	35	35.3	(27.0;50.8)
2005	91.7	(62.4;120.9)	53	57.8	(43.8;85.0)	27	29.5	(22.3;43.3)	27	29.5	(22.3;43.3)	34	37.1	(28.1;54.5)
2006	88.6	(55.4;121.8)	44	49.7	(36.1;79.4)	18	20.3	(14.8;32.5)	14	15.8	(11.5;25.3)	30	33.9	(24.6;54.2)
Place of birth														
France	152.6	(100.3;204.9)	86	56.4	(42.0;85.7)	37	24.2	(18.1;36.9)	47	30.8	(22.9;46.8)	55	36.0	(26.8;54.8)
foreign country	234.4	(158.9;309.9)	130	55.5	(42.0;81.8)	73	31.1	(23.6;45.9)	54	23.0	(17.4;34.0)	74	31.6	(23.9;46.6)
Region of diagnosis														
Paris area	198.1	(154.7;241.4)	139	70.2	(57.6;89.9)	79	39.9	(32.7;51.1)	79	39.9	(32.7;51.1)	82	41.4	(34.0;53.0)
other regions	188.9	(101.0;276.8)	77	40.8	(27.8;76.2)	31	16.4	(11.2;30.7)	22	11.6	(7.9;21.8)	47	24.9	(17.0;46.5)
Total	387	(271;503)	216	55.8	(42.9;79.7)	110	28.4	(21.9;40.1)	101	26.1	(20.1;37.3)	129	33.3	(25.6;47.6)

DOVIH : Mandatory notification system for HIV; EPF: French peri-natal survey; LaboVIH : Survey evaluating the screening activity for HIV in laboratories

$\hat{N}$: Estimate of the number of diagnoses; N_{obs} : Number of diagnoses observed; Compl : Completeness; 95% CI : 95% confidence interval for completeness

Figure 1



DOVIH : Mandatory notification system for HIV
EPF : French peri-natal survey
LaboVIH : Survey evaluating the screening activity for HIV in laboratories

Annexe 4

Hepatitis C virus genotype 3 and the risk of severe liver disease in a large population of drug users in France.

Journal of Medical Virology 82:1647–1654 (2010)

Hepatitis C Virus Genotype 3 and the Risk of Severe Liver Disease in a Large Population of Drug Users in France

Christine Larsen,[1]* Vanina Bousquet,[1] Elisabeth Delarocque-Astagneau,[1,2] Corinne Pioche,[1] Françoise Roudot-Thoraval,[3] the HCV surveillance steering committee, the HCV surveillance group and Jean-Claude Desenclos

[1]*Institut de veille sanitaire (InVS), Saint-Maurice, France*
[2]*Institut Pasteur, Unité d'épidémiologie des maladies émergentes, Paris, France*
[3]*Centre hospitalier universitaire, APHP-Henri Mondor, Créteil, France*

Although risk factors for cirrhosis in chronic hepatitis C virus (HCV) infection have been identified, the role of HCV-genotype 3 remains controversial, and limited data are available in drug users. The aim of the study was to assess risk factors for severe liver disease (cirrhosis/hepatocellular carcinoma) in HCV-infected drug users between 2001 and 2007 in France. Patients who reported drug use and who had been referred for HCV infection to hepatology centers from a national surveillance system were identified. The severity of liver disease was assessed clinically and histologically (Metavir score). Factors associated with severe liver disease were analyzed after estimating missing values by multiple imputation (MI). Of the 4,065 drug users naive to anti-HCV treatment who were referred to the 26 participating centers, 8.0% had severe liver disease, 25.7% were infected with HCV-genotype 3. Factors associated independently with an increased risk of severe liver disease were HCV-genotype 3 (adjusted odds ratio, multiple imputation (aOR_{MI}) = 1.6, [95% confidence interval, 95% CI: 1.2–2.1]), HIV infection (aOR_{MI} = 1.8, [1.2–2.8]), male sex (aOR_{MI} = 2.0, [1.4–2.8]), age over 40 years (aOR_{MI} = 2.1, [1.6–2.9]), history of excessive alcohol consumption (aOR_{MI} = 2.8, [2.1–3.7]), and duration of infection $\geq$18 years (aOR_{MI} = 2.9, [2.0–4.3]). This analysis shows that HCV-genotype 3 is associated with severe liver disease in drug users, independently of age, sex, duration of infection, alcohol consumption, and co-infection with HIV. These results are in favor of earlier treatment for drug users infected with HCV- genotype 3 and confirm the need for concomitant care for excessive alcohol consumption. ***J. Med. Virol. 82:1647–1654, 2010.*** © 2010 Wiley-Liss, Inc.

KEY WORDS: alcohol; cirrhosis; drug users; HCV genotype 3; HIV

INTRODUCTION

In France, the prevalence of anti-hepatitis C virus (HCV) in injecting drug users ranges from 55% to 60% [Jauffret-Roustide et al., 2009; Meffre et al., 2010]. In addition, injecting drug users accounted for more than two-thirds of the patients co-infected with HIV and HCV [Larsen et al., 2008].

Factors associated with progression of liver fibrosis have been well established, and are related to host factors (age at contamination, male sex), comorbidities (co-infection with HIV, heavy alcohol consumption), metabolic conditions (obesity, steatosis, and diabetes mellitus), and duration of infection [Massard et al.,

Christine Larsen, Vanina Bousquet, Elisabeth Delarocque-Astagneau have contributed equally.

The Hepatitis C surveillance steering committee (France): P. Couzigou (Bordeaux), P. Marcellin (Clichy), F. Roudot-Thoraval (Créteil), P. Hillon (Dijon), JP. Zarski (Grenoble), JP. Bronowicki (Nancy), E. Delarocque-Astagneau (Paris), C. Sylvain (Poitiers), D. Guyader (Rennes), O. Goria (Rouen).

The Hepatitis C surveillance group (France): D. Capron (Amiens); P. Cales, I. Hubert-Fouchard (Angers); V. Di Martino (Besançon); J. Foucher (Bordeaux); C. Guillemard (Caen); A. Abergel (Clermont-Ferrand); MP. Ripault (Clichy); D. Dhumeaux (Créteil); A. Minello (Dijon); A. Edouard (Fort de France); MN. Hilleret (Grenoble); V. Canva (Lille), V. Loustaud-Rati (Limoges); P. Pradat, C. Trepo (Lyon); JJ. Raabe (Metz); D. Larrey (Montpellier); J. Gournay (Nantes); E. Marine-Barjoan, A. Tran (Nice); X. Causse (Orléans); B. Nalpas, S. Pol (Paris); G. Thiefin (Reims); H. Danielou (Rennes); K. Barange (Toulouse); L. D'Alteroche, EH. Metman (Tours).

Grant sponsor: Institut de veille sanitaire (French Ministry of Health); Grant sponsor: Agence Nationale de Recherches sur le Sida et les hépatites virales (ANRS).

*Correspondence to: Christine Larsen, Institut de veille sanitaire, Département des maladies infectieuses, 12 rue du Val d'Osne, 94415 Saint-Maurice Cedex, France.
E-mail: c.larsen@invs.sante.fr

Accepted 20 April 2010
DOI 10.1002/jmv.21850
Published online in Wiley Online Library (wileyonlinelibrary.com)

2006; Mallat et al., 2008]. Furthermore, co-infection with HIV and a history of excessive alcohol consumption are linked to death at an earlier age in patients infected chronically with HCV as shown in a recent study on mortality related to chronic hepatitis C in France [Marcellin et al., 2008]. Viral induced steatosis linked to HCV genotype 3 has also emerged as a cofactor for progression of liver fibrosis in patients infected with HCV [Rubbia-Brandt et al., 2004]. HCV genotype 3 is also the most frequent genotype found among drug users infected with HCV in France, accounting for more than one third [Payan et al., 2005]. Excessive alcohol consumption is common in this population, too [Campbell et al., 2006]. However, limited data are available on factors associated with progression to liver fibrosis in former or current injecting drug users infected by HCV.

In addition, with new antiviral treatment options and efficacy, recommendations on antiviral treatment of drug users infected with HCV have been expanded in France [Dhumeaux et al., 2003] since the 2002 consensus conference on treatment of HCV infection.

In France, a surveillance system was set up to monitor changes over time of the epidemiologic and clinical characteristics of patients infected with HCV at the first referral to hepatology reference centers; nearly 3,900 cases of chronic hepatitis C, of whom 35% acknowledged drug use, have been recorded each year since 2001 by this system. The aim of the present study was to assess risk factors associated with severe liver disease in patients infected with HCV who had reported drug use using the data collected by this surveillance system.

METHODS

Study Population and Inclusion Criteria

Details of the methods have been described previously [Delarocque-Astagneau et al., 2005]. Briefly, this surveillance system based on voluntary participation of 26 of the 30 existing hepatology reference centers was set up in 2000 to monitor changes over time of the epidemiologic and clinical characteristics of patients infected with HCV at the time of their first referral. Hepatology reference centers are located in University hospitals throughout France. To be included, a case was defined as a patient with anti-HCV antibodies who was referred (first contact as an in- or out-patient) to any participating reference center. Patients can be referred by their general practitioner, by a specialist or by self-referral. Patients included in the analysis were those who reported having injected or snorted drugs at least once in their lifetime. To ascertain the first referral, patients with no prior histologic evaluation of their liver disease at referral were selected for the analysis. The surveillance system obtained ethical approval from the French data protection authority (CNIL).

Data Collection

A standardized notification form was used in all centers to collect data routinely, including risk factors for HCV transmission, year of the last HCV negative test (if available), date and the reasons for anti-HCV testing, HCV RNA serum status, HCV genotype, co-infections with HIV (anti-HIV serostatus) and hepatitis B virus (hepatitis B surface antigen serostatus), history of excessive alcohol consumption defined as more than 210 g of pure ethanol per week for women and 280 g for men, and the number of years with excessive consumption. The severity of liver disease was assessed either by liver biopsy (Metavir score [Bedossa and Poynard, 1996]) and/or by clinical evaluation including clinical examination, biochemical tests, and imaging (mainly abdominal ultrasound). The clinical evaluation was graded as follows: repeated normal ALT values defined as ALT levels below the upper limit of normal at three consecutive tests within six months, chronic hepatitis C, cirrhosis (compensated or not), and hepatocellular carcinoma. Cirrhosis at clinical evaluation was defined by the fibrosis Metavir score F4 when a liver biopsy was carried out or by the association of clinical signs, laboratory findings and ultrasound images used by all participating centers [Schuppan and Afdhal, 2008]. The diagnostic criteria used for hepatocellular carcinoma were those established during the Barcelona-2000 EASL Conference [Bruix et al., 2001].

The form was modified in 2004 to include data on non-invasive assessment of liver fibrosis and on antiviral therapy.

HCV genotyping was carried out by the referral laboratories of the participating centers. Most of these laboratories were involved in the French ANRS panel quality control study in 2000 [Lefrere et al., 2004] and the same standardized genotyping assays (Inno LiPA assay) or other methods (sequencing, primer-specific PCR) during the study period were used.

The database was checked regularly for duplicate reporting in and between participating reference centers.

The suspected year of HCV infection was defined by the year of the last HCV negative test performed during the drug-use period or year of first drug injection. Time from the suspected year of infection to the year of referral to the reference center was used as a surrogate of the duration of HCV infection. The median duration of infection was used to create a qualitative variable ($\geq$18 years; $<$18 years) in the univariate and multivariate analysis.

Patients included in the analysis were drug users with HCV RNA in the serum who were naive to anti-HCV treatment.

The outcome of severe liver disease was defined as cirrhosis (compensated or not) or hepatocellular carcinoma at the first referral, based on the clinical assessment.

Statistical Analysis

Factors associated with severe liver disease. Univariate analysis was carried out to assess the characteristics associated with severe liver disease at clinical

evaluation. Variables with P values <0.2 in univariate analysis were entered into the multivariate analysis. STATA® 9.2 statistical package was used for the analysis (Stata Corporation, College Station, TX).

