Approches statistiques causales en grande dimension pour la détection de signaux en pharmacovigilance : application aux notifications spontanées et aux données médico-administratives

Etienne Volatier

université PARIS-SACLAY

Approches statistiques causales en grande dimension pour la détection de signaux en pharmacovigilance : application aux notifications spontanées et aux données médico-administratives

High-dimensional causal statistical approaches for signal detection in pharmacovigilance: application to spontaneous reporting and medico-administrative data

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°570 Santé Publique (EDSP)
Spécialité de doctorat : Biostatistiques et data sciences
Graduate School : Santé Publique
Référent : Université de Versailles Saint-Quentin-en-Yvelines

Thèse préparée dans l'unité de recherche **CESP (Université Paris-Saclay, UVSQ, Inserm)**, sous la direction de **Pascale TUBERT-BITTER**, directrice de recherche, du co-encadrement d'**Ismaïl AHMED**, chargé de recherche

Thèse soutenue à Villejuif, le 24 janvier 2024, par

Etienne VOLATIER

Composition du Jury

Membres du jury avec voix délibérative

Philippe BROËT PU-PH, Université Paris-Saclay	Président
Antoine CHAMBAZ PU, Université Paris Cité	Rapporteur & Examinateur
Anne-Sophie JANNOT MCU-PH, Université Paris Cité, HDR	Rapporteure & Examinatrice
Mounia HOCINE MCU, Conservatoire national des arts et métiers, HDR	Examinatrice

THESE DE DOCTORAT

NNT : 2024UPASR001

Titre : Approches statistiques causales en grande dimension pour la détection de signaux en pharmacovigilance : application aux notifications spontanées et aux données médico-administratives

Mots clés : régression pénalisée, détection de signal, pharmacovigilance, Inférence causale, apprentissage statistique, données de grande dimension

La détection de signal est une première étape d'analyse de données exploratoire fondamentale pour la pharmacovigilance, le but final étant de suivre et d'évaluer la sécurité du médicament une fois sa mise sur le marché. Les bases de données utilisées aujourd'hui pour la détection de signal sont essentiellement les bases de notifications spontanées mais récemment l'utilisation des bases médico-administratives a été proposée. Devant le nombre croissant de notifications spontanées ainsi que la grande taille des bases médico-administrative en termes de nombre d'individus et de nombre de variables mesurées, la mise en place d'approches statistiques adaptées à la grande dimension est devenue essentielle.

La présence de facteurs de confusion souvent non mesurés, la poly-exposition et la grande dimension des expositions rendent néanmoins l'analyse complexe. Pour faire face à ces difficultés, nous proposons plusieurs méthodes issues de l'inférence causale en grande dimension pour la détection de signal appliquée sur les bases de notifications spontanées et les bases médico-administratives. Cette approche causale permet de tirer parti de modèles non paramétriques tels que les arbres de régression boostés pour l'ajustement aux données. Nous offrons ainsi des solutions alternatives aux méthodes linéaires généralisées avec pénalité LASSO couramment utilisées pour l'analyse de données en grande dimension. Nous intégrons à nos propositions des outils comme le score de propension en grande dimension, les études autocontrôlées ainsi que la g-computation et l'apprentissage ciblé.

Nos résultats obtenus à la fois par simulations et sur données réelles suggèrent que les approches proposées offrent des alternatives compétitives aux approches basées sur le LASSO. Par ailleurs, deux applications pour l'infarctus du myocarde et les lésions hépatiques aiguës ont pu mettre en évidence plusieurs signaux pertinents.

Title : High-dimensional causal statistical approaches for signal detection in pharmacovigilance: application to spontaneous reporting and medico-administrative data

Keywords : penalized regression, signal detection, pharmacovigilance, causality, statistical learning, high dimension

Signal detection is a fundamental first step in exploratory data analysis for pharmacovigilance, the ultimate aim being to monitor and evaluate drug safety once a drug has been marketed. Currently, the databases used for signal detection are mainly spontaneous reporting databases, but recently the use of medico-administrative databases has been proposed. In view of the growing number of spontaneous reports, and the large size of medico-administrative databases in terms of number of individuals and number of variables measured, it has become essential to implement statistical approaches adapted to large-scale analyses.

However, the presence of often unmeasured confounding factors, poly-exposure and the large number of exposures make analysis complex. To address these difficulties, we propose several methods based on high-dimensional causal inference for signal detection applied to spontaneous reporting and medico-administrative databases. This causal approach allows us to take advantage of nonparametric models such as boosted regression trees for data fitting. We thus offer alternative solutions to the generalized linear methods with LASSO penalty commonly used for high-dimensional data analysis. We integrate tools such as the high-dimensional propensity score, self-controlled studies, g-computation and targeted learning into our approaches.

Our results obtained both in simulations and on real data suggest that the proposed approaches offer competitive alternatives to LASSO-based approaches. In addition, two applications to myocardial infarction and acute liver injury highlighted several relevant signals.

REMERCIEMENTS

En premier lieu, j'adresse mes remerciements les plus forts à mes directeurs de thèse Pascale Tubert-Bitter et Ismaïl Ahmed. Ils ont, grâce à leur curiosité et leur rigueur scientifique, supervisé mes travaux avec pertinence, patience et bienveillance. Je ne commenterai pas longuement leurs qualités humaines, car tout le monde sait qu'elles sont exceptionnelles. Ils m'ont également permis de découvrir de nouvelles méthodes statistiques, ce qui m'a permis de progresser et d'élargir ma palette de connaissances tout au long de cette thèse.

Je tiens également à remercier le professeur Antoine Chambaz et Anne-Sophie Jannot d'avoir accepté d'être rapporteur de ce travail et de m'avoir transmis des remarques très précieuses et justes. Je remercie également Mounia Hocine et Philippe Broët d'avoir accepté de faire partie de mon jury, votre expertise va certainement permettre de construire des discussions intéressantes.

C'est lors d'un séminaire organisé par la Société Française de Statistiques que j'ai pu appréhender les approches causales grâce à des présentations très claires. Je tiens donc à remercier les organisateurs, conférenciers et participants de ces « Journées d'Etudes en Statistique, 2018 ».

Je remercie également mes collègues pour leur gentillesse et leur humour que ce soient les « anciens » : Anne, Emeline, Ghislaine, Hervé, Hong, Liliane, Lucas, Matthieu, Romain, Sidy, Sylvie,

Stéphanie … ou les « nouveaux » : Ana, Elise, Juliette, Lucas, Sidonie …

Ma famille est un roc sur lequel je peux compter à tout moment : mes parents Sophie et Jean-Luc, ma fratrie JB, Mylène la petite Lucie, Pierre et Maud, Claire et Jérémy m'apportent joie et bonheur tous les jours.

Mes amis m'ont soutenu tout au long de ce travail en m'apportant de la bonne humeur et des discussions légères ou profondes selon les moments. Je remercie particulièrement par ordre alphabétique : Camille, Damia, David, Emma, Flo, Gaëtan, Laurent, Najwa, Sébastien, Stéphane, Tessa, Thibaut et Timothé. La longévité de ce groupe d'amis qui date des années lycées pour la majorité est une de mes plus grandes fiertés et une très belle réalisation.

VALORISATION SCIENTIFIQUE

PUBLICATIONS

Publiée dans une revue internationale avec comité de relecture

E. Volatier, F. Salvo, A. Pariente, É. Courtois, S. Escolano, P. Tubert-Bitter*, I. Ahmed*, High-Dimensional Propensity Score-Adjusted Case-Crossover for Discovering Adverse Drug Reactions from Computerized Administrative Healthcare Databases. Drug Saf (2022), doi:10.1007/s40264-022-01148-5

Publication en lien avec la thèse publiée dans une revue internationale avec comité de relecture

É. Courtois, A. Pariente, F. Salvo, **E. Volatier**, P. Tubert-Bitter*, I. Ahmed*, Propensity Score-Based Approaches in High Dimension for Pharmacovigilance Signal Detection: an Empirical Comparison on the French Spontaneous Reporting Database. Front Pharmacol 9, 1010 (2018).

COMMUNICATION ORALE EN CONGRÈS INTERNATIONAL

E. Volatier, E. Courtois, S. Escolano, P. Tubert-Bitter*, I. Ahmed*, Combining case-crossover designs and propensity score approaches for the detection of adverse drug reactions, 29th International Biometric Conference (IBC 2018), Barcelona, Spain, 8-13 July 2018

TABLE DES MATIERES

1 Introduction

1.1 DETECTION DE SIGNAL EN PHARMACOVIGILANCE ET DONNEES DISPONIBLES

1.1.1 Objectifs de la détection de signal en pharmacovigilance

L'identification, le plus rapidement possible, des potentiels effets indésirables des médicaments constitue un champ de recherche majeur pour la santé publique. Les essais cliniques permettent une première identification des effets indésirables du médicament étudié. Néanmoins, ils sont généralement conduits dans des populations de patients homogènes, suivis sur des durées limitées, et de taille insuffisante pour observer des effets rares ou avec de longs délais de survenue. Pour pouvoir identifier des sous-populations à faible effectif et à risque n'ayant pu être détectées lors des essais cliniques, ou des risques liés à un changement dans l'utilisation du médicament, des approches de pharmacovigilance en population réellement traitée, à plus large échelle sont nécessaires. C'est l'objet de la phase IV du développement du médicament.

La pharmacovigilance a donc pour objectif la détection d'éventuels effets indésirables des médicaments, non détectés par les essais cliniques, une fois leur mise sur le marché. La pharmacovigilance contribue ainsi à une meilleure connaissance du risque pour prévenir d'éventuels effets indésirables et à terme améliorer l'offre de soin.

Un signal en pharmacovigilance consiste en une identification

d'un couple médicament-événement indésirable suspect qui nécessiterait une étude plus approfondie. Nous abordons cette problématique du point de vue statistique, pour lequel il s'agit donc d'identifier, dans des bases de données, comprenant beaucoup d'expositions médicamenteuses et d'événements indésirables, des couples médicament-événement indésirable surreprésentés. La détection « automatisée » de signal en pharmacovigilance peut donc être considérée comme une première approche d'analyse exploratoire. L'objectif final de la pharmacovigilance est d'évaluer la causalité de l'association médicament-événement indésirable et nécessite ainsi d'autres facteurs de preuves (Hill 1965) comme l'analyse de mécanismes biochimiques et pharmacologiques que la détection de signal ne peut fournir.

Actuellement, la détection de signal en pharmacovigilance est basée essentiellement sur les systèmes de notifications spontanées constitués des observations rapportées notamment par des professionnels de santé. Une notification concerne un patient donné pour lequel le notificateur suspecte le rôle d'un ou plusieurs médicaments dans la survenue d'un (ou plusieurs) événements indésirables. Il s'agit d'une suspicion et pas du tout d'un lien avéré.

Plus récemment l'utilisation de bases médico-administratives a été mise en avant afin de pallier les biais liés à la sous-notification et à l'absence de vrais témoins dans les bases de notifications spontanées, celles-ci n'étant constituées que de cas. Ces bases médico-

administratives permettent d'observer l'utilisation de médicaments ainsi que la survenue éventuelle d'événements indésirables. Ces bases couvrent quasiment l'ensemble de la population, à l'échelle d'une organisation de santé, d'une région ou d'un pays, utilisant le médicament et permettent la constitution de populations de témoins.