First, the general approach for multivariate analysis was used including only patients who had no missing data in any of the variables (complete case analysis). However, six out of the eight variables included in the multivariate model comprised 10–27% of missing values resulting in a loss of power and precision, complication in data handling and analysis and potentially biased odds ratios (OR) due to differences between the observed and unobserved data [Little and Rubin, 2002]. Therefore, missing values were estimated from the observed values using multiple imputation (MI) (imputation by chained equations, ice procedure STATA® 9.2) [Royston, 2007]. Several complete data sets (m = 30) were created. Each data set was analyzed separately, and the results were combined to obtain the OR [Little and Rubin, 2002]. The sensitivity of the OR to departures from the assumed missing mechanism was tested for three variables (alcohol consumption, co-infection with HIV, and HCV genotype 3) using a weighting approach developed by Carpenter et al. [2007].

RESULTS

Of the 5,004 drug users who were referred for HCV infection between 2001 and 2007 and who were naive to anti-HCV treatment, 4,236 were HCV RNA positive and were included in the analysis. The remaining 768 had either a negative (n = 455) or an unspecified (n = 313) result for HCV RNA.

Characteristics of the Study Population

Of the 4,236 patients positive for HCV RNA, 77% were male; the median age was 39 years and the median estimated duration of HCV infection was 18 years (Table I). Co-infections with HBV and HIV were diagnosed in 2% and 6%, respectively. A history of excessive alcohol consumption was reported for 43%. Among the 1,154 patients who reported heavy drinking in the past with an available duration of excessive consumption, 67% had drunk heavily for more than 5 years. HCV genotypes were mainly genotype 1 (36%) and 3 (25%), the distribution of HCV genotypes remained stable over the study period.

A liver biopsy was carried out in 53.6% of the drug users included in 2001–2003 and in 27.1% of those included in 2004–2007. In the latter period, 29.2% had an evaluation of fibrosis by non-invasive methods only.

A clinical evaluation of liver disease was available for 4,065 (96%) patients with HCV RNA, and severe liver disease was diagnosed in 8% of the patients. Cirrhosis was diagnosed in 301 patients and hepatocellular carcinoma in 10 patients. Of these 311 patients, 236 (76%) had an available HCV genotyping result: HCV genotypes 1 and 3 were found in 48% and 42%, respectively. Of the 164 patients with severe liver disease at clinical evaluation who underwent a liver biopsy, the Metavir score F4 was diagnosed for 94% including three out of the four patients with hepatocellular carcinoma (F3 for the remaining patient).

In univariate analysis the following variables were associated with a severe liver disease at referral (Table II): inclusion after 2003, male sex, age >40 years at referral, estimated duration of infection ≥18 years, history of excessive alcohol intake, and co-infections with HBV, HIV, and HCV genotype 3. Time between the first HCV positive test and referral was not significant.

In both the complete-case and the MI analyses, the independent factors associated with severe liver disease at clinical evaluation were male sex (adjusted OR_{MI} (aOR_{MI}): 2.0; 95% confidence interval, 95% CI, 1.4–2.8), age >40 years at referral (aOR_{MI}: 2.1; 95% CI, 1.6–2.9), duration of HCV infection ≥18 years (aOR_{MI}: 2.9; 95% CI, 2.0–4.3), history of excessive alcohol consumption (aOR_{MI}: 2.8; 95% CI, 2.1-3.7), and HCV genotype 3 (aOR_{MI}: 1.6; 95% CI, 1.2–2.1) (Table III). In the MI analysis, co-infection with HIV (aOR_{MI}: 1.8; 95% CI, 1.2–2.8) was associated significantly with severe liver disease. No significant interactions were found.

The univariate analysis for the factors associated with the F4 Metavir score, limited to the patients who had a liver biopsy (n = 1,725), showed similar results although the 95% CI were wider. However, co-infections with HIV and HBV were not significant statistically ($P > 0.2$; data not shown).

DISCUSSION

This multi-site study documents factors associated with severe liver disease in a large group of patients infected with HCV who reported drug use. The findings confirm the role of excessive alcohol consumption and co-infection with HIV and suggest that HCV genotype 3 is associated with an increased risk of severe liver disease independently of age, sex, duration of infection, and excessive alcohol consumption. To our knowledge this is the first study to address the risk factors of severe liver disease in a population of drug users infected with HCV at first referral.

Clinical studies have demonstrated a significant association between HCV genotype 3 infection and the presence of steatosis [Hui et al., 2002; Rubbia-Brandt et al., 2004; Cross et al., 2009]. In patients infected with HCV genotype 3, fat accumulation in the liver is correlated with the level of HCV replication [Adinolfi et al., 2001] and disappears following successful antiviral therapy [Castera et al., 2004]. Although there have been several studies providing evidence for the influence of steatosis on the progression of fibrosis regardless of the HCV genotype [Hu et al., 2004; Fartoux et al., 2005; Cross et al., 2009] or in patients infected with HCV genotype 3 [Castera et al., 2003, 2004; Rubbia-Brandt et al., 2004], the issue remains controversial. Indeed, other studies failed to show any association between steatosis and liver fibrosis [Asselah et al., 2003; Ryder et al., 2004; Perumalswami et al., 2006], including one

TABLE I. Main Characteristics of HCV RNA Positive Drug Users at First Referral to Hepatology Reference Centers in France, 2001–2007

Characteristics	All patients (n = 4,236)	HCV genotype 3 (n = 1,077)	Other HCV genotypes (n = 1,986)
Sex ratio (female/male)	1/3	1/3	1/3
Age, years	39 (10)	39 (11)	40 (10)
Duration of HCV infection[a], years	18 (12)	18 (12)	19 (12)
ALT level[b], xULN	1.5 (1.5)	2.0 (1.9)	1.5 (1.2)
History of excessive alcohol intake[c]			
Missing	467 (11.0)	119 (11.1)	212 (10.7)
No	1,952 (46.1)	499 (46.3)	908 (45.7)
Yes	1,817 (42.9)	459 (46.3)	866 (43.6)
HIV serostatus			
Missing	702 (16.6)	162 (15.0)	289 (14.6)
Negative	3,284 (77.5)	868 (80.6)	1,557 (78.4)
Positive	250 (5.9)	47 (4.4)	140 (7.0)
HBsAg serostatus			
Missing	685 (16.2)	150 (13.9)	285 (14.3)
Negative	3,466 (81.8)	906 (84.1)	1,660 (83.6)
Positive	85 (2.0)	21 (2.0)	41 (2.1)
HCV genotype			
Missing	1,173 (27.7)		
1 non a, b	407 (9.6)		
1a	694 (16.4)		
1b	439 (10.3)		
2	117 (2.8)		
3 non a	317 (7.5)		
3a	760 (17.9)		
4	325 (7.7)		
5	4 (0.1)		
Clinical stage at 1st referral			
Missing	171 (4.1)	31 (2.9)	58 (2.9)
Normal ALT values	713 (16.8)	116 (10.8)	356 (17.9)
Chronic hepatitis	3,000 (70.8)	815 (75.6)	1,419 (71.4)
Cirrhosis[d]	301 (7.1)	97 (9.0)	132 (6.7)
Hepatocellular carcinoma	10 (0.2)	3 (0.3)	4 (0.2)
Acute hepatitis	41 (1.0)	15 (1.4)	17 (0.9)
Liver biopsy at 1st referral	1,765 (41.7)	405 (37.6)	992 (49.9)
Fibrosis Metavir score at liver biopsy [Bedossa and Poynard, 1996]	1,725	396	976
F0	138 (8.0)	25 (6.3)	81 (8.3)
F1	735 (42.6)	136 (34.3)	455 (46.6)
F2	480 (27.8)	112 (28.3)	264 (27.1)
F3	195 (11.3)	62 (15.7)	94 (9.6)
F4	177 (10.3)	61 (15.4)	82 (8.4)

HCV, hepatitis C virus; IQR, interquartile range; SD, standard deviation; ALT, alanine aminotransferase; ULN, upper limit of normal; HIV, human immunodeficiency virus; HBsAg, hepatitis B surface antigen.
Data are no. (%) of patients or median values (interquartile range), unless otherwise indicated.
[a]Available for 3,626 (85.6%) HCV RNA+ drug users, for 932 (86.5%) infected with genotype 3, for 1,673 (84.2%) infected with other genotypes.
[b]Available for 3,938 (93.0%) HCV RNA+ drug users, for 1,023 (95.0%) infected with genotype 3, for 1,906 (96.0%) infected with other genotypes.
[c]>210 g/week for women and >280 g/week for men (g, gram of ethanol).
[d]Including decompensated cirrhosis.

study on a large cohort of patients infected with HCV [Perumalswami et al., 2006]. However, the comparability of the studies is questionable, as there were differences in the study populations and the risk factors studied. Furthermore, another study showed that steatosis had no association with fibrosis in patients infected with HCV genotype 3 [Bugianesi et al., 2006] suggesting other causes for the progression of fibrosis such as insulin resistance or viral induced liver inflammation.

Numerous studies have shown that a past or present history of excessive alcohol consumption was associated with more rapid progression to fibrosis [Westin et al., 2002; Tolstrup et al., 2009]. As expected in this particular population of patients who reported drug use, cirrhosis at the time of first referral was associated strongly with excessive alcohol consumption in the past. The definition of excessive alcohol consumption used in this study corresponds to more than 30 g/day for women and 40 g/day for men compared with 50 g/day for both sexes in many other studies. Of those who acknowledged excessive alcohol consumption in the past, more than half reported that the period lasted at least 5 years, allowing the development of fibrosis. In addition, more than one third of the patients who reported excessive alcohol consumption in the past, acknowledged current excessive consumption. This is consistent with the results reported by Campbell et al.

TABLE II. Factors Associated With Severe Liver Disease at Clinical Evaluation in HCV RNA Positive Drug Users at First Referral to Hepatology Reference Centers in France, 2001–2007

Factors	Patients (n = 4,065)	% SLD	Univariate analysis	
			OR (95% CI)	*P*
Period of inclusion				
2001–2003	2,211	6.8	1	
2004–2007	1,854	8.7	1.3 (1.0–1.6)	0.02
Sex				
Female	929	4.0	1	
Male	3,136	8.7	2.3 (1.6–3.3)	<0.001
Age				
≤40 years	2,302	3.7	1	
>40 years	1,763	12.8	3.8 (2.9–4.9)	<0.001
Time between 1st HCV+ test and referral				
<1 year	1,683	6.8	1	
≥1 year	1,987	7.7	1.1 (0.9–1.5)	0.3
Missing	395	10.6		
Duration of HCV infection at referral[a]				
<18 years	1,636	3.0	1	
≥18 years	1,843	12.5	4.7 (3.4–6.5)	<0.001
Missing	586	7.7		
History of excessive alcohol intake[b]				
No	1,881	4.1	1	
Yes	1,740	12.5	3.3 (2.5–4.3)	<0.001
Missing	444	3.6		
HBsAg				
Negative	3,366	7.8	1	
Positive	83	13.2	1.8 (0.9–3.4)	0.07
Missing	616	6.0		
HIV				
Negative	3,180	7.7	1	
Positive	240	12.9	1.8 (1.2–2.7)	0.004
Missing	645	5.6		
HCV genotype 3				
No	1,928	7.0	1	
Yes	1,046	9.6	1.4 (1.1–1.8)	0.02
Missing	1,091	6.9		

SLD, severe liver disease (cirrhosis, hepatocellular carcinoma); aOR, odds ratio; CI, confidence interval; HCV, hepatitis C virus; HBsAg, hepatitis B surface antigen; HIV, human immunodeficiency virus.
[a]Time from suspected year of infection to year of referral to the reference center. Suspected year of HCV infection is defined as year of the last HCV negative test performed during the drug-use period or year of first drug injection.
[b]>210 g/week for women and >280 g/week for men (g, gram of ethanol).

[2006], who found that in a population of drug users infected with HCV, knowledge about the detrimental effects of alcohol consumption related to HCV infection did not appear to influence alcohol use.

The analysis of the mortality data in the Swiss hepatitis C cohort showed recently that mortality was increased for individuals who reported drinking more than 40 g alcohol per day. HCV genotype 3 was associated also with a higher mortality although it did not remain significant in the multivariate analysis [Prasad et al., 2009].

HIV infection is known to modify the natural history of HCV by accelerating the rate of progression of fibrosis and the development of advanced fibrosis both before [Poynard et al., 2003] and since [Thein et al., 2008] the highly active antiretroviral therapy era. Interestingly, a study has shown advanced liver fibrosis to be associated with HCV genotype 3 in a population of patients co-infected with HIV and HCV [Barreiro et al., 2006]. Surprisingly, in the study by Barreiro et al. [2006], neither past nor current excessive alcohol consumption remained associated significantly with advanced liver fibrosis. However, the assessment of liver fibrosis in this study was based on a non-invasive method (elastometry) only.

Among the limitations of the present study is its cross-sectional design. Indeed, it is difficult to use cross-sectional data to estimate longitudinal parameters. However, although collected at the same time as the outcome of interest (severe liver disease), most of the variables assessed in this study reflect exposures that occurred prior to the outcome. Therefore, this study could be considered as a retrospective cohort, which is less subject to bias than a purely cross-sectional study [Rothman and Greenland, 1998]. Other factors shown to be associated with progression of fibrosis were not collected, such as daily cannabis smoking [Hezode et al., 2005] and insulin resistance related to the patient characteristics such as obesity and type 2 diabetes mellitus, and to the virus [Cua et al., 2008; Moucari et al., 2008; Petta et al., 2008]. However, Moucari et al. [2008] provided evidence that insulin resistance was

TABLE III. Factors Independently Associated With Severe Liver Disease at Clinical Evaluation in HCV RNA Positive Drug Users at First Referral to Hepatology Reference Centers in France, 2001–2007

	Multivariate analysis			
	Complete case (n = 2,294)		Multiple imputation (n = 4,065)	
Factors	aOR (95% CI)	*P*	aOR (95% CI)	*P*
Sex				
Female	1		1	
Male	2.0 (1.2–3.1)	0.005	2.0 (1.4–2.8)	<0.001
Age				
≤40 years	1		1	
>40 years	2.4 (1.6–3.6)	<0.001	2.1 (1.6–2.9)	<0.001
Duration of HCV infection at referral[a]				
<18 years	1		1	
≥18 years	2.9 (1.8–4.6)	<0.001	2.9 (2.0–4.3)	<0.001
History of excessive alcohol intake[b]				
No	1		1	
Yes	2.6 (1.8–3.6)	<0.001	2.8 (2.1–3.7)	<0.001
HCV genotype 3				
No	1		1	
Yes	1.4 (1.0–1.9)	0.05	1.6 (1.2–2.1)	0.003
HIV serostatus[c]				
Negative	—	—	1	
Positive	—	—	1.8 (1.2–2.8)	0.005

aOR, adjusted odds ratio; CI, confidence interval; HCV, hepatitis C virus; HBsAg, hepatitis B surface antigen; HIV, human immunodeficiency virus.
Severe liver disease (cirrhosis, hepatocellular carcinoma).
[a]Time from suspected year of infection to year of referral to the reference center. Suspected year of HCV infection is defined as year of the last HCV negative test performed during the drug-use period or year of first drug injection.
[b]>210 g/week for women and >280 g/week for men (g, gram of ethanol).
[c]HIV serostatus did not remain significant in the complete case analysis.

associated with HCV genotypes 1 and 4 in contrast to patients with HCV genotypes 2 and 3. This lower rate of insulin resistance in patients with HCV genotype 3 was shown also in another study [Hui et al., 2003]. The validity of the method used in the present study for assessing the severity of liver disease could be called into question since it was not based on histology. However, the classification used for clinical evaluation of the liver disease is likely to be homogeneous within the participating reference centers. Furthermore, as the indication for liver biopsy may be related to certain risk factors or the intention to begin a course of antiviral therapy, the use of the clinical stage for assessment of severe liver disease appeared more suited to the purpose of the study objectives. Also, the proportion of patients managed in a reference center who had a liver biopsy decreased dramatically from 45% in 2001 to 18% in 2006 [Delarocque-Astagneau et al., 2010]. Indeed, since 2002, physicians in France have been treating patients infected with HCV genotypes 2 and 3 or with HIV co-infection without performing a liver biopsy, in accordance with the 2002 French recommendations on HCV treatment [Dhumeaux et al., 2003]. The univariate analysis of the factors associated with the F4 Metavir score limited to patients who had a liver biopsy showed similar results, but the 95% CI were wider and co-infection with HIV was not significant statistically.

The present analysis was confronted with missing values as it was based on surveillance data. In order to keep all cases in the analysis, a MI method was applied to the data. MI is a well-established method for the analysis of data sets with missing values. As the multivariate analysis will include all cases, MI provides consistency in results [Little and Rubin, 2002]. Thus, the association between HCV genotype 3 and severe liver disease found in the complete case analysis was reinforced by the MI analysis. Also, HIV infection was associated significantly with severe liver disease after MI analysis. In addition, the sensitivity analysis performed on HCV genotype 3, co-infection with HIV and alcohol consumption allowed to be confident of the robustness of the estimations [Carpenter et al., 2007][data not shown]. A recent analysis of the data from the Swiss hepatitis C cohort found HCV genotype 3 as a factor associated with progression of liver fibrosis, which reinforced the validity of the results [Bochud et al., 2009].

These results suggest that in patients infected chronically with HCV who reported drug use, early therapeutic intervention should be recommended for those patients who are infected with HCV genotype 3 and who acknowledge excessive alcohol consumption in the past, since progression to cirrhosis may be more rapid. Therefore, integrated treatment of HCV and alcohol addiction is essential. Thus, collaboration between healthcare providers who specialize in the management of addictions and those who specialize in the management of HCV infection should be promoted.

ACKNOWLEDGMENTS

We thank the clinicians and patients from all the participating hepatology reference centers and Dr. Elisabeth Couturier for her critical reviews of the manuscript.

REFERENCES

Adinolfi LE, Utili R, Andreana A, Tripodi MF, Marracino M, Gambardella M, Giordano M, Ruggiero G. 2001. Serum HCV RNA levels correlate with histological liver damage and concur with steatosis in progression of chronic hepatitis C. Dig Dis Sci 46:1677–1683.

Asselah T, Boyer N, Guimont MC, Cazals-Hatem D, Tubach F, Nahon K, Daikha H, Vidaud D, Martinot M, Vidaud M, Degott C, Valla D, Marcellin P. 2003. Liver fibrosis is not associated with steatosis but with necroinflammation in French patients with chronic hepatitis C. Gut 52:1638–1643.

Barreiro P, Martin-Carbonero L, Nunez M, Rivas P, Morente A, Simarro N, Labarga P, Gonzalez-Lahoz J, Soriano V. 2006. Predictors of liver fibrosis in HIV-infected patients with chronic hepatitis C virus (HCV) infection: Assessment using transient elastometry and the role of HCV genotype 3. Clin Infect Dis 42: 1032–1039.

Bedossa P, Poynard T. 1996. An algorithm for the grading of activity in chronic hepatitis C. The METAVIR Cooperative Study Group. Hepatology 24:289–293.

Bochud PY, Cai T, Overbeck K, Bochud M, Dufour JF, Mullhaupt B, Borovicka J, Heim M, Moradpour D, Cerny A, Malinverni R, Francioli P, Negro F. 2009. Genotype 3 is associated with accelerated fibrosis progression in chronic hepatitis C. J Hepatol 51:655–666.

Bruix J, Sherman M, Llovet JM, Beaugrand M, Lencioni R, Burroughs AK, Christensen E, Pagliaro L, Colombo M, Rodes J. 2001. Clinical management of hepatocellular carcinoma. Conclusions of the Barcelona-2000 EASL conference. European Association for the Study of the Liver. J Hepatol 35:421–430.

Bugianesi E, Marchesini G, Gentilcore E, Cua IH, Vanni E, Rizzetto M, George J. 2006. Fibrosis in genotype 3 chronic hepatitis C and nonalcoholic fatty liver disease: Role of insulin resistance and hepatic steatosis. Hepatology 44:1648–1655.

Campbell JV, Hagan H, Latka MH, Garfein RS, Golub ET, Coady MH, Thomas DL, Strathdee SA. 2006. High prevalence of alcohol use among hepatitis C virus antibody positive injection drug users in three US cities. Drug Alcohol Depend 81:259–265.

Carpenter JR, Kenward MG, White IR. 2007. Sensitivity analysis after multiple imputation under missing at random: A weighting approach. Stat Methods Med Res 16:259–275.

Castera L, Hezode C, Roudot-Thoraval F, Bastie A, Zafrani ES, Pawlotsky JM, Dhumeaux D. 2003. Worsening of steatosis is an independent factor of fibrosis progression in untreated patients with chronic hepatitis C and paired liver biopsies. Gut 52:288–292.

Castera L, Hezode C, Roudot-Thoraval F, Lonjon I, Zafrani ES, Pawlotsky JM, Dhumeaux D. 2004. Effect of antiviral treatment on evolution of liver steatosis in patients with chronic hepatitis C: Indirect evidence of a role of hepatitis C virus genotype 3 in steatosis. Gut 53:420–424.

Cross TJ, Quaglia A, Hughes S, Joshi D, Harrison PM. 2009. The impact of hepatic steatosis on the natural history of chronic hepatitis C infection. J Viral Hepat 16:492–499.

Cua IH, Hui JM, Kench JG, George J. 2008. Genotype-specific interactions of insulin resistance, steatosis, and fibrosis in chronic hepatitis C. Hepatology 48:723–731.

Delarocque-Astagneau E, Roudot-Thoraval F, Campese C, Desenclos JC, The Hepatitis CSSS. 2005. Past excessive alcohol consumption: A major determinant of severe liver disease among newly referred hepatitis C virus infected patients in hepatology reference centers, France, 2001. Ann Epidemiol 15:551–557.

Delarocque-Astagneau E, Meffre C, Dubois C, Pioche C, Le Strat Y, Roudot-Thoraval F, Hillon P, Silvain C, Dhumeaux D, Desenclos JC. 2010. The impact of the prevention programme of hepatitis C over more than a decade: The French experience. J Viral Hepat 17: 435–443.

Dhumeaux D, Marcellin P, Lerebours E. 2003. Treatment of hepatitis C. The 2002 French consensus. Gut 52:1784–1787.

Fartoux L, Chazouilleres O, Wendum D, Poupon R, Serfaty L. 2005. Impact of steatosis on progression of fibrosis in patients with mild hepatitis C. Hepatology 41:82–87.

Hezode C, Roudot-Thoraval F, Nguyen S, Grenard P, Julien B, Zafrani ES, Pawlotsky JM, Dhumeaux D, Lotersztajn S, Mallat A. 2005. Daily cannabis smoking as a risk factor for progression of fibrosis in chronic hepatitis C. Hepatology 42:63–71.

Hu KQ, Kyulo NL, Esrailian E, Thompson K, Chase R, Hillebrand DJ, Runyon BA. 2004. Overweight and obesity, hepatic steatosis, and progression of chronic hepatitis C: A retrospective study on a large cohort of patients in the United States. J Hepatol 40:147–154.

Hui JM, Kench J, Farrell GC, Lin R, Samarasinghe D, Liddle C, Byth K, George J. 2002. Genotype-specific mechanisms for hepatic steatosis in chronic hepatitis C infection. J Gastroenterol Hepatol 17:873–881.