1.1.2 Utilisation des bases de notifications spontanées

La collecte des notifications spontanées s'effectue à l'échelle mondiale par l'OMS (World Health Association (WHO)), qui dispose d'une base de données de notifications spontanés nommée Vigibase (Lindquist 2008) qui comprends plus de 30 millions de notifications recueillies depuis 1968 (Uppsala Monitoring Centre 2023). A l'échelle européenne, l'agence européenne des médicaments (European Medicines Agency (EMA) dispose elle-aussi d'une base de données. Cette base, Eudravigilance, agrège les données de vigilance des pays européens (EMA 2018) et comprend plus de 12 millions de notifications. Enfin, à l'échelle nationale, en France l'Agence nationale de sécurité du médicament et des produits de santé (ANSM) coordonne la Base Nationale de PharmacoVigilance (BNPV) qui centralise plus de 900 000 notifications spontanées recueillies par le réseau des centres régionaux de pharmacovigilance (« La surveillance renforcée des médicaments » 2023). Aux Etats-Unis un système équivalent, FAERS (FDA Adverse Event Reporting System), est mis en place par la FDA (Food and Drug Administration) et comporte plus de 24 millions de notifications.

Une notification peut présenter plusieurs médicaments suspectés ainsi que plusieurs effets indésirables. Des informations sur les caractéristiques du patient, les dosages, le délai de survenue, l'issue, peuvent potentiellement être fournies. Une part importante des effets indésirables potentiels ne sont pas notifiés : on considère que même pour des événements indésirables graves qui sont les mieux notifiés, moins de 10% remontent dans les systèmes (Hazell et Shakir 2006). La notification dépend aussi de la capacité à faire le diagnostic d'un lien possible, ce qui est plus difficile lorsque l'événement survient à distance de l'exposition. Surtout, elle varie au cours du temps, du cycle de vie du médicament (meilleure notification pour les médicaments récents), de ce qu'on connaît déjà de ses effets indésirables, et aussi de phénomènes de crise qui peuvent survenir amenant des biais de notoriété. Enfin, ces bases sont anonymisées et la jointure avec d'autres bases de données comme les bases médico-administratives ou des cohortes n'est pas possible.

Face à ces très grands ensembles de données, plusieurs systèmes de pharmacovigilance se sont dotés de méthodes statistiques de génération de signaux dont l'objectif est d'identifier des couples médicaments-événements indésirables dont la présence est anormalement élevée par rapport à ce qui serait attendu compte tenu de l'information présente dans la base. Ces couples suspects sont appelés signaux du fait des limites inhérentes au recueil des notifications spontanées et une évaluation par des experts pharmacologues est nécessaire afin de conclure à la plausibilité de ces

derniers. Les méthodes actuellement utilisées par les principaux systèmes de pharmacovigilance reposent sur l'analyse des notifications spontanées résumées sous la forme d'une très grande table de contingence croisant l'ensemble des médicaments et des événements indésirables. Ces méthodes reposent sur des mesures de disproportionnalité comparant les nombres observés de notifications pour un couple médicament – événement indésirable donné au nombre attendu qui aurait été observé en cas d'indépendance entre l'ensemble des médicaments et des événements indésirables. Les méthodes de disproportionnalité diffèrent par le modèle de probabilité utilisé, la mesure d'association utilisée pour ranger les signaux ainsi que le seuil de détection utilisé. Les méthodes de disproportionnalités existantes ont été développés dans un cadre fréquentiste ou dans un cadre bayésien. Les principales méthodes fréquentistes sont le Reporting Odds Ratio (ROR) (van Puijenbroek et al. 2002) et le Proportional Reporting Ratio (PRR) (Evans, Waller, et Davis 2001). Elles consistent à calculer respectivement un odds ratio ou un risque relatif pour chaque couple médicament – événement indésirable. Le PRR est implémenté par l'agence européenne du médicament pour la base Eudravigilance. Les méthodes bayésiennes incluent le Multi-item Gamma Poisson Shrinker utilisé par le FAERS (DuMouchel 1999) et l'Information Component (aussi appelée Bayesian Component Propagation Neural Network) utilisé sur les données de Vigibase (Bate et al. 1998). Comme son nom l'indique, le Multi-item Gamma Poisson Shrinker repose sur un modèle de type Gamma-Poisson. L'Information Component repose quant à lui sur un

modèle beta-binomial.

L'évolution récente de la recherche méthodologique en détection de signal automatisée tend à exploiter la très grande taille des données, avec la prise en compte de leur structure de correlation en développant des approches multivariées. Elles ciblent un événement indésirable donné par l'ajustement d'un modèle typiquement logistique intégrant beaucoup de variables « candidates » d'exposition. Pour ce faire, des régressions pénalisées adaptées à la grande dimension type LASSO (Tibshirani 1996) sont à l'œuvre (Caster, O. et al. 2010; Ahmed, Pariente, et Tubert-Bitter 2018).

1.1.3 Intérêt récent pour l'utilisation des bases de données médico-administratives

Plus récemment, l'attention des chercheurs s'est portée sur l'utilisation des bases de données médico-administratives. En France, le Système National des Données de Santé (SNDS) contient des informations précises sur les remboursements des médicaments et les hospitalisations pour la quasi-totalité de la population française. Son utilisation connaît un essor important en pharmacoépidémiologie, c'est-à-dire pour mener des études épidémiologiques en vie réelle, ciblant une association suspecte à évaluer (Luu et al. 2019). Elle est encore peu envisagée pour des objectifs de détection de signal même si des travaux pionniers ont déjà été réalisés, certains avec l'ambition de mettre en œuvre des modèles longitudinaux complexes (Ryan et al.

2013; Kulldorff et al. 2013; Arnaud et al. 2018; Morel et al. 2017; Demailly et al. 2020; Sabatier et al. 2022).

L'avantage de ces bases est qu'il n'y a pas de sous-déclaration potentielle des effets indésirables en comparaison aux bases de notifications spontanées (Moride et al. 1997). Par ailleurs ces bases permettent d'avoir une vue objective des expositions et événements de santé car elles ne reposent pas sur des déclarations mais sur des données de remboursement et des séjours hospitaliers. En revanche, de nombreuses difficultés pour exploiter ces bases subsistent.

Tout d'abord, ces bases médico-administratives ont été conçues comme des bases comptables afin de permettre l'analyse de l'activité médicale. Des difficultés de codage des diagnostics peuvent exister. Elles manquent d'information sur le statut général de la santé des patients (Wells et al. 2013) et d'importants facteurs de confusion comme le tabagisme, l'obésité, l'activité physique…. Dans ce cas une exposition protectrice peut se trouver statistiquement associée à un effet indésirable. De manière générale, comme les populations traitées sont plus fragiles que les populations non traitées elles sont également plus susceptibles d'avoir un effet indésirable. Il est difficile de démêler le rôle du médicament de celui de l'état de santé, un défi classique en pharmacoépidémiologie.

Néanmoins, ces bases sont très riches en information et récemment il a été proposé d'utiliser les nombreuses variables présentes dans ces bases pour constituer un proxy de l'état de santé

globale de l'individu (Sebastian Schneeweiss et al. 2009). Il est donc nécessaire d'adapter la méthodologie en prenant compte à la fois de la richesse des informations contenues dans la base mais également de la nature incomplète des données disponibles.

Un travail important a été effectué précédemment dans le cadre de l'OMOP (Ryan et al. 2013) puis plus tard de l'OHDSI afin de développer l'usage des bases médico-administratives pour la détection de signal en pharmacovigilance. Des algorithmes basés sur les méthodes autocontrôlées (Simpson et al. 2013), un design épidémiologique intéressant pour prendre en compte les facteurs de confusion, ont été développés pour traiter la grande dimension des expositions. L'inclusion d'une pénalisation au modèle de régression logistique conditionnelle associée aux méthodes autocontrôlées est possible et permet d'assurer la convergence de la procédure d'estimation en grande dimension.

1.1.4 Détection de signal et inférence causale

Les démarches d'inférence causale se concentrent principalement sur l'estimation d'un effet traitement donné. Cet effet traitement correspond à une question scientifique d'intérêt (par exemple l'effet d'un médicament sur une maladie). L'objectif de la détection de signal n'est pas d'aboutir à une interprétation causale mais de mettre en emphase certaines relations médicaments événéments indésirables nécessitant une investigation plus approfondie par la suite. Néanmoins des travaux récents ont abordé la

question de l'apport de l'utilisation de méthodes issues du champ de l'inférence causale pour le cadre de la détection de signal, que ce soit à partir de notifications spontanées (Tatonetti et al. 2012; Courtois et al. 2018) ou à partir de données médico-administratives (Demailly et al. 2020). Les méthodes étudiées dans ces travaux reposent sur la construction de scores de propension ou de scores prognostiques en grande dimension, des scores qui ont été développés dans le domaine de la causalité (Sebastian Schneeweiss et al. 2009; Hansen 2008; Kumamaru et al. 2016).

1.2 OBJECTIFS DE THESE

1.2.1 Construction et évaluation de nouvelles approches de détection de signal basées sur des outils issus d'approches statistiques de la causalité

Les approches d'inférence causale ont été initialement proposées pour l'estimation d'effet pouvant faire l'objet d'une interprétation causale. Pour permettre cette interprétation, certaines hypothèses sont effectuées. Avec les bases de données utilisées dans le cadre de ce travail, aucune de ces hypothèses ne peut être justifiée. Néanmoins, des outils statistiques employés en inférence causale méritent d'être testés dans le cadre de la détection de signal qui constitue une analyse exploratoire mais qui implique potentiellement des approches algorithmiques pour permettre une détection de

signaux automatisée. En particulier, le développement récent de méthodologies issues de l'apprentissage statistique s'associe bien avec les approches d'inférence causale. Dans ce cadre, plusieurs travaux méthodologiques ont déjà été proposés comme l'utilisation du score de propension en grande dimension (S. Schneeweiss et al. 2009), de la g-computation (Robins 1986) et de la méthodologie de l'apprentissage ciblé (Laan et Rose 2011; Schuler et Rose 2017). En revanche, ces estimateurs sont essentiellement utilisés pour évaluer un ou peu de couples médicament-événements indésirables.

Un objectif de cette thèse, est de proposer et d'évaluer l'utilisation de plusieurs de ces outils dans un cadre exploratoire pour enrichir l'ensemble des méthodes automatisées déjà employées en détection de signal pour la pharmacovigilance.

L'évaluation et l'éventuelle validation des différentes approches de détection de signal que ce soit sur les bases de notifications spontanées ou sur les bases médico-administratives est difficile car pour la quasi-intégralité des effets indésirables il n'existe pas de « gold standard » pour annoter les signaux obtenus (Hauben, Aronson, et Ferner 2016). Idéalement, l'évaluation de la pertinence des signaux renvoyés nécessite une étroite collaboration avec des experts pharmacologues.

Dans ce travail de thèse nous évaluerons les méthodes sur des critères statistiques mesurés sur des études de simulations. Nous proposerons également des études sur données réelles avec une

évaluation effectuée par les experts n'ayant pas pris part à la construction des méthodes et travaillant en aveugle sans savoir quelles méthodes ont généré quels résultats. Par ailleurs nous utiliserons également une base de connaissances pour un événement indésirable donné (M. Chen et al. 2016) établie indépendamment de la détection de signal en pharmacovigilance pour évaluer certains résultats toujours sur données réelles.