Hui JM, Sud A, Farrell GC, Bandara P, Byth K, Kench JG, McCaughan GW, George J. 2003. Insulin resistance is associated with chronic hepatitis C virus infection and fibrosis progression [corrected]. Gastroenterology 125:1695–1704.

Jauffret-Roustide M, Le SY, Couturier E, Thierry D, Rondy M, Quaglia M, Razafandratsima N, Emmanuelli J, Guibert G, Barin F, Desenclos JC. 2009. A national cross-sectional study among drug-users in France: Epidemiology of HCV and highlight on practical and statistical aspects of the design. BMC Infect Dis 9:113.

Larsen C, Pialoux G, Salmon D, Antona D, Le SY, Piroth L, Pol S, Rosenthal E, Neau D, Semaille C, Delarocque-Astagneau E. 2008. Prevalence of hepatitis C and hepatitis B infection in the HIV-infected population of France, 2004. Euro Surveill 13.

Lefrere JJ, Roudot-Thoraval F, Lunel F, Alain S, Chaix ML, Dussaix E, Gassin M, Izopet J, Pawlotsky JM, Payan C, Stoll-Keller F, Thibault V, Trabaud MA, Bettinger D, Bogard M, Branger M, Buffet-Janvresse C, Charrois A, Defer C, Laffont C, Lerable J, Levayer T, Martinot-Peignoux M, Mercier B, Rosenberg AR. 2004. Expertise of French laboratories in detection, genotyping, and quantification of hepatitis C virus RNA in serum. J Clin Microbiol 42:2027–2030.

Little RJ, Rubin DB. 2002. Statistical analysis with missing data. 2nd edition. Hoboken, NJ: John Wiley & Sons, Inc.

Mallat A, Hezode C, Lotersztajn S. 2008. Environmental factors as disease accelerators during chronic hepatitis C. J Hepatol 48:657–665.

Marcellin P, Pequignot F, Delarocque-Astagneau E, Zarski JP, Ganne N, Hillon P, Antona D, Bovet M, Mechain M, Asselah T, Desenclos JC, Jougla E. 2008. Mortality related to chronic hepatitis B and chronic hepatitis C in France: Evidence for the role of HIV coinfection and alcohol consumption. J Hepatol 48:200–207.

Massard J, Ratziu V, Thabut D, Moussalli J, Lebray P, Benhamou Y, Poynard T. 2006. Natural history and predictors of disease severity in chronic hepatitis C. J Hepatol 44:S19–S24.

Meffre C, Le Strat Y, Delarocque-Astagneau E, Dubois F, Antona D, Lemasson JM, Warszawski J, Steinmetz J, Coste D, Meyer JF, Leiser S, Giordanella JP, Gueguen R, Desenclos JC. 2010. Prevalence of hepatitis B and hepatitis C virus infections in France in 2004: Social factors are important predictors after adjusting for known risk factors. J Med Virol 82:546–555.

Moucari R, Asselah T, Cazals-Hatem D, Voitot H, Boyer N, Ripault MP, Sobesky R, Martinot-Peignoux M, Maylin S, Nicolas-Chanoine MH, Paradis V, Vidaud M, Valla D, Bedossa P, Marcellin P. 2008. Insulin resistance in chronic hepatitis C: Association with genotypes 1 and 4, serum HCV RNA level, and liver fibrosis. Gastroenterology 134: 416–423.

Payan C, Roudot-Thoraval F, Marcellin P, Bled N, Duverlie G, Fouchard-Hubert I, Trimoulet P, Couzigou P, Cointe D, Chaput C, Henquell C, Abergel A, Pawlotsky JM, Hezode C, Coude M, Blanchi A, Alain S, Loustaud-Ratti V, Chevallier P, Trepo C, Gerolami V, Portal I, Halfon P, Bourliere M, Bogard M, Plouvier E, Laffont C, Agius G, Silvain C, Brodard V, Thiefin G, Buffet-Janvresse C, Riachi G, Grattard F, Bourlet T, Stoll-Keller F, Doffoel M, Izopet J, Barange K, Martinot-Peignoux M, Branger M, Rosenberg A, Sogni P, Chaix ML, Pol S, Thibault V, Opolon P, Charrois A, Serfaty L, Fouqueray B, Grange JD, Lefrere JJ, Lunel-Fabiani F. 2005. Changing of hepatitis C virus genotype patterns in France at the beginning of the third millenium: The GEMHEP GenoCII Study. J Viral Hepat 12:405–413.

Perumalswami P, Kleiner DE, Lutchman G, Heller T, Borg B, Park Y, Liang TJ, Hoofnagle JH, Ghany MG. 2006. Steatosis and progression of fibrosis in untreated patients with chronic hepatitis C infection. Hepatology 43:780–787.

Petta S, Camma C, Di M V, Alessi N, Cabibi D, Caldarella R, Licata A, Massenti F, Tarantino G, Marchesini G, Craxi A. 2008. Insulin resistance and diabetes increase fibrosis in the liver of patients with genotype 1 HCV infection. Am J Gastroenterol 103:1136–1144.

Poynard T, Mathurin P, Lai CL, Guyader D, Poupon R, Tainturier MH, Myers RP, Muntenau M, Ratziu V, Manns M, Vogel A, Capron F, Chedid A, Bedossa P. 2003. A comparison of fibrosis progression in chronic liver diseases. J Hepatol 38:257–265.

Prasad L, Spicher VM, Negro F, Rickenbach M, Zwahlen M. 2009. Little evidence that hepatitis C virus leads to a higher risk of mortality in the absence of cirrhosis and excess alcohol intake: The Swiss Hepatitis C Cohort Study. J Viral Hepat 16:644–649.

Rothman K, Greenland S. 1998. Types of epidemiologic studies. In: Modern epidemiology. Philadelphia: Lippincott Raven. pp. 67–78.

Royston P. 2007. Multiple imputation of missing values: Further update of ice, with an emphasis on interval censoring. Stata Journal 7:445–464.

Rubbia-Brandt L, Fabris P, Paganin S, Leandro G, Male PJ, Giostra E, Carlotto A, Bozzola L, Smedile A, Negro F. 2004. Steatosis affects chronic hepatitis C progression in a genotype specific way. Gut 53: 406–412.

Ryder SD, Irving WL, Jones DA, Neal KR, Underwood JC. 2004. Progression of hepatic fibrosis in patients with hepatitis C: A prospective repeat liver biopsy study. Gut 53:451–455.

Schuppan D, Afdhal NH. 2008. Liver cirrhosis. Lancet 371:838–851.

Thein HH, Yi Q, Dore GJ, Krahn MD. 2008. Natural history of hepatitis C virus infection in HIV-infected individuals and the impact of HIV in the era of highly active antiretroviral therapy: A meta-analysis. AIDS 22:1979–1991.

Tolstrup JS, Gronbaek M, Tybjaerg-Hansen A, Nordestgaard BG. 2009. Alcohol intake, alcohol dehydrogenase genotypes, and liver damage and disease in the Danish general population. Am J Gastroenterol 104:2182–2188.

Westin J, Lagging LM, Spak F, Aires N, Svensson E, Lindh M, Dhillon AP, Norkrans G, Wejstal R. 2002. Moderate alcohol intake increases fibrosis progression in untreated patients with hepatitis C virus infection. J Viral Hepat 9:235–241.

Annexe 5

Practical considerations for sensitivity analysis after multiple imputation applied to epidemiological studies with incomplete data.

Practical considerations for sensitivity analysis after multiple imputation applied to epidemiological studies with incomplete data.

Soumis à BMC Medical Research & Methodology

Authors:
Christine Larsen, Institut de Veille Sanitaire, Département des Maladies Infectieuses.
James Carpenter, London School of Hygiene & Tropical Medicine, Medical Statistics Unit.
Jean-Claude Desenclos, Institut de Veille Sanitaire, Direction Scientifique.
Yann Le Strat, Institut de Veille Sanitaire, Département des Maladies Infectieuses.

Corresponding author:
Dr. Vanina Heraud-Bousquet
Mail: Institut de Veille Sanitaire, Département des maladies infectieuses,12 rue du Val d'Osne, 94415 St Maurice, France.
Email: v.bousquet@invs.sante.fr
Telephone: 00 33 1 41 79 69 76
Fax: 00 33 1 41 79 68 02

Running head: Sensitivity analysis after multiple imputation

Financial support: This work was supported by the 'Agence Nationale de Recherches sur le Sida et les Hépatites Virales' (ANRS) (grant number NM/DF/1754).

ABSTRACT

Background

Multiple Imputation as usually implemented assumes that data are Missing At Random (MAR), meaning that the underlying missing data mechanism, given the observed data, is independent of the unobserved data. To explore the sensitivity of the inferences to departures from the MAR assumption, we applied the method proposed by Carpenter *et al.* (2007). This approach aims to approximate inferences under a Missing Not At random (MNAR) mechanism by reweighting estimates obtained after multiple imputation where the weights depend on the assumed degree of departure from the MAR assumption.

Methods

The method is illustrated with epidemiological data from a surveillance system of hepatitis C virus (HCV) infection in France during the 2001-2007 period. The subpopulation studied included 4343 HCV infected patients who reported drug use. Risk factors for severe liver disease were assessed. After performing complete-case and multiple imputation analyses, we applied the sensitivity analysis to 3 risk factors of severe liver disease: past excessive alcohol consumption, HIV co-infection and infection with HCV genotype 3.

Results

In these data, the association between severe liver disease and HIV was underestimated, if given the observed data the chance of observing HIV status is high when this is positive. Inference for two other risk factors were robust to plausible local departures from the MAR assumption.

Conclusions

We have demonstrated the practical utility of, and advocate, a pragmatic widely applicable approach to exploring plausible departures from the MAR assumption post multiple imputation. We have developed guidelines for applying this approach to epidemiological studies.

BACKGROUND

Missing data are ubiquitous in epidemiological and clinical research, and in consequence there is increasing interest in appropriate statistical methods, principally multiple imputation (MI) [1, 2]. Multiple imputation techniques available in standard statistical software [3, 4] enable parameter estimation under the assumption that missing data are missing at random (MAR), meaning that the missingness mechanism depends on observed data only, and given these no longer on the missing data [5].

Incomplete datasets are usually addressed by a complete-case (CC) analysis restricted to individuals that have no missing data in any of the variables required for the analysis. For etiologic analyses, a complete-case approach leads to a loss in power, but gives valid results if the probability of being a complete-case is independent of the outcome, given the covariates in the model [5, 6]. However, if the missingness mechanism depends on the outcome, given the covariates, a complete-case analysis will be systematically biased, even under the MAR assumption [7, 8]. Conversely, MI allows individuals with incomplete data to be included in the analysis. It yields valid and efficient inferences under the MAR assumption, even if the missingness mechanism is related to the outcome, provided the imputation model is appropriate [5].

Missing data may also be due to a Missing Not At Random (MNAR) mechanism, also termed non-ignorable, meaning that, given the observed data (including the outcome), the missingness mechanism depends on unobserved data. In practice, it is impossible to distinguish between MAR and MNAR data [9]. When performing multiple imputation under MAR, the estimate of the regression coefficient of a covariate with missing values can be subject to bias when the missingness mechanism of the covariate is MNAR, whether this MNAR mechanism depends on the outcome variable or not [6, 7]. The extend of this bias is often greater the stronger the dependence of the missingness mechanism on the outcome [10]. Sensitivity analysis is useful in such cases.

Specifically, where the missingness mechanism for one or more of the covariates depends on the response in the model of interest, a MI analysis assuming MAR is preferable to a CC analysis, especially if additional variables, not in the model of interest, can be included in the imputation model to increase the plausibility of the MAR assumption [11-14]. Nevertheless, the missingness mechanism may additionally depend on unseen values of a covariate, and the estimates of the coefficient of this covariate may be sensitive to this.