1.2.2 Prise en compte de facteurs de confusion par l'utilisation de modèles non paramétriques ou semi-paramétriques

Le modèle linéaire généralisé est à la base de la majeure partie des méthodes de détection de signal en pharmacovigilance récemment proposées. Dans ce cadre, où la matrice des données d'expositions est creuse, c'est-à-dire présente de nombreuses valeurs nulles, et constituée de données binaires, les modèles linéaires avec pénalisation, sont souvent assez efficaces.

Les modèles d'apprentissage statistique non paramétriques telles que les ensembles d'arbres sont potentiellement plus flexibles que les modèles linéaires généralisés et très adaptés à ce type de données (Breiman 2001). On peut donc supposer que leur utilisation permettra d'améliorer la performance en termes de prédictions mais également peut-être en termes de selection de variables. Ils peuvent être à la base d'estimateurs issus de l'inférence causale s'appuyant sur la construction de scores de propension, de g-computation ou

d'apprentissage ciblé.

Un objectif de cette thèse est de proposer et d'évaluer des approches dans un cadre d'apprentissage ciblé ou basée sur la g-computation sur les bases de notifications spontanées.

Par ailleurs dans le cadre des données issues de bases de données médico-administratives la présence de facteurs de confusion non mesurés pose problème y compris dans le cadre de la détection de signal. Récemment, l'emploi de méthodes auto-contrôlées (Simpson et al. 2013) a été proposé pour tenir compte de facteurs de confusion non mesurés mais invariant dans le temps. Si l'usage de tels modèles est courant en faible dimension, des problématiques existent en grande dimension pour tenir compte des nombreuses covariables présentes dans les bases médico-administratives. L'usage de tels modèles auto-contrôlés associés à des approches d'apprentissage statistique non paramétrique et à des outils issus de l'inférence causale sera étudié.

1.3 PLAN DE THESE

Dans un premier chapitre, nous présenterons les bases théoriques de différentes méthodes d'apprentissage statistique. Ces méthodes étant construites initialement plus pour la prédiction que pour l'inférence, nous étudierons comment obtenir des mesures d'importance des variables en couplant ces méthodes avec la théorie

de l'inférence causale.

Dans un deuxième chapitre, nous étudierons le schéma d'étude cas-croisé dans le cadre des bases médico-administratives. Dans ce schéma nous étudierons l'inclusion du score de propension en grande dimension. Une application à l'infarctus du myocarde sur données réelles sera proposée en plus d'une étude de simulations.

Dans un troisième chapitre, nous comparerons différentes approches de détection de signal sur bases de notifications spontanées. En particulier, les modèles paramétriques basées sur les modèles linéaires généralisés avec pénalisation LASSO seront comparés à des mesures d'importance des variables issues de modèles non paramétriques telles que décrites dans le premier chapitre. Une étude de simulation ainsi qu'une étude en cas réel sur les lésions hépatiques aiguës seront proposées.

2 METHODES D'APPRENTISSAGE STATISTIQUE EN GRANDE DIMENSION POUR LA DETECTION DE SIGNAL

Les principaux systèmes de pharmacovigilance reposent actuellement sur des méthodes de disproportionalité. Ces méthodes présentent l'avantage d'être assez peu complexes et coûteuses en calcul. Néanmoins, la représentation agrégée sous la forme d'une grande table de contingence des notification spontanées induit la perte de l'information individuelle, or les notifications spontanées mentionnent le plus souvent plusieurs médicaments suspects. En pratique, ces notifications se retrouvent dupliquées dans la table autant de fois que de médicaments impliqués, ce qui induit un biais car ces notifications ont alors un poids plus important.

Pour tenir compte de ces potentielles multi-expositions, des méthodes basées sur l'analyse de la structure de données sous forme de deux matrices creuses ont été développées et reposent sur des modèles de régressions pénalisés (Caster, O. et al. 2010; Ahmed, Pariente, et Tubert-Bitter 2018; Courtois, Tubert-Bitter, et Ahmed 2021). La première matrice creuse associe à chaque notification en ligne les médicaments suspects en colonnes. La deuxième matrice creuse associe à chaque notification en ligne les effets indésirables en colonnes. Ainsi ces deux matrices sont de grande dimension. Comme le nombre de médicaments suspects est faible au regard de l'ensemble des traitements possibles et de la même manière le nombre d'effets indésirables est faible au regard de l'ensemble des effets indésirables possibles, ces deux matrices comportent essentiellement des zéros. Pour permettre les traitements informatiques sur ces données de grande dimension, seules les coordonnées des valeurs non nulles sont

stockées en mémoire. La représentation de tels objets est appelée matrice creuse et dans le langage de programmation R une telle structure de données est implémentée dans le package Matrix.

Récemment, dans le domaine de l'apprentissage statistique de nombreux algorithmes prédictifs ont été développés tels que les modèles basés sur les ensembles d'arbres (Breiman 2001; Friedman 2001). Ces modèles sont plus flexibles pour l'ajustement aux données que les modèles linéaires généralisés ce qui leur confère potentiellement des avantages du point de vue de la prédiction. En effet, l'utilisation de modèles basés notamment sur des arbres permettrait de prendre en compte automatiquement et plus facilement des sous-populations fragiles ou des interactions entre expositions médicamenteuses.

En revanche, avec ce type d'approches, il est plus difficile de déterminer quelles sont, de manière globale, les variables ayant le plus d'importance dans la prédiction de la variable réponse. C'est pourquoi ces méthodes sont souvent considérées comme des « boîtes noires », c'est-à-dire comme efficaces pour la prédiction mais difficilement utilisables pour l'interprétation (Zhao et Hastie 2021).

Différentes approches ont été proposées pour pallier ce problème d'interprétabilité et en particulier des approches issues du domaine de l'inférence causale (Zhao et Hastie 2021). En effet, dans le domaine de l'inférence causale des développements récents se sont portés sur les approches d'estimation non paramétriques. Ces

approches font la distinction entre les paramètres du modèle statistique servant à l'ajustement aux données et un ou plusieurs paramètres d'intérêt servant à l'interprétation et définis préalablement. En particulier, la démarche de l'apprentissage ciblé propose de séparer l'ajustement du modèle aux données (ajustement réalisé par des méthodes d'apprentissage statistique pouvant être « boîtes noires ») avec l'estimation des paramètres servant à l'interprétation (Laan et Rose 2011).

Dans ce chapitre, nous présenterons dans un premier temps les développements récents autour des méthodes paramétriques pour la détection de signal en pharmacovigilance. Nous présenterons ensuite des méthodes d'apprentissage statistique basées sur les ensembles d'arbres ainsi que des mesures d'importance de variables proposées pour ces méthodes. Enfin, nous présenterons le cadre de l'inférence causale appliqué au contexte de la détection de signal en pharmacovigilance, plus précisement la g-computation, le score de propension et l'apprentissage ciblé.

2.1 NOTATIONS UTILISEES

On utilise les notations suivantes : soit $\boldsymbol{X}$ la matrice d'expositions médicamenteuses comportant n notifications et p médicaments. Cette matrice est composée uniquement de valeurs binaires caractérisant l'occurrence de chaque exposition médicamenteuse pour chaque notification.

Soit Y l'occurrence de l'événement indésirable d'intérêt comportant n valeurs binaires.

On considère le j-ème traitement d'intérêt comme une colonne de $\boldsymbol{X}$ que l'on note X_j et l'ensemble des autres $p-1$ traitements comme une matrice notée $\boldsymbol{X}_{-j}$. Pour un individu i on note ses expositions médicamenteuses X_{ij} et $\boldsymbol{X}_{i(-j)}$ pour le traitement d'intérêt et les autres expositions respectivement.

2.2 APPROCHES DERIVEES DU MODELE LINEAIRE GENERALISE AVEC AJOUT D'UNE PENALISATION

Les approches issues de modèles linéaires généralisés constituent une grande partie des méthodes utilisées pour mesurer la contribution d'une variable ou d'une exposition pour la prédiction d'une variable réponse. Les avantages du modèle linéaire généralisé sont la simplicité, l'interprétabilité directe des coefficients et de bonnes performances lorsque l'hypothèse de linéarité est respectée. Par ailleurs même lorsque cette hypothèse n'est pas tout à fait satisfaite le modèle est souvent assez robuste (Christodoulou et al. 2019). Dans le cadre de données d'expositions binaires (ce qui est le cas pour les données de notifications spontanées) les modèles linéaires sont souvent compétitifs en termes de prédictions même en étant comparés à des algorithmes plus sophistiqués (Cowling et al. 2021).

En revanche, en grande dimension, le modèle linéaire généralisé

peut être, malgré sa simplicité, non identifiable. En termes de calculs numériques, l'algorithme d'optimisation permettant d'estimer les coefficients peut présenter une absence de convergence. Pour remédier à ce problème l'emploi d'une pénalisation permet d'obtenir la convergence des coefficients du modèle au prix d'un biais dans la valeur des coefficients estimés.

Dans le cadre d'une variable réponse binaire la régression logistique est utilisée. Pour un échantillon de taille n, (X_i, Y_i), i variant de 1 à n, la log vraisemblance pénalisée du modèle est ainsi donné par :

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} y_i \log\left(\frac{1}{1+\exp(-\boldsymbol{\beta} X_i)}\right) + (1 - y_i) \log\left(1 - \frac{1}{1+\exp(-\boldsymbol{\beta} X_i)}\right) + \lambda\, p(\boldsymbol{\beta})$$

Lorsque des propriétés de parcimonie (*sparsity*) du modèle sont recherchées, la pénalité LASSO s'impose très souvent (Hastie, Friedman, et Tibshirani 2001). Elle est définie de la manière suivante :

$$p(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\boldsymbol{\beta_j}|,$$

avec p le nombre de prédicteurs considéré. Cette pénalité permet de contraindre de nombreux coefficients à une valeur de zéro. Appliquée à la détection de signal, on considère comme sélectionnées les variables présentant un coefficient estimé strictement positif. Une

difficulté avec ce type d'approche réside dans le choix de la pénalité λ qui détermine la parcimonie du modèle final. Avec un objectif de prédiction, cette pénalité est classiquement choisie par validation croisée. Néanmoins, l'utilisation de la régression LASSO dépasse désormais largement le cadre de la prédiction seule et de nombreux travaux méthodologiques se sont penchés sur son utilisation pour la sélection de variables ou l'inférence (Wasserman et Roeder 2009; Meinshausen et Bühlmann 2010; J. Chen et Chen 2008; Lockhart et al. 2014; Kammer et al. 2022).

L'utilisation du LASSO pour la détection de signaux en pharmacovigilance a été proposé initialement par Caster et al. (Caster et al. 2010). Ce travail a notamment illustré comment l'utilisation de cette approche multivariée pouvait pallier les limites des approches de disproportionalité tels que « l'effet de masquage » (Maignen et al. 2014), ce dernier résultant de l'analyse des données de notifications spontanées sous la forme d'une table de contingence. Plus récemment, Ahmed et al. (Ahmed, Pariente, et Tubert-Bitter 2018), en s'inspirant de l'algorithme Stability Selection (Meinshausen et Bühlmann 2010) ont proposé une approche reposant sur la combinaison du LASSO et d'un échantillonnage boostrap prenant en compte la nature déséquilibrée de la variable réponse (l'événement indésirable). Enfin, encore plus récemment, Courtois et al. se sont intéressés à l'utilisation du lasso adaptatif (Zou 2006) dans le cadre de la détection de signal et ont notamment proposé et comparé plusieurs choix de pondérations (Courtois, Tubert-Bitter, et Ahmed 2021).