Knowledge of the direction and extent of this sensitivity is important when drawing conclusions from an analysis. The method we present here allows such sensitivity analysis to be performed rapidly after MI under MAR.

In the statistical literature, both selection models [15] and pattern-mixture models [16] have been proposed for the analysis of data under MNAR assumptions [17, 18]. Here, our focus is on selection models, which describe assumptions about the mechanisms causing the missing data and then work through the consequences for inference from the model of interest. Unfortunately, methods for such sensitivity analysis are not implemented in standard statistical software and in their full generality are computationally complex. Thus they are little used in practice [1].

However, a computationally much more straightforward approach to local sensitivity analysis, following MI under MAR, has been proposed by Carpenter *et al.* [19, 20].This 'selection-based' approach explores the robustness of inference under local departures from the MAR assumption, meaning that the sensitivity to departure from the MAR assumption can be calculated from the observed data without estimating a full non-ignorable model [21]. Parameter estimates obtained from the imputed datasets assuming MAR are reweighted to represent the distribution of imputations under a MNAR mechanism. Consequently, inferences obtained under the MAR and MNAR assumptions can be compared to assess the robustness of inferences to local departures from the MAR hypothesis.

This method is attractive as it is easy to implement after performing MI, and it has not been reported for observational data to our knowledge. We have therefore applied this method to surveillance data for hepatitis C viral infection collected in France [22]. As a result of this, we further propose guidelines on the use of the method for observational data.

METHODS

The hepatitis C virus (HCV) surveillance system is based on 26 participating hepatology reference centers out of the 31 located in university hospitals throughout France [23]. Since 2000, it has enrolled patients at first referral with HCV chronic infection to monitor changes in characteristics of HCV infection. A standardized questionnaire is used to collect epidemiological (date of first referral and last HCV negative test, circumstance of HCV antibody testing, and risk factors), clinical, biological (HCV RNA serum status, HIV and HBV co-infection), and history of excessive alcohol consumption data.

Among the 4,343 cases that reported having injected or snorted drugs at least once in their whole life, we assessed risk factors associated with severe liver disease (SLD) at first referral by multivariate logistic regression. The outcome of interest (SLD) was defined as cirrhosis or hepatocellular carcinoma at first referral, as assessed by biochemical tests and morphological evaluation [24].

PRELIMINARY ANALYSES

Details of the study design and the initial analyses have already been described [22, 23]. Six out of the 9 variables retained for the multivariate analysis were incomplete, with a range of missing values from 10 to 26% (Table 1). In the CC analysis, multivariate logistic analysis was reduced to 1,858 individuals (43% of total cases) having no missing data in any of the 9 variables of the analysis. Consequently, we estimated missing values through multiple imputation by chained equations using Stata's user written program *ice* [4] (STATA ® 9.2,

Stata Corporation, College Station, Texas, USA). This computationally convenient method is being increasingly used in epidemiology, and does not require any direct assumption on the joint distribution of the variables [25, 26]. The imputation algorithm is based on a set of univariate imputation models which, in turn, regress one variable on all the other covariates and the outcome [27].

The variables in the imputation model were limited to the 9 variables retained after the univariate analysis. No additional (auxiliary) variables were included because they had either too many missing values or were insufficiently related to the missingness mechanism. A total of 30 imputed datasets were generated. The initial study exploring the risk factors of SLD was performed using a joint analysis of these 30 imputed datasets [22]. Further imputations were performed subsequently for the sensitivity analysis.

SENSITIVITY ANALYSIS METHOD

Consider a variable (covariate or response) *Y* with missing values. We denote by Y_i the value of *Y* for the individual *i*. Let R_i be an observation indicator variable equal to 1 if Y_i is observed and 0 if otherwise. We assume a logistic model relating the probability of observing *Y* to the underlying (but potentially unseen) value of *Y* itself, adjusted for a vector *X* of covariates:

$$\text{logit}\,\Pr(R_i = 1) = \alpha + \beta X_i + \delta Y_i. \qquad (1)$$

Under this parametric form assumption, if $\delta = 0$, given the fully observed data, the mechanism causing the missing data of *Y* does not depend on *Y*, so that the missing data are MAR. On the contrary, if $\delta \neq 0$, the missingness mechanism depends on the potentially missing *Y*, even taking into account the information in the observed data. Thus the data are MNAR.

In practice, the above logistic regression cannot be performed since, by definition, we do not observe Y_i when R_i=0. This implies that a value for δ must be chosen, and its effect on

inferences from the model of interest explored. With the method we investigate, this can be done using weights which are a simple function of δ and the imputed data. We next give an intuitive explanation of the approach.

Suppose M datasets are created by a MI method assuming MAR. For each dataset, we denote by $\hat{\theta}_m$ the estimate of the parameter of interest (e.g. a regression coefficient). Multiple imputation assuming MAR results in several point estimates which, under Rubin's rules, are simply averaged for final inference. Thus, the usual MI estimate of θ is expressed by:

$\hat{\theta}_{MAR} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m$ (see Appendix for its estimated variance).

Carpenter's approach works by replacing this simple average by a weighted average, where estimates arising from imputations that are more likely under MNAR are upweighted relative to the others. Under the logistic model for the missingness mechanism described in (1), Carpenter *et al.* show the weights take a particularly simple form [20].

The model (1) hypothesises that, after adjusting for other observed variables, the chance of observing Y per unit change in Y has log-odds ratio δ. Then the weight, noted $\tilde{w}_m(\delta)$ for imputation *m*, (*m=1,…,M),* is equal to $\exp\left[-\delta\sum_{i\in I_Y} Y_i^m\right]$, where $Y_i^m \ (i \in I_Y)$ is the imputed value of Y for the individual *i* in the dataset *m*, and I_Y is the set of individuals with Y unobserved. The exponential form of the weights comes from the logistic link in equation (1). Normalized weights calculated for each imputed dataset are expressed by $w_m(\delta) = \frac{\tilde{w}_m(\delta)}{\sum_{k=1}^{M}\tilde{w}_k(\delta)}$.

The MNAR estimate of θ is defined by $\hat{\theta}_{MNAR}(\delta) = \sum_{m=1}^{M} w_m(\delta)\cdot\hat{\theta}_m$ (see Appendix for its estimated variance).

Note that if data are MAR, then $\delta = 0$, and all imputations are equally weighted as in Rubin's original rules.

To gain an intuition for these weights, if δ is positive the chance of observing Y is greater for more positive *Y*. Thus in the data after imputation under MAR, imputations with small Y will be under-represented. The weights correct this by up-weighting (relative to the other imputed data sets) estimates from imputed data sets where the sum of the imputed values of *Y* is small.

Below, we present MAR estimates for the HCV dataset and explore their robustness to MNAR as δ moves away from zero. We further propose practical guidelines for selecting a δ value where possible (or at least a plausible range of values for δ).

FRAMEWORK FOR SENSITIVITY ANALYSIS

Among variables retained in the multivariate analysis (Table 1), we focused on the missingness mechanisms of 3 binary variables. We now discuss epidemiological hypotheses about these mechanisms for each variable in turn.

Alcohol consumption: Reporting alcohol consumption may be prone to a social desirability effect, even when past consumption is accounted for. We hypothesized that former heavy drinkers were less likely to report their past alcohol consumption.

HIV infection: HIV serostatus could be assessed either by a previous HIV test, where available, or by a test at first referral. Since the prevalence of HCV-HIV co-infection in hepatology reference centers is ~8% and HIV testing is quite systematic, the physician may consider that patients are mainly HCV mono-infected when no positive HIV test is available. We hypothesized that HIV testing is less often reported when patients are HIV negative.

HCV genotype 3: HCV genotypes are tested by the referral laboratories of the participating centers. Genotyping might depend on the physicians' attitudes, but probably not on the unobserved values of the genotype. We nevertheless explored the MNAR assumption.

Consequently, we focused the sensitivity analysis on these 3 variables to assess the robustness of estimates obtained after MI using Carpenter's method [20]. The sensitivity analysis was applied to each variable separately in turn.

PRACTICAL CONSIDERATIONS FOR SENSITIVITY ANALYSIS

Although it is recommended to impute at least 50 datasets [20], we chose to impute 1000 datasets using the Stata *ice* program to illustrate the features of the method. This is double the number used by Carpenter *et al.* [19] in a simulation study to test the method; although imputations are computationally cheap, beyond 1000 the gain, in terms of increased range of the imputation estimates $\hat{\theta}_m$, is small.

One way to select a value for δ is to formally elicit plausible values from experts [28]. An alternative is to explore a range of values consistent with hypotheses concerning the missing data mechanism, such as those outlined above.

We propose the following 4-step approach for choosing an appropriate value for δ, and illustrate this using the HCV genotype 3 variable, before applying it to the other variables. Our focus is sensitivity analysis for the parameter of interest i.e. the coefficient of genotype 3 in the post MI multivariate logistic regression explaining SLD (Table 1, rightmost column). Using previous notation, $\hat{\theta}_m$ is the MAR estimated logistic regression coefficient of genotype 3 in the imputed dataset *m=1,…,M*.

Procedure for choosing δ

Step 1: Logistic regression to explore the missingness mechanism

Generate an indicator variable for the covariate in question being missing, and use logistic regression to assess association with the outcome and other covariates.

Illustrating with genotype 3, we generate a missing indicator equal to 1 if genotype 3 is observed and 0 otherwise. Using imputed values, we then fit a multivariate logistic regression

model to explain the genotype 3 missing indicator; in this model we include the outcome (SLD) and all the covariates included in the initial analysis model, genotype 3 excepted.

The results, shown in Table 2, suggest the missingness mechanism for genotype 3 depends on age and disease duration, but given these is independent of SLD, the outcome in the analysis model.

Step 2: Graphical determination of a delta value

The theoretical justification of the method rests on importance sampling [29]. When using importance sampling, it is not recommended to put all the weight on one, or very few values. The implication is that we should restrict the range of δ. Consistent with this, we recommend the following criteria: values of δ should be such that the maximum normalized weight is around 0.5, and at least 5 normalized weights are above 1/M (the weight when $\delta = 0$)[20]. Thus our MNAR estimate will draw on information from at least 5 imputations, the minimum typically advised in practice. It also reflects practically relevant, yet appropriately local, departures from MAR.

In practice, we recommend presenting this information in a graph such as Figure 1. The left panel shows a histogram of the sum of the imputed values for genotype 3. Extreme values are 340 in imputed dataset n°921 and 480 in dataset n°771. The right panel indicates normalized weights for each of the M=1000 datasets by δ value. The maximum normalized weight corresponds to the dataset(s) in which the sum of imputed values of *Y* is minimal (dataset n°921) when δ>0 or maximal (dataset n°771) when δ<0. When δ=0, the normalized weight is equal to 1/M because all the $\tilde{w}_m(0)$ are equal to 1.

Figure 2 shows the central part of the right panel of Figure 1. Following our recommendation above, we retain positive or negative δ values that correspond to a maximum normalized weight of ~0.5. This gives a range of [-0.2 to 0.15]. Even at the end of this range, more than 5 normalized weights are > 0.001. The central part of the hatched zone (defined subjectively,

although an objective criteria could be set down *a-priori* if desired) corresponds to departures from MAR for which the weights are still approximately equal, so that MAR and MNAR inferences are essentially the same.

Step 3: Choice of sign of δ

Here, we choose δ to be either the upper or lower end of the range identified in step 2. For HCV genotype, equation (1) shows the relation between the sign of δ and the assumed missingness mechanism: for positive δ, the adjusted odds for observing genotype increases if a person's HCV is of genotype 3; for negative δ the converse. In this instance, consistent with the results from step 1 (Table 2), we selected $\delta = 0.15$. This means that the adjusted odds of missing data for genotype 3 is 1.2 (exp(0.15)) times greater for individuals infected by a genotype 3 strain than for those infected by other genotypes. For this variable, experience does not strongly suggest a positive or negative δ, and results for both are presented below.