2.3 APPRENTISSAGE STATISTIQUE PAR ENSEMBLE D'ARBRES

Les modèles de forêts aléatoires ont été proposés initialement par Breiman (Breiman 2001). Ils sont utilisés pour des tâches de prédictions quand la forme paramétrique de la dépendance entre variable réponse et expositions est laissée libre. Ces modèles sont formés d'un ensemble d'arbres, arbres qui pris individuellement sont des prédicteurs peu robustes mais progressent quand ils sont groupés en grand nombre. Friedman propose par la suite une alternative qui repose sur un algorithme glouton : le gradient boosting (Friedman 2001).

Le gradient boosting peut être vu comme une descente de gradient dans l'espace des prédictions tandis que la forêt aléatoire utilise le bootstrap ainsi qu'une sélection aléatoire de variables pour générer de l'aléa et ainsi obtenir de la diversité dans l'ensemble d'arbres.

Pour construire un arbre, on part de la racine vers les feuilles et en choisissant à chaque nœud une variable j et un seuil s de manière à maximiser le gain potentiel (réduction de la fonction de perte) (T. Chen et Guestrin 2016) défini comme :

$$Gain(j,s) = G_{\text{gauche}}(j,s)^2 + G_{\text{droite}}(j,s)^2 - G_{\text{sans noeud}}(j,s)^2 \quad (\text{Eq. } 1)$$

Les valeurs G_{gauche} et G_{droite} correspondent à la réduction de la fonction de perte obtenue en séparant le jeu de données en deux parties selon que les valeurs prises par la $j-ème$ variable soient inférieures ou supérieures au seuil s. Une illustration est donnée en Figure 1. La valeur $G_{\text{sans noeud}}(j,s)$ correspond à la perte empirique sans effectuer de séparation. A chaque nœud on choisit la variable j et le seuil s qui maximisent $\text{Gain}(j,s)$.

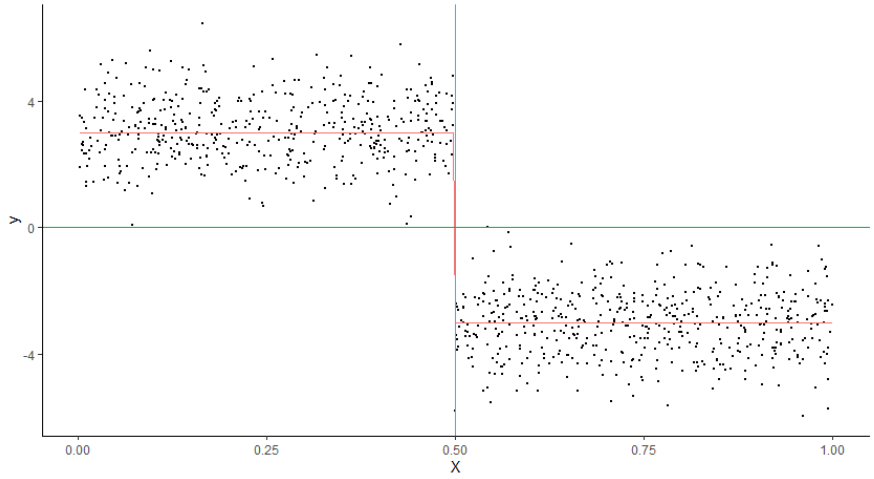

Figure 1: Illustration d'une séparation selon la variable X au seuil 0.5 (ligne bleue). $G_{sans\ noeud}(X,0.5)$ correspond à la somme des écarts quadratiques entre les points en noir et la ligne verte. Tandis que $G_{gauche}(X,0.5)+G_{droite}(X,0.5)$ correspond à la somme des écarts quadratiques entre les points en noir et la ligne rouge.

En termes de choix de fonction de perte, dans le cas d'une variable cible binaire on utilise de préférence l'entropie croisée :

$$L(y,\hat{p}) = -\frac{\sum_{i=1}^{n} y_i \log(\hat{p}_i) + (1-y_i)\log(1-\hat{p}_i)}{n}$$

où y correspond aux valeurs de la variable réponse et $\hat{p}$ correspond aux prédictions et n au nombre d'observations.

En utilisant la fonction sigmoïde comme fonction lien, on obtient comme prédiction $\hat{p}$ pour chaque individu, en notant q la valeur du paramètre individuel à optimiser pour se rapprocher d'un minimum de L :

$$\hat{p} = \frac{1}{1+\exp(-\hat{q})}$$

Afin de minimiser L, on peut calculer un gradient pour chaque q correspondant donc à chaque prédiction (avant passage au lien sigmoïde) :

$$\frac{\partial L}{dq}(y,q)_{|\,q=\hat{q}} = \frac{\partial}{dq}\left(-y\log\left(\frac{1}{1+\exp(-q)}\right) - (1 - \mathrm{y})\log\left(1-\frac{1}{1+\exp(-q)}\right)\right)$$

$$\frac{\partial L}{dq}(y,q)_{|\,q=\hat{q}} = \frac{\partial}{dq}\left(-y\log(1+\exp(-q)) - (1-y)\log(1+\exp(q))\right)$$

$$\frac{\partial L}{dq}(y,q)_{|\,q=\hat{q}} = exp(-q)\,y(\frac{1}{(1+exp(-q)}) - (1-y)\ exp(q)\ (\frac{1}{(1+exp(q)})$$

$$\frac{\partial L}{dq}(y,q)_{|\,q=\hat{q}} = y(1-\hat{p}) - (1-y)\hat{p}$$

$$\frac{\partial L}{dq}(y,q)_{|\,q=\hat{q}} = y - \hat{p}$$

Un calcul similaire nous donne la Hessienne pour chaque individu :

$$\frac{\partial^2 L}{dq^2}(y,q)_{|\,q=\hat{q}} = \frac{\partial}{dq}\left(-\frac{1}{1+\exp(-q)}\right)$$
$$\frac{\partial^2 L}{dq^2}(y,q)_{|\,q=\hat{q}} = \frac{\exp(-q)}{1+\exp(-q)}\frac{1}{1+\exp(-q)}$$
$$\frac{\partial^2 L}{dq^2}(y,q)_{|\,q=\hat{q}} = \hat{p}(1-\hat{p})$$

On note alors ces valeurs :

$$g = y - \hat{p}$$
$$h = \hat{p}(1-\hat{p})$$

On note que h est de dimension égale à celle de $\hat{p}$ car d'une part elle ne dépend que de q et d'autres part les individus étant distribués indépendamment la valeur q associée à un individu n'influe pas sur les autres valeurs de q.

Pour un sous ensemble d'individus, on a alors en notant i l'indicateur associée à chaque individu présent dans le sous-ensemble et n le nombre total d'individus présents dans le sous-ensemble :

$$G = -\sum_{i=1}^{n} g_i$$

et

$$H = \sum_{i=1}^{n} h_i$$

Contrairement à l'optimisation « classique », on ne cherche pas à optimiser des paramètres directement tels que le choix de variable j et le seuil s. Ces paramètres sont obtenus par une recherche quasi-exhaustive sur l'ensemble des valeurs possibles maximisant l'équation (1) avec :

$$G_{\text{gauche}}(j, s) = \sum_{i=1}^{n} \mathbf{1}_{X_{ij} < s} g_i$$

et

$$H_{\text{gauche}}(j, s) = \sum_{i=1}^{n} \mathbf{1}_{X_{ij} < s} h_i$$

L'algorithme pour la construction d'un arbre est le suivant :

Itérer :

1. Chercher la variable j et le seuil s maximisant le $Gain(j, s)$ potentiel.
2. Diviser le jeu de données en deux parties selon j, s.
3. Reprendre 1 et 2. Une fois atteint un critère d'arrêt, construire une feuille et y associer la valeur minimisant L dans cette feuille.

Deux méthodes existent principalement pour construire une

forêt à partir d'arbres : le bagging et le boosting.

Le bagging (Breiman 2001) consiste à tirer aléatoirement un échantillon bootstrap et une sélection de variables puis construire un arbre basé sur chaque échantillon bootstrap. L'estimateur est ensuite constitué par une moyenne de tous les arbres.

Pour le gradient boosting (Friedman 2001) les arbres sont construits séquentiellement sur tout l'échantillon, chaque arbre reprend les prédictions fournis par les arbres précédents et à partir de ce point de départ cherche à se rapprocher d'un minimum de la perte empirique. Celà consiste à l'étape $k+1$ à construire un arbre régressant tous les résidus obtenus à l'étape k à l'aide de toutes les variables disponibles. Pour permettre plus de robustesse et privilégier les modèles utilisant un plus grand nombre d'arbres, un paramètre de taux d'apprentissage α compris entre 0 et 1 peut être utilisé pour limiter la vitesse d'apprentissage.

En termes de prédiction, le gradient boosting est généralement un peu plus précis que les forêts aléatoires mais demande plus de contrôle des hyper paramètres. Comme ces modèles sont flexibles, pour empêcher le sur apprentissage, il convient d'utiliser la validation croisée à la fois comme critère d'arrêt du modèle d'apprentissage statistique mais également pour garantir que les prédictions obtenues ne proviennent pas d'individus ayant servi à établir le modèle. Pour les algorithmes d'apprentissage statistique basés sur le gradient boosting, il est préférable de considérer comme valeur de départ la moyenne de

la variable réponse.

Une limite de ces modèles est que les prédictions, étant conçues par une moyenne d'estimateurs, sont parfois concentrées vers une moyenne de la variable réponse et ne sont pas précises pour des valeurs extrêmes. Pour corriger ce phénomène une correction basée sur la régression non paramétrique isotone peut être appliquée (Niculescu-Mizil et Caruana 2005).

2.3.1 Approches basées sur une combinaison de plusieurs algorithmes

Récemment, en apprentissage statistique de nombreux travaux ont été menés pour le développement d'algorithmes prédictifs. Dans le cadre de la classification supervisée, les forêts aléatoires le gradient boosting (T. Chen et Guestrin 2016), l'apprentissage profond ou encore les machines à vecteurs de support sont autant d'approches ayant le même objectif : prédire une variable réponse du mieux possible à partir de variables explicatives en minimisant une fonction de perte qui est souvent l'entropie croisée. Contrairement à la régression logistique ces méthodes ne présentent pas de coefficients facilement interprétables. Ils sont ainsi utilisés le plus souvent pour des tâches purement prédictives.

Le superlearning aussi appelé « stacking » (Wolpert 1992; Laan et Rose 2011) est un méta algorithme effectuant une pondération de plusieurs algorithmes d'apprentissage statistique. Il fonctionne en entraînant tout d'abord les modèles d'apprentissage statistique

individuellement sur chaque sous-ensemble d'une partition des données. Puis on conserve les prédictions obtenues sur les sous-ensembles n'ayant pas servi à l'apprentissage. Enfin, on optimise une pondération des différents algorithmes sur les jeux de données obtenus en regroupant les partitions de données n'ayant pas servi à l'apprentissage (Figure 2).

Généralement le superlearner obtient de meilleurs résultats en termes de prédictions que le meilleur de chacun des sous algorithmes utilisés pour sa construction (Džeroski et Ženko 2004). Néanmoins, lorsqu'un algorithme le composant est nettement moins bon en prédictions que les autres algorithmes, le résultat peut être inférieur. Il est donc important de vérifier la performance des algorithmes pris individuellement pour s'assurer de bons résultats.

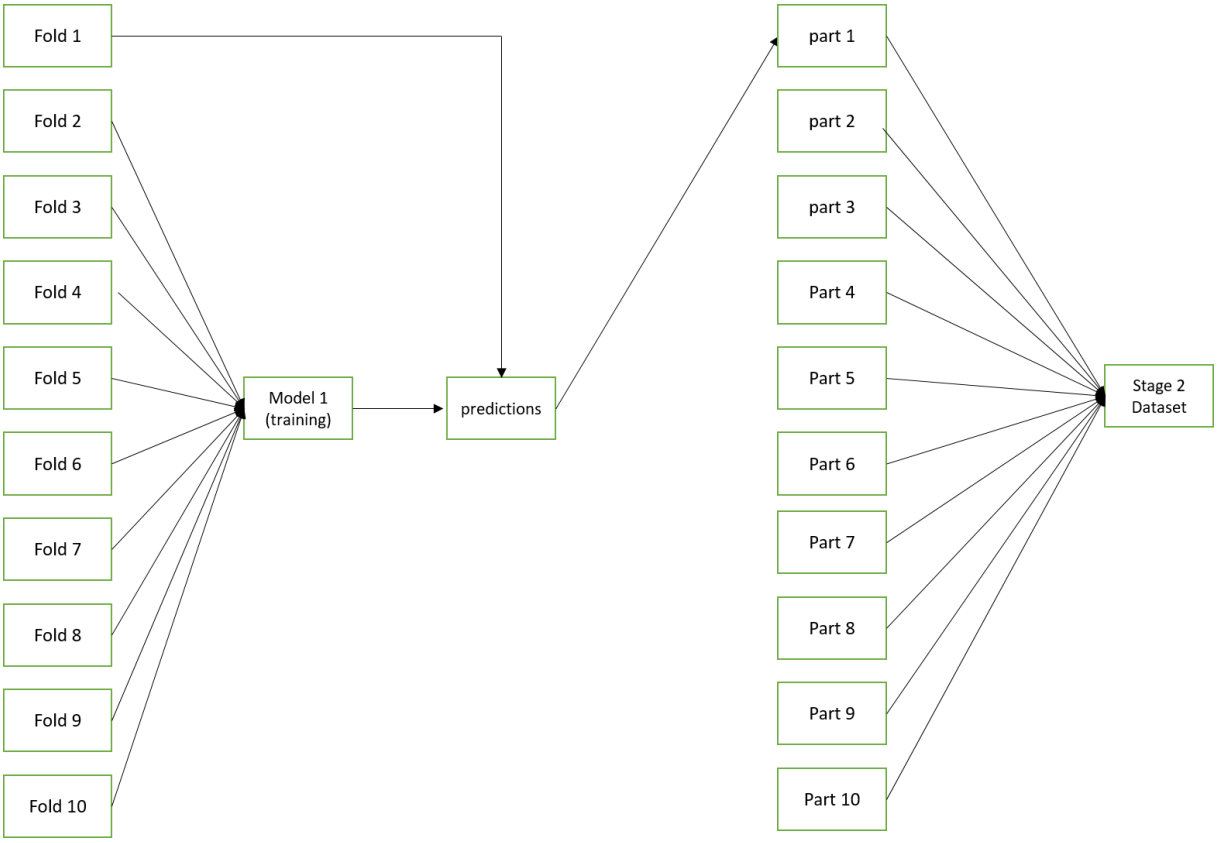

Figure 2: Principe d'entrainement de chaque algorithme composant le superlearner. Chaque partie est tour à tour exclue de l'ensemble d'apprentissage mais des prédictions sont générées qui forment une partie du deuxième jeu de données sur lequel est entraîné la pondération.

2.4 MESURES D'IMPORTANCE DES VARIABLES DANS LE CADRE DE MODELES STATISTIQUES NON PARAMETRIQUES

La flexibilité des modèles d'arbres de régression boostés ou des forêts aléatoires est un avantage certain pour la prédiction mais rend difficile l'interprétation. En particulier, le rôle précis de chaque variable est compliqué à interpréter quand le nombre d'arbres devient important. Plusieurs mesures d'importance des variables sont malgré tout proposées. Il faut alors distinguer des mesures locales (Lundberg et Lee 2017) qui correspondent à l'interprétation d'une prédiction pour un individu particulier de mesures globales (Natekin et Knoll 2013) qui correspondent à la mesure d'impact d'une variable sur l'ensemble de la population étudiée.

Le rôle de la détection de signal étant de se baser sur l'ensemble des observations pour juger de la présence d'un signal, les mesures d'importance globales sont privilégiées. Les mesures d'importance globale des variables les plus utilisées sont basées sur la somme des gains, en termes de réduction de la fonction de perte, effectués à chaque fois qu'une variable est choisie pour déterminer le critère de décision dans un nœud. Ces mesures ne peuvent pas être interprétées comme effets causaux ou comme analogues aux coefficients de la régression logistique. Une limite à ces mesures est aussi liée au fait que les algorithmes utilisés ont tendance à sélectionner plus facilement les variables présentant un grand nombre de modalités ou une plus forte

prévalence par rapport au reste des variables, ce qui peut conduire à des biais lors de l'analyse de l'importance respective des variables dans la prédiction (Strobl et al. 2007; Z. Zhang, Zhang, et Li 2023). Par ailleurs, si aucune contrainte n'est donnée au signe des poids associés aux feuilles cette approche ne permet pas de donner un sens à l'association entre une variable explicative et la variable réponse ce qui invalide l'emploi d'une telle approche pour la détection de signal où seules les variables positivement associées à l'occurrence de l'événement indésirable constituent des signaux potentiels.

2.5 OUTILS STATISTIQUES ISSUS DE L'INFERENCE CAUSALE ET DETECTION DE SIGNAL

Les approches d'inférence causale ont été développées pour permettre l'estimation d'un effet traitement sur une population définie. Des méthodes récentes, comme l'apprentissage ciblé (Laan et Rose 2011), permettent de mesurer cet effet traitement en utilisant des méthodes d'apprentissage statistique non paramétriques et également d'obtenir des tests statistiques. L'inférence causale se base sur trois hypothèses : la positivité c'est-à-dire la probabilité d'être traité ou non-traité n'est jamais nulle pour tous les individus, l'absence de facteurs de confusion non mesurés et la cohérence (Cole et Hernán 2008). Dans le cadre de ce travail, nous étudions l'application de méthodes issues de l'inférence causale à la détection de signal. L'objectif n'est donc pas l'interprétation causale mais juste une analyse

exploratoire. L'estimation de paramètres n'est qu'un prétexte pour pouvoir sélectionner des couples médicament-événements indésirable d'intérêt pour les experts pharmacologues.

Ainsi, en considérant une variable d'exposition dans nos bases de données comme traitement, les méthodes d'inférence causale développées comme l'apprentissage ciblé permettraient de construire une mesure d'importance des variables.

Généralement, les études d'inférence causale ont un seul traitement à étudier. Dans notre cas, l'ensemble des médicaments sera étudié. Comme la cardinalité de l'ensemble des combinaisons de traitements est extrêmement importante une simple boucle sur l'ensemble des traitements sera effectuée. Ainsi à chaque traitement correspond un unique paramètre et dans notre cadre pour chaque traitement fixé tous les autres traitements sont considérés comme facteurs de confusion sur lesquels il faut construire un ajustement.

2.5.1 Définition de quantités d'intérêt à estimer en utilisant le modèle des issues potentielles

Même sans objectif d'interprétation causale, le modèle des issues potentielles introduit par Rubin (Rubin 2005) offre un cadre pour la détection de signal en définissant des quantités d'intérêt qui permettent d'étudier la relation exposition médicamenteuse événement indésirable.

Les variables aléatoires dont sont issues les données disponibles sont notées typiquement P_0. De cette distribution multivariée est issu un jeu de données constitué de n individus comportant le triplé de variables $O_i = (Y_i, X_{ij}, X_{i(-j)})$ avec $i = 1, \dots, n$. Dans cette notation Y correspond à l'occurrence de l'effet indésirable, X_j est le traitement d'intérêt et $\boldsymbol{X_{-j}}$ l'ensemble des co-prescriptions le tout pour un total de n individus indépendants et identiquement distribués avec *i* variant de 1 à n la variable dénotant les individus. Le modèle des issues potentielles compare pour chaque individu la variable $Y1$ avec la variable $Y0$ qui représentent la valeur de la variable réponse, l'individu étant respectivement traité ou non traité. Seule une de ces variables est observée car un individu ne peut être traité et non traité à la fois. On représente cette situation par le tableau de données suivant (Tableau 1**Erreur ! Source du renvoi introuvable.**) :

Tableau 1: Modèle des issues potentielles, les points d'interrogation indiquent les valeurs manquantes.

Individu	X_j	$Y1$	$Y0$	$\boldsymbol{X_{-j}}$
1	1	1	?	$(X_{11}, \dots, X_{1(j-1)}, \dots, X_{1(j+1)}, \dots, X_{1p})$
2	1	0	?	$(X_{21}, \dots, X_{2(j-1)}, \dots, X_{2(j+1)}, \dots, X_{2p})$
3	0	?	1	$(X_{31}, \dots, X_{3(j-1)}, \dots, X_{3(j+1)}, \dots, X_{3p})$
4	0	?	1	$(X_{41}, \dots, X_{4(j-1)}, \dots, X_{4(j+1)}, \dots, X_{4p})$
…	…	…	…	…
$n-1$	1	1	?	$(X_{(n-1)1}, \dots, X_{(n-1)(j-1)}, \dots, X_{(n-1)(j+1)}, \dots, X_{(n-1)p})$
n	0	?	0	$(X_{n1}, \dots, X_{n(j-1)}, \dots, X_{n(j+1)}, \dots, X_{np})$

Ainsi, on estimera les valeurs manquantes de ce tableau à l'aide entre autres d'un modèle statistique non paramétrique f. Cela nous permettra d'obtenir, dans un deuxième temps et en utilisant ce modèle, une estimation du paramètre d'intérêt en comparant $Y1$ à $Y0$ avec les valeurs manquantes imputées par des estimations. On répétera ensuite l'approche en bouclant sur j, c'est-à-dire que chaque médicament deviendra tour à tour le médicament d'intérêt.

On notera que comme p est grand $(p > 1000)$ les 2^p $\text{combinaisons possibles}$ de traitements ne peuvent être étudiées exhaustivement. Ainsi l'approche proposée en effectuant une boucle est un compromis permettant à la fois d'étudier l'ensemble des médicaments et de proposer des tests statistiques sans nécessiter un temps de calcul trop important. Une correction pour les tests multiples sur les p tests pourra être effectuée basée sur l'approche de Benjamini-Hochberg afin de limiter le taux de fausses découvertes (Benjamini et Hochberg 1995).

2.5.2 Paramètres d'intérêt

Un paramètre d'intérêt pouvant être étudié est l'odds-ratio marginal :

$$OR = \frac{E(Y1)(1 - E(Y0))}{E(Y0)(1 - E(Y1))}.$$

L'objectif de l'étude n'est pas d'interpréter ce paramètre mais simplement de juger si une variable est importante ou non dans la construction d'un modèle non paramétrique prédisant la variable réponse. Le paramètre OR est donc d'une part une manière de juger de l'importance d'un traitement par rapport à la variable Y et d'autre part d'apporter un sens à l'association. Une exposition significativement associée à un OR supérieur à 1 étant considérée comme signal.