Step 4: Graphical diagnostic

The re-weighting method is for local sensitivity analysis, which means that the distributions of the parameter of interest under MAR and MNAR should overlap (albeit they have different means). This will not generally hold for non-local sensitivity analysis. The 'range' of such local sensitivity analyses will depend on the between imputation variance of the estimator, which is indirectly related to the proportion of missing observations. To assess whether, at the chosen value of δ, this holds we propose (i) a plot of normalized weight w_m against $\hat{\theta}_m$, $m=1,\ldots,M$ and (ii) a plot of the estimate under MNAR as the number of imputations increases. In both plots, if the method is to give reliable results, MNAR estimates should be supported within the distribution of $\hat{\theta}_m$ obtained by MI under MAR. If all the weight is

accruing to estimates at the end of the range of $\hat{\theta}_m$, this is consistent with the mean of the MNAR distribution lying outside the range of MAR estimates, i.e. a 'non-local' departure from MAR.

For the HCV genotype variable the results are shown in Figure 3. The left-hand panel plots the normalized weights versus $\hat{\theta}_m$ for each imputed data, using δ=0.15 (recall $\hat{\theta}_m$ is the regression coefficient estimate obtained under the MAR assumption for each imputed dataset). The right panel plots the MNAR estimate calculated using *n* imputations against the number of imputed datasets noted *n* (*n=10,…,M*) and defined by:

$$\hat{\theta}_{MNAR}(\delta, n) = \frac{\sum_{m=1}^{n} w_m(\delta) \cdot \hat{\theta}_m}{\sum_{m=1}^{n} w_m(\delta)}.$$

In this case we see that (i) the MNAR estimate appears to settle down as the number of imputations increases and (ii) the MNAR estimates distribution seems to be well supported by the MAR distribution (indicated by the 'rug' on the right side of the plot).

RESULTS

The complete-case and MI (assuming MAR) analysis are shown in Table 3. Here we also give the results of sensitivity analysis for the following three variables: HCV genotype 3, HIV serostatus and history of excessive alcohol consumption.

For HCV genotype 3, we derived the value of the sensitivity parameter δ above, to illustrate our four step approach. We applied the same approach to HIV serostatus and alcohol consumption.

For alcohol consumption, step 1 showed the probability of observing this depends on the outcome (SLD status). Step 2 identified the range for δ of [-0.4;0.4] (left panel of Figure 4). Consistent with step 1 and experience, the odds of observing alcohol intake is higher if it is not excessive, we chose δ = -0.4. The interpretation is that, after adjustment for other

variables, the odds of observing alcohol history is reduced among those with a history of excessive intake by 0.7=exp(-0.4).

For HIV serostatus, step 1 showed the probability of observing this depends on the outcome (SLD status). Step 2 identified a range of [-0.4;0.7] (right panel of Figure 4). Taking the results from step 1, and given that in similar contexts the chance of observing HIV infection is higher for HIV positive individuals, we chose $\delta = 0.7$. The interpretation is that, after adjustment for other variables, the odds of observing HIV serostatus is 2.0=exp(0.7) times higher if HIV serostatus is positive.

For these three variables, the diagnostic in step 4 was acceptable. Adjusted odds ratios (OR) are shown in Table 3. Note the same multivariate model including sex, age, duration of HCV infection, alcohol consumption, genotype 3, and HIV serostatus, was applied for each analysis. The sensitivity analysis was applied to each of the 3 variables in turn.

Two criteria are useful to interpret the adjusted odds ratios in the 3 analyses:
(1) The coefficient of variation (CV) of the OR gives its normalized measure of dispersion. For the 3 variables, it is clearly reduced after MI and remains stable after reweighting.
(2) The variation rate (VR) assesses the relative change between the $\hat{OR}_{MNAR}$ and the $\hat{OR}_{MAR}$, and is defined by $VR_{SA} = 100 \times (\hat{OR}_{MNAR} - \hat{OR}_{MAR}) / \hat{OR}_{MAR}$. Similarly, we define a variation rate named VR_{MI} that displays the relative variation of the OR obtained after CC and MI analyses. VR_{MI} varies from 9.7% for genotype 3 to 15.5% for HIV and 22% for alcohol. VR_{SA} is given for the value of δ selected for each variable. Its value is relatively small for alcohol (1.3%) and genotype 3 (3.5%) but larger for HIV at 6.6%. The VR_{SA} is relatively stable as δ varies in [-1;1] for alcohol and genotype 3, but continues to increase for HIV (Figure 5).

DISCUSSION

With missing data, all analyses and corresponding inferences rest on inherently untestable assumptions about the missingness mechanism. Therefore, sensitivity analyses, where we explore the robustness of inferences as assumptions change, are important.

The method presented here enables rapid local sensitivity analysis to inferences obtained via MI under MAR. It works by upweighting imputations which are more plausible under MNAR; under a logistic model for the missingness mechanism, these weights take a particularly simple form.

While the sensitivity analysis is local, it nevertheless provides important information on the duration and impact of departures from MAR on inference, while avoiding the computational complexity of full joint modeling. Its accuracy for local sensitivity analysis has been confirmed elsewhere [20, 30].

Here, we have developed and illustrated the practical utility of the approach, proposing a 4-step process for choosing a value for the sensitivity parameter. We now discuss the results. Note that all three variables are binary, so the scale for delta is the same.

For genotype 3, step 1 of our process shows that among individual with complete records on variables apart from genotype 3, the probability of observing this variable does not appear to be related to the outcome in the model of interest (severe liver disease) (Table 2). The sensitivity analysis allows this probability additionally to depend on the underlying value of genotype 3 (present or absent). Table 3 and Figure 5 show inference is insensitive to this, indeed for plausible delta the estimate moves back towards the complete case estimate, consistent with what would be expected if the additional MNAR dependence does not materially change the lack of dependence of the chance of observing genotype 3 on severe liver disease.

Regarding alcohol consumption, our hypothesis was that patients will be less willing to report past excessive alcohol consumption because of the associated social stigma. However,

the literature is not unanimous on this [31-34]. Reporting alcohol consumption is strongly related to the sociodemographic characteristics [35] that can be included in the imputation model in order to reduce non-response bias. We only included age and sex in the imputation model because other sociodemographic variables were not related to the missingness mechanism. Our step 1 showed that the probability of observing alcohol does depend on the outcome, after taking into account other covariates. This is consistent with the relatively large change in the alcohol covariate under MAR (Table 3). The sensitivity analysis allows this probability additionally to depend on the alcohol value. Table 3 and Figure 5 show inference is relatively insensitive to this, which is consistent with what would be expected if the additional MNAR dependence does not materially change the dependence of the chance of observing alcohol consumption on severe liver disease.

For HIV co-infection, hepatologists in reference centers tend to consider their patients as being HCV mono-infected because HCV-HIV co-infected patients are usually referred to infectious diseases departments in France. Under the default assumption of mono-infection, there would thus usually be less impetus to test for or record HIV status. Thus we felt it was more likely to be observed if it was present, i.e. serostatus was positive, hence our positive value for δ. Step 1 showed that the probability of observing HIV status does depend on outcome, after taking into account other covariates. This is consistent with the relatively large change in the coefficient under MAR. The sensitivity analysis allows this probability additionally to depend on the HIV status. Table 3 and Figure 5 show inference is sensitive to this, suggesting that if the mechanism is MNAR with increased chance of observing HIV status for those who are positive, the association in the model of interest is stronger and more significant than analysis under MAR would suggest.

In summary, for these data collected from a French surveillance system for hepatitis C infection, CC analysis is plausibly biased, as the data suggest dependence of the chance of observing values on the outcome, even given the covariates. Thus analysis under MAR, via

MI, is preferable. Our sensitivity analysis shows that for local departures from MAR, inference for the genotype 3 and alcohol consumption is little changed, while the effect of HIV status is underestimated if, given the observed data, the chance of observing HIV status is higher when this is positive.

The approach we have described here can also be applied to explore the situation when there is an interaction between, say, the response (disease status) and the chance of a risk factor being observed. In this case we may have two sensitivity parameters, one for each group, or possibly a single parameter representing the difference between these. Since evidence for this has been found [36] this is a natural area for future work.

CONCLUSION

This sensitivity analysis provides a fast, albeit approximate, way to assess the robustness of inferences to the MAR assumption, avoiding the need for further imputation and model fitting to the imputed datasets. In this paper we have proposed a 4-step process for using this method in practice. We have demonstrated the application of this method and the interpretation of the results. Faced with non-trivial proportions of missing data, we encourage readers to apply the method in their own analyses.

Appendix

Let M the number of imputed datasets, $\hat{\theta}_m$ the estimated parameter of interest in the imputed dataset m, $m=1,\ldots,M$, and $\hat{\sigma}_m^2$ its associated variance estimate. The estimated MAR variance of $\hat{\theta}_{MAR}$ is:

$\hat{V}_{MAR}\left(\hat{\theta}_{MAR}\right)=\hat{V}_W\left(\hat{\theta}_{MAR}\right)+\left(1+\frac{1}{M}\right)\cdot\hat{V}_B\left(\hat{\theta}_{MAR}\right)$, where $\hat{V}_W\left(\hat{\theta}_{MAR}\right)=\frac{1}{M}\sum_{m=1}^{M}\hat{\sigma}_m^2$ and

$\hat{V}_B\left(\hat{\theta}_{MAR}\right)=\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\theta}_m-\hat{\theta}_{MAR}\right)^2$.

The estimated MNAR variance of $\hat{\theta}_{MNAR}$ for a chosen δ value is:

$\hat{V}_{MNAR}\left(\hat{\theta}_{MNAR}(\delta)\right)\approx\hat{V}_W\left(\hat{\theta}_{MNAR}(\delta)\right)+(1+\frac{1}{M})\cdot\hat{V}_B\left(\hat{\theta}_{MNAR}(\delta)\right)$, where $\hat{V}_W\left(\hat{\theta}_{MNAR}(\delta)\right)=\sum_{m=1}^{M}w_m(\delta)\cdot\hat{\sigma}_m^2$

and $\hat{V}_B\left(\hat{\theta}_{MNAR}(\delta)\right)=\sum_{m=1}^{M}w_m(\delta)\cdot\left(\hat{\theta}_m-\hat{\theta}_{MNAR}(\delta)\right)^2$.

Reference List

1. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR: **Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls**. *BMJ* 2009, **338:**b2393.

2. Kenward MG, Carpenter J: **Multiple imputation: current perspectives**. *Stat Methods Med Res* 2007, **16:**199-218.

3. Horton NJ, Kleinman KP: **Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models**. *Am Stat* 2007, **61:**79-90.

4. Royston P: **Multiple imputation of missing values: furter update of ice, with an emphasis on categorical variables.** *Stata J* 2009, **9:**466-477.

5. Little RJA, Rubin DB: *Statistical analysis with missing data.* 2nd ed. New York: Wiley; 2002.

6. Spratt M, Carpenter J, Sterne JA, Carlin JB, Heron J, Henderson J, Tilling K: **Strategies for multiple imputation in longitudinal studies**. *Am J Epidemiol* 2010, **172:**478-487.

7. Allison PD: *Missing data*. Iowa City: Sage Publication; 2002.

8. Vach W, Blettner M: **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**. *Am J Epidemiol* 1991, **134:**895-907.

9. Potthoff RF, Tudor GE, Pieper KS, Hasselblad V: **Can one assess whether missing data are missing at random in medical studies?** *Stat Methods Med Res* 2006, **15:**213-234.

10. White IR, Carlin JB: **Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values**. *Stat Med* 2010, **29:**2920-2931.

11. Collins LM, Schafer JL, Kam CM: **A comparison of inclusive and restrictive strategies in modern missing data procedures**. *Psychol Methods* 2001, **6:**330-351.

12. Raghunathan TE: **What do we do with missing data? Some options for analysis of incomplete data**. *Annu Rev Public Health* 2004, **25:**99-117.