2.5.3 Estimation d'un effet traitement par g-computation

En utilisant des modèles non paramétriques, il y a rupture de la bijection entre modèle statistique et paramètres d'intérêt typiquement observée en inférence statistique paramétrique. Pour conduire l'estimation de paramètres en utilisant un modèle statistique non paramétrique, il faut construire une projection de l'espace fonctionnel d'où est issu le modèle statistique vers l'espace des paramètres. Pour cela la g-computation peut être utilisée (Robins 1986). Cette approche consiste à calculer des moyennes de prédictions établies par le modèle en forçant certaines variables en entrée à des valeurs fixées. Par exemple, pour mesurer un effet traitement on peut forcer la variable traitement à 1 et obtenir en moyennant un estimateur de $E\left(E\left(Y \middle| X_j = 1, \boldsymbol{X}_{-j}\right)\right)$. Un tel estimateur peut être construit par 1/n $\sum_{i=1}^{n} f(1, \boldsymbol{X}_{-j})$ où f est le modèle non paramétrique, n le nombre d'observations, $\boldsymbol{X}_{-j}$ les facteurs de confusion et où on impose la valeur 1 à la variable X_j.

2.5.4 Le rôle du score de propension

On appelle score de propension la probabilité de traitement étant donnés les facteurs de confusion,

$$P\left(X_{ij} = 1 \middle| \boldsymbol{X}_{i(-j)}\right)$$

Le rôle du score de propension est de prendre en compte le fait que la distribution de $\boldsymbol{X}_{(-j)}$ chez les traités puisse varier comparativement à la distribution chez les non traités.

En inférence causale, on peut montrer que quand $\boldsymbol{X}_{(-j)}$ vérifie le critère backdoor tel que défini par Pearl, c'est-à-dire que $\boldsymbol{X}_{(-j)}$ n'est pas descendant de X_j et bloque les chemins détournés entre Y et X_j dans le graphe causal (Pearl 1993) alors :

$$E\,(Y1\,) \;=\; E\,(E(Y\,|X_j =\; 1, \boldsymbol{X}_{(-j)})) \;=\; E(\frac{\mathbf{1}(\,X_j = 1)}{P(X_j = 1|\boldsymbol{X}_{(-j)})}Y)$$

et

$$E\,(Y0\,) \;=\; E\,(E(Y\,|X_j =\; 0, \boldsymbol{X}_{(-j)})) \;=\; E(\frac{\mathbf{1}(\,X_j = 0)}{P(X_j = 0|\boldsymbol{X}_{(-j)})}Y)$$

Ainsi on dispose de deux formes pour estimer l'effet traitement. Par exemple, chez les traités, la première forme $E\,(E(Y\,|X_j =\; 1, \boldsymbol{X}_{(-j)}))$ est une forme construite à partir d'un modèle prédisant y à partir de X. Une deuxième forme $E\left(\frac{\mathbf{1}(\,X_j=1)}{P\left(X_j = 1\middle|\boldsymbol{X}_{(-j)}\right)}Y\right)$ est basée sur l'inverse du

score de propension. L'intérêt du score de propension est donc de pouvoir corriger l'estimation de l'effet traitement avec un estimateur complémentaire de celui basé simplement sur la régression.

Il existe classiquement quatre manières de prendre en compte le score de propension : l'ajustement, la stratification, l'appariement et la pondération (Austin 2011).

L'ajustement consiste à inclure le score de propension comme variable supplémentaire dans le modèle de régression. La stratification est proche de l'ajustement et consiste à ajuster sur des quantiles du score de propension. L'appariement groupe les cas avec un certain nombre de témoins ayant un score de propension similaire puis conduit une analyse sur données appariées. La pondération consiste à créer des poids dépendant du score de propension et à effectuer une régression en utilisant ces poids. Plusieurs pondérations ont été proposées, la plus fréquente étant basée sur l'inverse du score de propension. Plus récemment, dans le cadre paramétrique de la régression logistique des pondérations plus stables ont été proposées comme les « overlap weights » qui sont construites sans passage à l'inverse du score de propension (Li, Thomas, et Li 2019).

2.5.5 Estimateur doublement robuste

On a vu qu'il y a deux formules pour l'estimation de l'effet traitement. Il est possible d'en tirer avantage pour construire un estimateur doublement robuste. Un estimateur doublement robuste

est consistant (c'est-à-dire correct asymptotiquement) si ou bien le modèle de régression ou le modèle du score de propension est correct. Un tel estimateur est constitué par la formule suivante :

$$\widehat{dr} = \hat{Q}\left(X_j, \boldsymbol{X}_{(-j)}\right) + \hat{\mathrm{r}}$$

Le premier terme est l'estimateur issu de la g-computation pour le modèle de la variable réponse. Le deuxième terme est défini, en notant $g1$ le score de propension, pour les traités et les non-traités respectivement par :

$$\hat{\mathrm{r}}_1 = \frac{\sum_{i=1}^{n} \mathbf{1}(X_{ij} = 1) \frac{1}{g1(X_{ij} = 1 | \boldsymbol{X}_{i(-j)})} (Y - \hat{Q}(X_{ij}, \boldsymbol{X}_{i(-j)}))}{\sum_{i=1}^{n} \mathbf{1}(X_{ij} = 1) \frac{1}{g1(X_{ij} = 1 | \boldsymbol{X}_{i(-j)})}}$$

$$\hat{\mathrm{r}}_0 = \frac{\sum_{i=1}^{n} \mathbf{1}(X_{ij} = 0) \frac{1}{1 - g1(X_{ij} = 1 | \boldsymbol{X}_{i(-j)})} (Y - \hat{Q}(X_{ij}, \boldsymbol{X}_{i(-j)}))}{\sum_{i=1}^{n} \mathbf{1}(X_{ij} = 0) \frac{1}{1 - g1(X_{ij} = 1 | \boldsymbol{X}_{i(-j)})}}$$

On voit que si le modèle f est correct alors quel que soit le score de propension l'estimateur de l'effet traitement est consistant. Inversement si le score de propension est correct l'éventuel biais de $\hat{Q}$ est corrigé par le score de propension et les résidus du premier modèle.

1.1.1.1 Apprentissage ciblé

L'estimateur du maximum de vraisemblance ciblé (Targeted

Maximum likelihood estimator, TMLE) est un estimateur doublement robuste. A ce titre, il intègre donc étape d'estimation d'un score de propension $\widehat{ps}$. Ce score de propension permet de construire deux variables : $H(1, \boldsymbol{X_{-j}}) = \frac{X_j}{\widehat{ps}}$ et $H(0, \boldsymbol{X_{-j}}) = \frac{1-X_j}{1-\widehat{ps}}$ afin de réajuster le modèle de la variable réponse $E(Y|X_j, \boldsymbol{X_{-j}})$ prioritairement chez les individus traités ressemblant à des non traités et chez les individus non traités ressemblant à des traités. Dans le cadre de l'approche d'apprentissage ciblé, ces deux variables sont injectées dans un modèle de régression logistique. On appelle cette étape le « targeted step » (Laan et Rose 2011) :

$$E(Y|X_j, \boldsymbol{X_{-j}}) = \frac{1}{1 + \exp(-\text{logit}\left(\overline{Q^0}(X_j, \boldsymbol{X_{-j}})\right) - \varepsilon_0 H(0, \boldsymbol{X_{-j}}) - \varepsilon_1 H(1, \boldsymbol{X_{-j}}))},$$

avec $\text{logit}(x) = \log(\frac{x}{1-x})$ et $\overline{Q^0}(X_j, \boldsymbol{X_{-j}})$ la prédiction initiale obtenue par un modèle d'apprentissage appliquée à la variable Y.

L'estimation des coefficients $(\varepsilon_0,\ \varepsilon_1)$ associés aux variables H permet de construire un modèle « ciblé » par rapport au paramètre d'intérêt, c'est-à-dire permettant la construction d'un estimateur final possédant de bonnes propriétés pour l'estimation du paramètre d'intérêt. A partir de :

$$\overline{Q^1}(X_j, \boldsymbol{X}_{-j}) = \frac{X_j}{1 + \exp\left(-\text{logit}\left(\overline{Q^0}(1, \boldsymbol{X}_{-j})\right) - \frac{\widehat{\varepsilon_1}}{\widehat{ps}}\right)} + \frac{1 - X_j}{1 + \exp(-\text{logit}\left(\overline{Q^0}(0, \boldsymbol{X}_{-j})\right) - \frac{\widehat{\varepsilon_0}}{1 - \widehat{ps}})}$$

Les quantités d'intérêt peuvent être ré-estimées :

$$\widehat{\mu_1{}^1} = \frac{\sum_{i=1}^{n} \overline{Q^1}\left(1, \boldsymbol{X}_{\boldsymbol{i}(-j)}\right)}{n},$$

$$\widehat{\mu_0{}^1} = \frac{\sum_{i=1}^{n} \overline{Q^1}\left(0, \boldsymbol{X}_{\boldsymbol{i}(-j)}\right)}{n},$$

D'où,

$$\widehat{OR}^1 = \frac{\widehat{\mu_1{}^1}(1 - \widehat{\mu_0{}^1})}{\widehat{\mu_0{}^1}(1 - \widehat{\mu_1{}^1})}.$$

Théoriquement, la construction de cet estimateur et l'utilisation du score de propension lui confèrent de meilleures propriétés par rapport à la g-computation pour un unique traitement telles que la double robustesse. La double robustesse assure la consistance de l'estimateur si au moins un des deux modèles est correct : ou bien celui pour le score de propension, ou bien celui pour la variable réponse.

L'application de l'apprentissage ciblé dans le cadre de la détection de signal en grande dimension constitue une proposition qui est en revanche plus lourde à mettre en œuvre du point de vue du calcul informatique que la g-computation si on étudie un grand

nombre de traitements.

1.1.1.2 Intervalles de confiance

Le bootstrap permet de calculer des intervalles de confiance pour $\widehat{OR}^1$. Néanmoins cette approche est assez lourde en temps de calculs car il faudrait calculer un score de propension pour chaque traitement et pour chaque échantillon bootstrap.

Une autre possibilité pour calculer cet intervalle de confiance pour $\widehat{OR}^1$ est donnée par l'utilisation de la courbe d'influence de l'estimateur. La courbe d'influence (CI) correspond pour chaque individu à la variation de l'estimation obtenue quand le poids associé à un individu change infinitésimalement.

Mathématiquement, pour un estimateur $\hat{\theta}$, la courbe d'influence de $\hat{\theta}$ en x correspond à la limite :

$$CI_{\hat{\theta}}(x) = \lim_{\varepsilon \to 0} \frac{\hat{\theta}(P_\varepsilon^x) - \hat{\theta}(P_0)}{\varepsilon},$$

où P_0 correspond à la distribution de probabilité du jeu de données $O_i = (Y_i, X_{ij}, X_{i(-j)})$ et $P_\varepsilon^x = (1-\varepsilon)P_0 + \varepsilon\delta_x$ en notant δ_x la distribution de Dirac en x et pour tout $\varepsilon \neq 0$.

En se basant sur la théorie des statistiques asymptotiques, pour un estimateur consistant et asymptotiquement linéaire comme celui induit par le TMLE, la variance de la courbe d'influence est proportionnelle à la variance de l'estimateur :

$$\mathrm{VAR}\left(\widehat{OR}^1\right) = \frac{\mathrm{VAR(CI)}}{n}.$$

Cela permet par la suite de construire un intervalle de confiance asymptotique :

$$IC_{95\%}\left(\log(\widehat{OR}^1)\right) = \log(\widehat{OR}^1) \pm 1.96\sqrt{\frac{\mathrm{VAR}(CI)}{n}}.$$

Par ailleurs l'estimateur induit par le TMLE se base sur la courbe d'influence efficace de l'OR ce qui signifie qu'il est efficace asymptotiquement, si correctement spécifié, dans la classe des estimateurs sans biais asymptotiquement linéaires. De la même manière qu'un estimateur paramétrique peut être efficace dans la classe des estimateurs sans biais (Wasserman 2004) c'est-à-dire atteindre la borne de Fréchet-Darmois-Cramer-Rao, l'estimateur du TMLE atteint une borne asymptotique d'optimalité dans le cadre d'estimateurs non paramétriques.

1.1.1.3 Importance de la validation croisée

L'apprentissage ciblé permet l'utilisation de modèles flexibles tels que les arbres de régression boostés que ce soit pour l'estimation de $\overline{Q^0}\left(X_j, \boldsymbol{X}_{-j}\right)$ ou du score de propension. Ainsi, même si nous effectuons la sélection d'hyper paramètres par validation croisée pour nos modèles estimant le score de propension et la variable réponse, les arbres de régression boostés étant très flexibles, il peut y avoir sur-apprentissage. Il est donc souvent recommandé, en utilisant des

modèles d'apprentissage statistique, d'effectuer les prédictions servant ensuite à la construction du deuxième modèle ciblé de manière « out of fold » (Levy 2018). Nous définissons ainsi l'estimateur $\widehat{OR^1}_{\mathrm{CV}}$ construit de la manière suivante :

1. Créer une partition de l'échantillon global en K sous parties P_k pour k variant de 1 à K.
2. Pour chaque sous partie P_k construire deux estimateurs pour la variable réponse et pour le score de propension entrainés sur tout l'échantillon à l'exception de P_k.
3. Prédire la variable réponse et le traitement à l'aide de ces modèles pour chaque individu de P_k. Regrouper toutes les prédictions dans un même jeu de données.
4. Effectuer le « targeted step »
5. Estimer $\widehat{OR^1}_{CV}$

Cette approche est potentiellement lourde du point de vue du temps de calcul mais permet d'éviter le sur-apprentissage et d'obtenir des estimations plus robustes.

3 SCORE DE PROPENSION ET DESIGN AUTOCONTROLE POUR LA DETECTION POUR LA DETECTION DE SIGNAL A PARTIR DES BASES MEDICO-ADMINISTRATIVES

1.2 INTRODUCTION ET CONTEXTE

Aujourd'hui, alors que la plupart des systèmes de détection de signal en pharmacovigilance, que ce soit au niveau national ou mondial, sont basés sur les bases de notifications spontanées émises par les médecins ou pharmaciens un intérêt grandissant se porte sur les bases de données médico-administratives. Ces bases de données médico-administratives ont déjà fait l'objet de premières études pour la détection de signal (Pariente et al. 2007; Arnaud et al. 2018; Salvo et al. 2013).

En France, le système national des données de santé (SNDS) forme une base qui contient l'information à propos des remboursements de médicaments et des hospitalisations pour une grande partie de la population. Ainsi, cette base présente potentiellement une grande valeur pour la pharmacovigilance. L'échantillon généralisé des bénéficiaires (EGB) constituait un échantillon représentatif de cette base réduit au 1/97ème. Certains travaux sur cet échantillon ont déjà été effectués au niveau national (Arnaud et al. 2018; Demailly et al. 2020).

A l'international, sur d'autres bases de données médico-administratives, l'OMOP (Observational Medical Outcomes Partnership), puis le successeur à cette organisation l'OHDSI (Observational Health Data Sciences and Informatics), ont développé d'importants travaux pour la mise en forme des données et leur

exploitation pour la détection de signal en pharmacovigilance par des méthodes statistiques (Ryan et al. 2013). Une large étude comparative de sept méthodes automatisées de détection de signal a notamment été conduite. Chaque méthode a été évaluée sur la base d'un ensemble de référence comprenant des données sur 399 médicaments couplés à quatre événements indésirables. Cette étude a mis en avant l'intérêt des méthodes autocontrôlées pour la détection de signal sur les bases médico-administratives notamment la méthode de la série de cas (Farrington 1995). En ne travaillant que sur des cas, c'est-à-dire par comparaison de périodes intra-sujet, les méthodes autocontrôlées présentent l'avantage de contrôler tous les facteurs fixes de confusion.

Ces méthodes sont le plus souvent utilisées en pharmaco-épidémiologie sur des jeux de données de dimension faible. Une de ces méthodes est la série de cas (SCCS) qui est une méthode comparant l'occurrence de la variable réponse entre des périodes où l'individu est exposé et des périodes sans esposition. Outre la série de cas le cas-croisé (case-crossover) est notamment utilisé pour identifier des expositions pouvant déclencher des effets aigus et est utilisé fréquemment pour les études traitant de l'infarctus du myocarde, un événement ou l'utilisation d'expositions post-effet indésirable (telle qu'utiliser par la série de cas) n'est pas acceptable (Maclure 1991).

Les bases de données médico-administratives manquent généralement d'information sur l'état de santé général des patients. Des facteurs de confusion importants tels que le statut tabagique, le

surpoids ou l'activité physique ne sont pas observés dans ces bases. En revanche ces bases contiennent de nombreuses covariables (co-expositions médicamenteuses ou données d'hospitalisation) pouvant permettre d'approcher l'état de santé des patients. Il s'agit d'une approximation empirique construite à partir de la grande richesse des données.

L'intérêt théorique des méthodes autocontrôlées est de pouvoir à partir des individus cas seulement de prendre en compte les facteurs de confusion statiques, c'est à dire non dépendant du temps, les sujets étant leur propre témoin. C'est pourquoi elles ont été considérées pour l'utilisation de ces bases médico-administratives.

Dans le domaine des études pharmaco épidémiologiques. Schneeweiss *et al.* (Sebastian Schneeweiss et al. 2009) ont émis l'hypothèse que le grand nombre de variables présentes pouvait constituer un proxy des facteurs de confusion non observés. Ainsi un algorithme nommé high dimensional propensity score (HDPS) fut construit pour prioriser les covariables associées marginalement à la fois au traitement d'intérêt et à la variable réponse mesurant l'occurrence de l'événement indésirable.

L'utilisation de cet algorithme HDPS couplée à une méthode autocontrôlée n'a pas encore été étudiée pour la détection de signal en pharmacovigilance. Ainsi, nous proposons de tester l'efficacité de la combinaison de ces deux approches en présence, notamment, de facteurs de confusion non dépendant du temps et non observés, et

d’autres facteurs de confusion potentiellement dépendant du temps mais observés ou pouvant être approchés dans les bases médico-administratives.

Il est clair que cet algorithme ne donnera pas lieu à une interprétation causale à cause de la nature des données utilisées. De plus, l’odds-ratio obtenu ne peut être assimilé aux odds-ratio marginaux obtenus par les méthodes d’inférence causale et le cas-croisé est également biaisé pour les expositions présentant une tendance dans le temps. Néanmoins, pour la détection de signal cette approche peut être intéressante car elle permettrait de corriger certains comme le biais d’indication.

Les deux principaux objectifs de ce travail sont donc d’étendre l’utilisation du cas-croisé au contexte de la détection de signal avec de nombreuses expositions et de présenter une nouvelle méthode nommée Propensity Score adjusted Case-Crossover (PS-CC) qui utilise le cas-croisé comme schéma d’étude avec le score de propension en grande dimension.

Un avantage potentiel de cette approche est de permettre de fixer le seuil de détection au regard de critères statistiques s’appuyant sur la théorie des tests multiples tels que le False Discovery Rate (FDR) (Benjamini et Hochberg 1995). Cet aspect est important car le choix du seuil de détection affecte directement la quantité de travail d'évaluation à effectuer par les experts pharmacovigilants. En comparaison, la définition du niveau de régularisation pour effectuer

la sélection de variables avec des approches de régression pénalisées telles que le LASSO reste un domaine de recherche en cours.

L'approche PS-CC a été évaluée à la fois par des simulations de Monte Carlo et par une étude de cas réels. Elle a été comparée (i) à un schéma d'étude cas-croisé univarié (U-CC) appliqué à tous les médicaments examinés et (ii) à une approche cas-croisé multivariée basée sur le LASSO (LASSO-CC). L'étude de cas réels a porté sur l'événement indésirable infarctus aigu du myocarde et a été réalisée sur les données de l'EGB pour la période allant du 1er janvier 2010 au 31 décembre 2016. Les trois approches ont été appliquées à la cohorte et les signaux générés par au moins une méthode ont été évalués par deux experts en sécurité des médicaments.

3.1 METHODES

3.1.1 Notations

On rappelle que les notations suivantes sont utilisées : soit $\boldsymbol{X}$ la matrice d'expositions médicamenteuses comportant n notifications et p médicaments. Soit Y l'occurrence de l'événement indésirable comportant des valeurs binaires. On considère le j-ème traitement d'intérêt comme une colonne de $\boldsymbol{X}$ que l'on note X_j et l'ensemble des autres $p-1$ traitements comme une matrice notée $\boldsymbol{X}_{-j}$. Pour un individu i on note ses expositions médicamenteuses X_{ij} et $\boldsymbol{X}_{i(-j)}$ pour le traitement d'intérêt et les autres expositions respectivement.

Dans cette partie, on introduit en plus la notion de période de temps. Ainsi on notera $\boldsymbol{X_t}$ la matrice d'expositions médicamenteuses durant la période t, avec t variant de 1 à T. On introduit X_{ijt} l'exposition à une exposition spécifique j pour un individu i durant la période t. La notation $\boldsymbol{X}_{i(-j)t}$ dénote l'ensemble des autres expositions et Y_{it} la survenue de l'événement indésirable durant cette période et pour l'individu i.

3.1.2 Le schéma d'études cas-croisé

Le schéma d'étude cas-croisé a été introduit par Maclure (Maclure 1991). Comme tous les schémas d'études autocontrôlés il se base uniquement sur les cas (individus ayant subi l'événement indésirable) pour évaluer l'association statistique entre une exposition d'intérêt et l'occurrence de l'évènement indésirable. Pour cela, l'exposition durant la période cas (période située juste avant la survenue de l'évènement indésirable) est comparée aux expositions durant les périodes témoins situées avant la période cas (voir Figure 3). Le choix de périodes témoins proches de la période cas que ce soit en termes de temporalité ainsi qu'en termes de risques est important (Mittleman et Mostofsky 2014) afin de pouvoir justifier de l'échangeabilité des périodes pour un individu donné. Toutes les périodes doivent être séparées par des périodes « wash-out » pour éviter les autocorrélations entre périodes étudiées.