13. Schafer JL: **Multiple imputation in multivariate problems when the imputation and analysis model differ.** *Stat Neer* 2003, **57:**19-35.

14. Wood AM, White IR, Royston P: **How should variable selection be performed with multiply imputed data?** *Stat Med* 2008, **27:**3227-3246.

15. Diggle MG, Kenward MG: **Informative drop-out in longitudinal data analysis.** *Appl Statist* 1994, **43:**49-93.

16. Hogan JW, Laird NM: **Mixture models for the joint distribution of repeated measures and event times**. *Stat Med* 1997, **16:**239-257.

17. Curran D, Molenberghs G, Thijs H, Verbeke G: **Sensitivity analysis for pattern mixture models**. *J Biopharm Stat* 2004, **14:**125-143.

18. Molenberghs G, Verbeke G: *Models for discrete longitudinal data.* New York: Springer; 2005.

19. Carpenter J, Rucker G, Schwarzer G: **Assessing the Sensitivity of Meta-analysis to Selection Bias: A Multiple Imputation Approach**. *Biometrics* 2011, **67:**1066-1072.

20. Carpenter JR, Kenward MG, White IR: **Sensitivity analysis after multiple imputation under missing at random: a weighting approach**. *Stat Methods Med Res* 2007, **16:**259-275.

21. Troxel AB, Ma G, Heitjan DF: **An index of local sensitivity to nonignorability**. *Stat Sin* 2004, **14:**1221-1237.

22. Larsen C, Bousquet V, Delarocque-Astagneau E, Pioche C, Roudot-Thoraval F, Desenclos JC: **Hepatitis C virus genotype 3 and the risk of severe liver disease in a large population of drug users in France**. *J Med Virol* 2010, **82:**1647-1654.

23. Delarocque-Astagneau E, Roudot-Thoraval F, Campese C, Desenclos JC, The Hepatitis CSSS: **Past excessive alcohol consumption: a major determinant of severe liver disease among newly referred hepatitis C virus infected patients in hepatology reference centers, France, 2001**. *Ann Epidemiol* 2005, **15:**551-557.

24. Schuppan D, Afdhal NH: **Liver cirrhosis**. *Lancet* 2008, **371:**838-851.

25. Lee KJ, Carlin JB: **Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation**. *Am J Epidemiol* 2010, **171:**624-632.

26. Van Buuren S., Boshuizen HC, Knook DL: **Multiple imputation of missing blood pressure covariates in survival analysis**. *Stat Med* 1999, **18:**681-694.

27. Van Buuren S: **Multiple imputation of discrete and continuous data by fully conditional specification**. *Stat Methods Med Res* 2007, **16:**219-242.

28. White IR, Carpenter J, Evans S, Schroter S: **Eliciting and using expert opinions about dropout bias in randomized controlled trials**. *Clin Trials* 2007, **4:**125-139.

29. Rubin DB: **The calculation of posterior distributions by data augmentation (in discussion of Tanner MA, Wong WH).** *J Am Stat Assoc* 1987, **82:**543-546.

30. Carpenter JR, Kenward MG: **Missing data in randomised clinical trials-a practical guide. Methodology Portfolio, Public Health, Epidemiology & Biostatistics, University of Birmingham, Edgbaston, Birmingham, UK.**; 2007.

31. Pernanen K: *Validity of survey data on alcohol use.* New York: Wiley; 1974.

32. Knibbe R: **Measuring drinking context**. *Alcohol Clin Exp Res* 1998, **22:**15S-20S.

33. Lemmens PH, Tan ES, Knibbe RA: **Bias due to non-response in a Dutch survey on alcohol consumption**. *Br J Addict* 1988, **83:**1069-1077.

34. Lahaut VM, Jansen HA, van de MD, Garretsen HF: **Non-response bias in a sample survey on alcohol consumption**. *Alcohol Alcohol* 2002, **37:**256-260.

35. Van Oers JA, Bongers IM, van de Goor LA, Garretsen HF: **Alcohol consumption, alcohol-related problems, problem drinking, and socioeconomic status**. *Alcohol Alcohol* 1999, **34:**78-88.

36. Lewden C, Jacqmin-Gadda H, Vilde JL, Bricaire F, Waldner-Combernoux A, May T, Cuzin L, Lang JM, Leport C, Chene G: **An example of nonrandom missing data for hepatitis C virus status in a prognostic study among HIV-infected patients**. *HIV Clin Trials* 2004, **5:**224-231.

Table 1: Multivariate logistic regression of factors associated with severe liver disease.

Complete case and multiple imputation analyses were applied to a population of HCV-RNA positive drug users newly referred in hepatology reference centres in France, 2001-2007.

				Multivariate analysis	
	Patients (n = 4 343)	% SLD	% missing data	Complete Case (n = 2 130) aOR* (95% CI*)	Multiple Imputation (n = 4 343) aOR* (95% CI*) *M*=30 imputed datasets
Factors					
Period of inclusion					
2001-2003	2330	7.0			
2004-2007	2013	9.5			
Sex					
Female	993	4.2		1.0	1.0
Male	3350	9.3		1.8 [1.1,3.0]	2.0 [1.4,2.9]
Age					
≤ 40 years	2435	3.9		1.0	1.0
> 40 years	1908	13.6		2.2 [1.5,3.3]	2.3 [1.7,3.1]
Time between 1st HCV+ test and referral					
< 1 year	1728	6.7			
≥ 1 year	2163	8.7			
Missing	452	11.5	10.4		
Duration of HCV infection at referral†					
< 18 years	1709	3.0		1.0	1.0
≥ 18 years	2002	12.5		3.1 [2.0,5.1]	2.6 [1.8,3.7]
Missing	632	8.2	14.6		
History of excessive alcohol intake‡					
No	2015	4.5		1.0	1.0
Yes	1847	13.2		2.6 [1.8,3.7]	2.8 [2.2,3.7]
Missing	481	4.4	11.1		
HbsAg status					
Negative	3570	8.3		1.0	
Positive	89	13.5		2.4 [1.0,5.9]	
Missing	684	6.7	15.7		
HIV serostatus					
Negative	3342	8.2			1.0
Positive	294	14.0			1.8 [1.2,2.6]
Missing	707	5.7	16.3		
HCV genotype 3					
No	2083	7.2		1.0	1.0
Yes	1117	10.3		1.5 [1.1,2.0]	1.6 [1.3,2.1]
Missing	1143	7.8	26.3		

*aOR, adjusted Odds Ratio; CI, confidence interval;
SLD, severe liver disease (cirrhosis, hepatocellular carcinoma); HCV, hepatitis C virus; HBsAg, hepatitis B surface antigen; HIV, human immunodeficiency virus
† Time from suspected year of infection to year of referral to the reference centre. Suspected year of HCV infection is defined as year of the last HCV negative test performed during the drug-use period or year of first drug injection
‡ >210g/week for women and >280g/week for men

Table 2: Multivariate regression to explain the missing indicator of genotype 3 using covariates.

Genotype 3 missing indicator	Regression coefficients	SE*	*P**
Severe liver disease †	-0.05	0.13	0.72
Age	0.16	0.09	0.06
Sex	0.04	0.08	0.63
Disease duration ‡	0.19	0.08	0.04
Delay of referral [4]	0.05	0.08	0.53
Alcohol consumption [5]	-0.005	0.07	0.94
HIV serostatus	0.02	0.14	0.90
HbsAg status	-0.14	0.23	0.52

* *P*, pvalue; SE, standard error.
† Cirrhosis or hepatocellular carcinoma
‡ Time from suspected year of infection to year of referral to the reference centre. Suspected year of HCV infection is defined as year of the last HCV negative test performed during the drug-use period or year of first drug injection
[4] Time between testing and first referral
[5] >210g/week for women and >280g/week for men

Table 3: Multivariate analysis for the complete case, multiple imputation and sensitivity analysis, with *M*=1000 imputed data sets.
Covariates included in the model were: sex, age, duration of HCV infection, alcohol consumption, genotype 3 and HIV serostatus. Sensitivity analysis: the weighting process was applied to alcohol consumption, genotype 3 and HIV serostatus indepedently.

		Complete Case (CC)			Multiple Imputation (MI)					Sensitivity Analysis (SA)			
	Missing values %	aOR 95% CI	SE	CV (%)	aOR 95% CI	SE	CV (%)	VR_{MI} (MI *vs* CC) (%)	δ	aOR* 95% CI	SE	CV (%)	VR_{SA} (SA *vs* MI) (%)
Alcohol consumption	11.1	2.32 [1.66,3.23]	0.39	17	2.82 [2.18,3.66]	0.37	13	21.86	-0.40	2.86 [2.21,3.70]	0.37	13	1.29
Genotype3	26.3	1.51 [1.10,2.07]	0.24	16	1.66 [1.27,2.16]	0.23	14	9.70	0.15	1.60 [1.23,2.06]	0.21	13	3.56
HIV serostatus	16.3	1.56 [0.92,2.62]	0.41	27	1.80 [1.24,2.61]	0.34	19	15.52	0.70	1.91 [1.32,2.76]	0.36	19	6.12

Note. aOR, odds ratio adjusted on sex, age, duration of HCV infection, alcohol consumption and HIV serostatus; aOR*, odds ratio obtained from the MI adjusted odds ratio estimates
CI, confidence interval; CV, coefficient of variation of the aOR; VR_{MI}, variation rate of the aOR for CC and MI analyses; VR_{SA}, variation rate of the aOR for MI and senstitivity analyses.

Figure 1:

Graphical determination of a delta value for the variable genotype 3.
Left panel: histogram of the sum of genotype 3 imputed values for each data set and for M = 1000 bases ; extreme values of this sum are 340 in imputed dataset n°921 and 480 in imputed dataset n°771.
Right panel: normalized weights (w_m) for each imputed dataset according to δ.

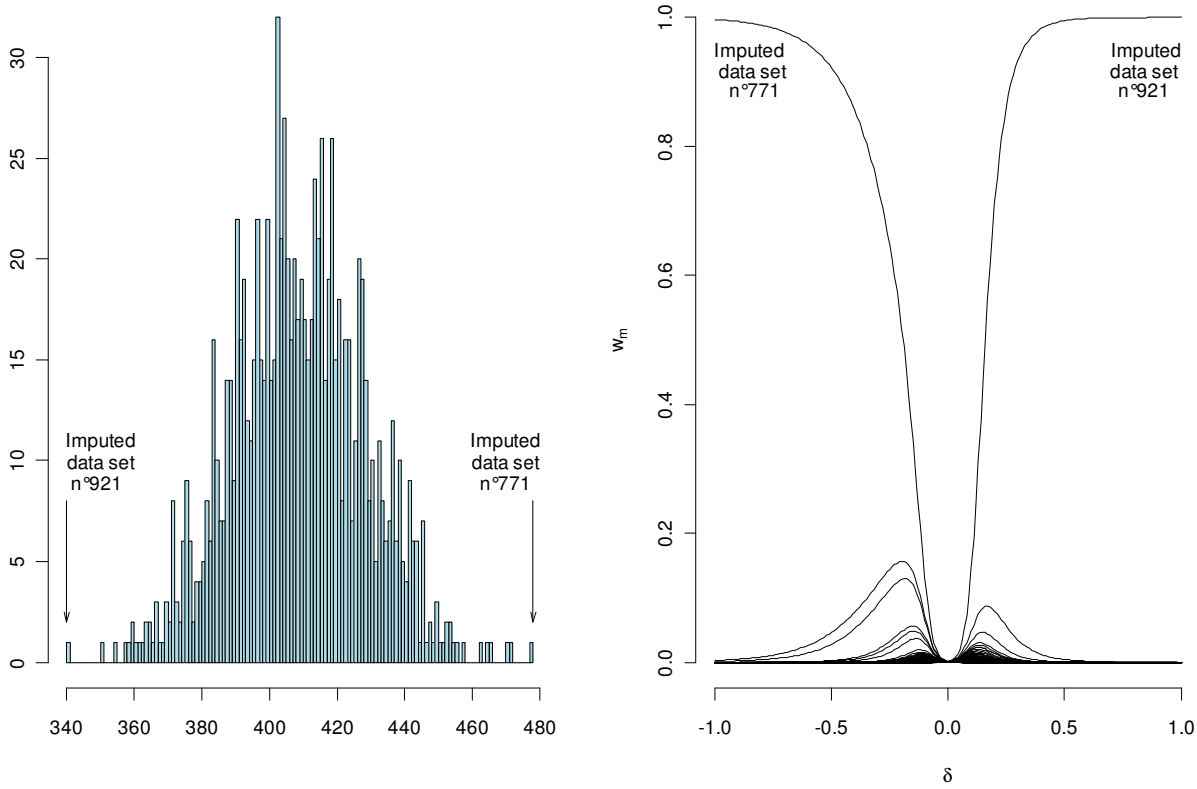


Figure 2:

Normalized weights (w_m) for each imputed dataset according to δ for the variable genotype 3.
The hatched zone delineates values of δ corresponding to maximum normalized weights equal to 0.5.

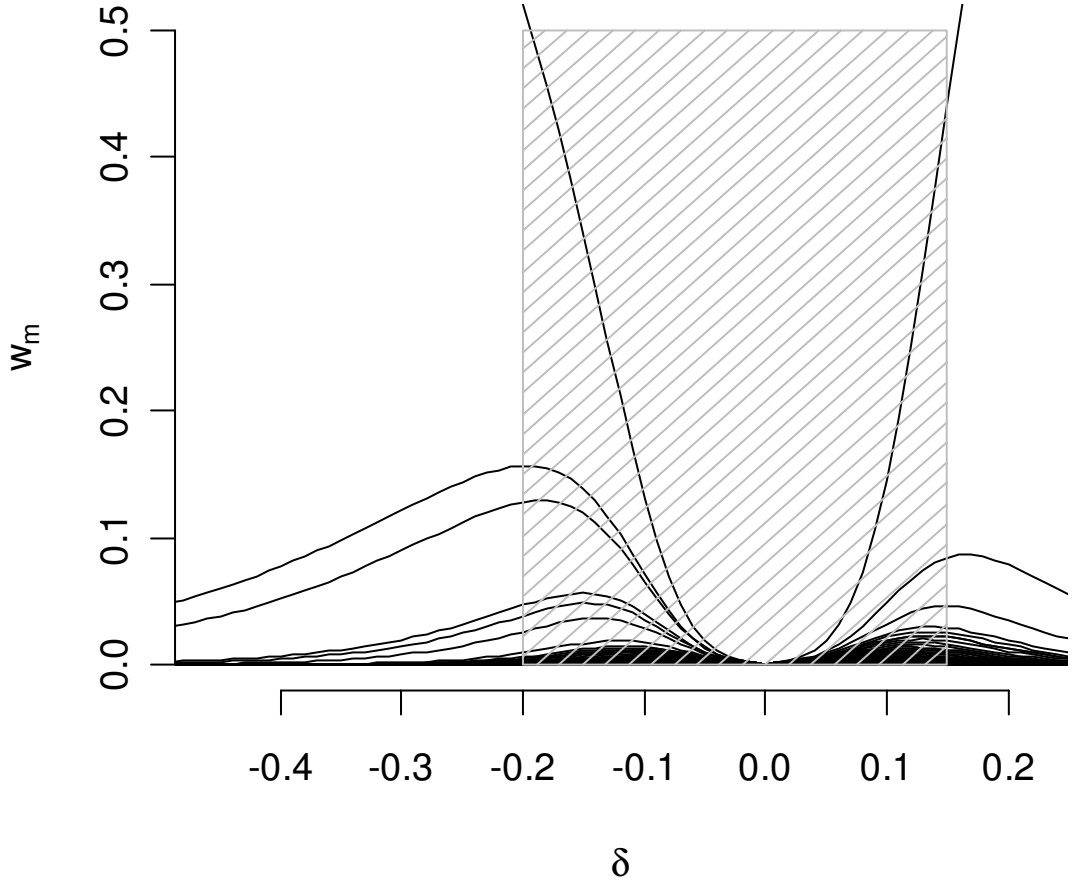


Figure 3:

Analysis of the variable genotype 3 with δ = 0.15.

Left panel: normalized weights (w_m) versus $\hat{\theta}_m$ (estimated logistic regression coefficient of genotype 3 in the imputed dataset *m*), for each imputed data set. The dash line represents $\hat{\theta}_{MAR}$ (mean of $\hat{\theta}_m$ over the 1000 imputed datasets).

Right panel: running estimate, calculated as the moving average of the $\hat{\theta}_{MNAR}$ according to the number of imputed datasets. On the right axis is plotted the 'rug' of the 1000 estimates $\hat{\theta}_m$ for each imputed dataset. The dash line represents $\hat{\theta}_{MAR}$ (mean of $\hat{\theta}_m$ over the 1000 imputed datasets).

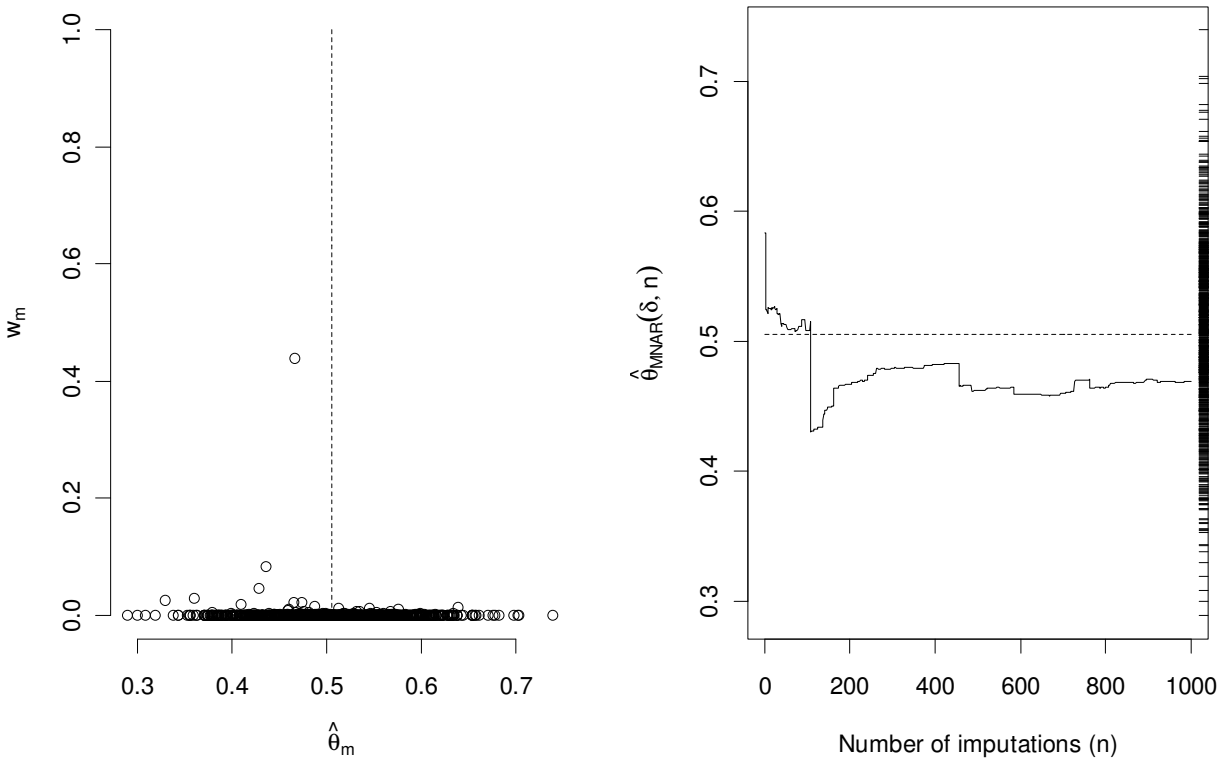


Figure 4:

Normalized weights (w_m) for each imputed dataset according to δ for the variables alcohol consumption and HIV serostatus.
Left panel: the interval for δ is restrained to [-0.4;0.4].
Right panel: the interval for δ is [-0.4;0.7].

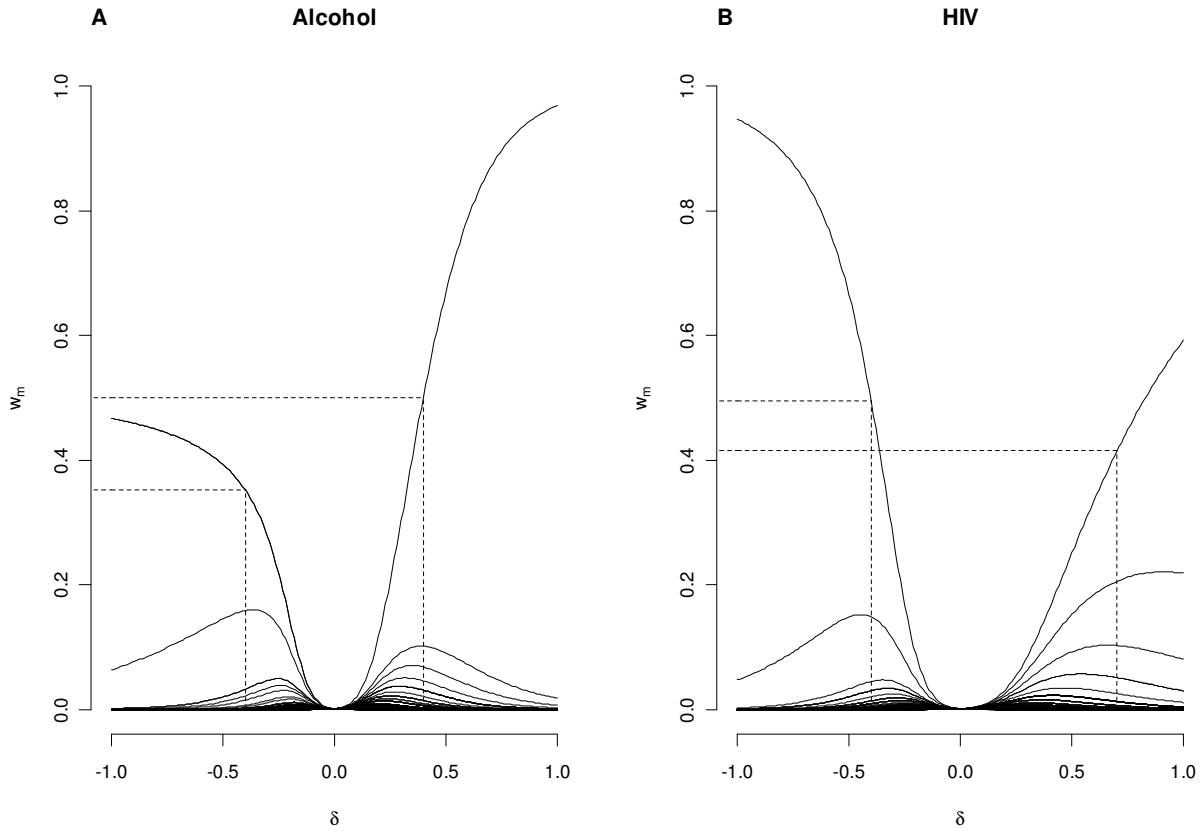


Figure 5:

Variation rate according to δ after sensitivity analysis (VR_{SA}) for genotype 3, alcohol consumption and HIV serostatus.

The black points correspond to the VR_{SA} calculated for the value of δ retained for each variable (genotype 3 δ=0.15, alcohol δ=-0.4 and HIV δ=0.7).