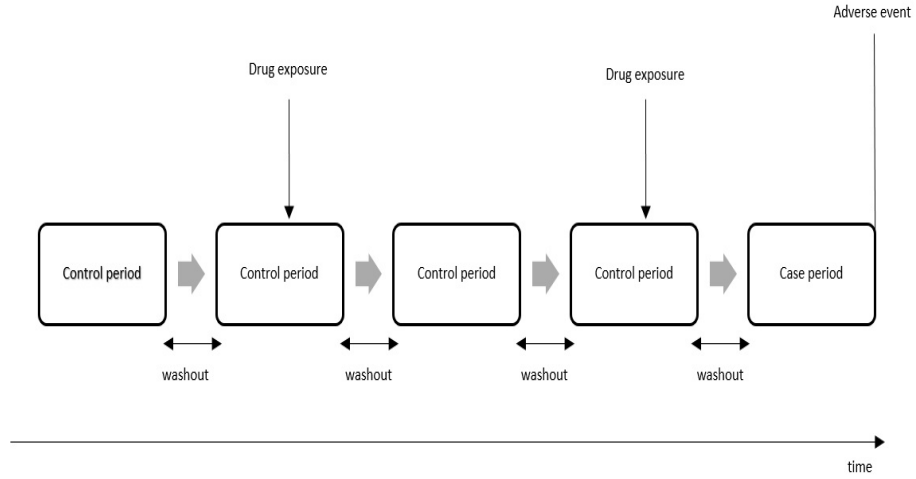

Figure 3 : le schéma d'étude du cas-croisé.

Ainsi ce schéma d'étude est un schéma cas-témoin apparié avec un appariement effectué sur l'individu, et une période cas située juste avant la survenue de l'événement indésirable et plusieurs périodes témoins. Les comparaisons sont effectuées uniquement de manière intra-individu. Le test de Cochran–Mantel–Haenszel et plus récemment la régression logistique conditionnelle sont utilisés pour effectuer l'inférence statistique. Comme notre étude vise à travailler sur de multiples expositions nous développerons uniquement la régression logistique conditionnelle.

3.1.3 Régression logistique conditionnelle

La régression logistique conditionnelle reprend la forme de la régression logistique mais en incluant à la place de l'intercept global un intercept dépendant d'une variable d'appariement. Cette approche est importante car utilisée dans les études cas-témoins matchés, les séries de cas et les cas-croisés, qui correspondent à des schémas d'études utilisés pour la détection de signal.

Pour un individu i avec $i = 1, \dots, n$, et une période $t = 1, \dots, T$, l'occurrence de la variable réponse est donnée par :

$$\Pr(Y_{it} = 1 | X_{ijt}) = \frac{\exp(\alpha_i + \beta X_{ijt})}{1 + \exp(\alpha_i + \beta X_{ijt})}$$

où l'intercept dépendant de la variable d'appariement est noté α_i et la mesure d'association de l'exposition d'intérêt avec l'événement indésirable est notée β.

L'estimation de l'intercept α_i n'est pas d'intérêt et seul β est important y compris pour la détection de signal. Par ailleurs on sait qu'il ne peut y avoir qu'un seul cas pour chaque strate. En conditionnant sur cette information on peut dériver la log-vraisemblance conditionnelle (Hardin 2005) :

$$l\left(\beta | \sum_{t=1}^{T} Y_{it} = 1\right) = \sum_{i=1}^{n} \left[\sum_{t=1}^{T} Y_{it} \beta X_{ijt} - \log\left(\sum_{t=1}^{T} \exp(\beta X_{ijt})\right)\right]$$

On note alors que cette formule ne dépend plus des paramètres α_i. L'estimation du coefficient β peut se faire à partir de cette formule par maximisation de la vraisemblance conditionnelle.

Pour la régression logistique conditionnelle univariée ou multivariée sous réserve de convergence un test statistique peut être effectué.

On note que le nombre de périodes T est le même pour tous les individus dans le cadre de notre étude.

3.1.3.1 Régression logistique conditionnelle multivariée

On distinguera la régression logistique conditionnelle multivariée en faible dimension de la régression logistique conditionnelle en grande dimension. Dans la première sous des hypothèses de faibles corrélations entre variables, on peut simplement étendre le modèle présenté au paragraphe précédent en considérant un vecteur $\beta = (\beta_1, \dots, \beta_p)$ d'associations statistiques entre les p variables considérées et l'événement indésirable.

En grande dimension (avec plus de $p > 100$ expositions), une telle approche ne converge pas forcément. Il existe plusieurs stratégies pour permettre d'opérer une inférence statistique dans ce contexte. Une approche simple est d'effectuer une boucle de la méthode univariée décrite dans la section précédente pour chaque exposition d'intérêt et d'effectuer une correction pour les tests multiples par la

suite. Une telle approche ne prend pas compte des éventuelles corrélations entre variables.

3.1.3.2 Méthode LASSO pour la régression logistique conditionnelle

Une meilleure approche pour gérer les instabilités numériques de la régression logistique multivariée en grande dimension est d'ajouter à la log vraisemblance conditionnelle une pénalité afin de favoriser un modèle robuste (Simpson et al. 2013). Une pénalité utilisée classiquement en grande dimension est la pénalité LASSO. Une telle pénalité associée à la série de cas a déjà été proposé dans le cadre des travaux menés par l'OMOP.

La log vraisemblance conditionnelle pénalisée s'écrit :

$$l_\lambda\left(\beta|\sum_{t=1}^{T} Y_{it}=1\right) = \sum_{i=1}^{n}\left[\sum_{t=1}^{T} Y_{it}\beta^T \boldsymbol{X}_{it} - \log\left(\sum_{t=1}^{T}\exp\left(\beta^T \boldsymbol{X_{ip}}\right)\right)\right] - \lambda\sum_{j=1}^{m}|\beta_j|.$$

L'utilisation d'un tel modèle permet la détection de signal car certains coefficients seront estimés comme valant exactement 0. Ainsi seules les expositions médicamenteuses présentant un coefficient positif strictement seront considérées comme signaux potentiels.

Le choix de λ contrôle la complexité du modèle, une valeur faible favorise un modèle avec peu de covariables sélectionnées tandis qu'une valeur plus forte favorise un modèle avec beaucoup de

coefficients présentant une valeur non nulle. Plusieurs heuristiques existent pour choisir la valeur de ce paramètre, la plus courante étant d'utiliser la validation croisée pour sélectionner la valeur λ associée au modèle généralisant le mieux en dehors de l'ensemble d'entraînement.

Néanmoins un tel choix a tendance à sélectionner un peu trop de variables (Courtois, Tubert-Bitter, et Ahmed 2021). Nous choisissons par la suite de recourir à une heuristique basée sur un critère d'information d'Akaike, et ainsi de choisir la valeur λ^{opt} minimisant ce critère suivant :

$$\lambda^{\text{opt}} = \text{argmin}_{\lambda}\ 2K(\lambda) - 2l_{\lambda}(\hat{\beta})$$

Dans cette formule K est le nombre de prédicteurs actifs c'est-à-dire dont les coefficients associés sont différents de zéro et le modèle choisi est le modèle associé à la pénalité λ^{opt}.

3.1.4 Proposition d'une méthode basée sur le score de propension en grande dimension

Comme il n'est pas forcément possible, en grande dimension, d'ajuster sur toutes les covariables sans pénaliser la log-vraisemblance conditionnelle, nous proposons de réduire l'ensemble des covariables d'ajustement à une seule, le score de propension (PS), et d'ajuster le modèle sur cette variable. Les méthodes de score de propension sont principalement utilisées dans les études de cohorte. L'approche PS-CC est construite par analogie à l'utilisation du score de propension dans

le cas d'études cas-témoins. Pour ce faire, nous utilisons pour calculer le score de propension les périodes témoins uniquement en considérant l'occurrence de l'événement indésirable comme rare. Cette méthode d'estimation du score est moins biaisée que celle utilisant l'ensemble des données disponibles (Månsson et al. 2007).

Pour une exposition médicamenteuse d'intérêt $j \in \{1, \dots, p\}$, nous cherchons à identifier dans un premier temps l'ensemble U des facteurs de confusion potentiels. En utilisant la méthode HDPS établie par Schneeweiss *et al.* (Sebastian Schneeweiss et al. 2009), les facteurs de confusion potentiels sont sélectionnés sur la base de la formule de Bross qui est une mesure des associations marginales entre chacune des covariables notée c et la variable réponse Y d'une part et la variable d'exposition j d'autre part :

$$BIAS(c, j) = \frac{P_{c1}(RR_{cY} - 1) + 1}{P_{c0}(RR_{cY} - 1) + 1}$$

Dans cette formule, P_{c1} et P_{c0} désignent respectivement la prévalence de la variable c chez les exposés à j et les non-exposés à j et RR_{cY} désigne le risque relatif entre c et Y.

Dans un deuxième temps un score de propension est calculé à partir des variables sélectionnées par HDPS. Pour cette étape on utilise uniquement les périodes témoins conformément à la méthode des témoins uniquement décrite par Månsson *et al.* (Månsson et al. 2007). Cette exclusion des périodes cas permet de réduire le biais dû à la

surreprésentation de la période cas dans le cadre de l'étude d'un événement rare.

L'utilisation de méthodes d'apprentissage statistique est possible pour calculer le score de propension et permet de sélectionner automatiquement les interactions entre variables, prédictives de la variable réponse. Nous proposons d'utiliser une méthode non paramètrique, à savoir des arbres de régression boostés.

La troisième étape consiste à ajuster sur le score de propension la régression logistique conditionnelle selon la formule suivante.

$$\Pr\left(Y_{it} = 1 | X_{ijt}\right) = \frac{\exp\left(\alpha_i + \beta_j X_{ijt} + \gamma_j \widehat{ps}_{it}\right)}{1 + \exp\left(\alpha_i + \beta_j X_{ijt} + \gamma_j \widehat{ps}_{it}\right)},$$

et d'obtenir une statistique de test pour β_j.

On boucle ensuite sur l'ensemble des variables d'intérêt pour obtenir une mesure d'importance pour chaque variable et on peut corriger la p-value pour tenir compte des tests multiples et utiliser une méthode comme celle de Benjamini Hochberg (Benjamini et Hochberg 1995) pour obtenir un seuil de détection correspondant à un False Discovery Rate (FDR) visé.

En résumé, l'algorithme PS-CC est composé des étapes suivantes pour une exposition j :

1. Utilisation de HDPS pour trouver les traitements sur lesquels ajustés parmi $\boldsymbol{X}_{-j}$. On note $\boldsymbol{X}^*_{-j}$ l'ensemble de ces traitements.
2. Pour réduire le biais présent à cause de l'utilisation d'un schéma cas croisé (assimilable à un schéma cas-témoin), on élimine ensuite les périodes cas avant d'estimer le score de propension.
3. On entraine un modèle d'apprentissage statistique pour estimer le score de propension noté, pour un individu i et une période de temps t, $\widehat{ps}_{it} = P\left(X_{ijt} = 1 \middle| \boldsymbol{X}^*_{i(-j)t}\right)$.
4. Enfin on ajuste un modèle de régression logistique conditionnelle en ajustant sur ce score de propension.

Pour obtenir une estimation pour chacun des traitements il faut ensuite effectuer une boucle pour j variant de 1 à p. Une correction sur les tests multiples peut ensuite être effectuée.

3.2 ETUDE DE SIMULATIONS

Nous proposons une étude de simulations pour comparer les méthodes présentées dans la partie précédente à savoir le cas-croisé univarié avec boucle sur l'ensemble des expositions médicamenteuses (U-CC), le cas-croisé avec pénalité LASSO (LASSO-CC) et l'approche basée sur le score de propension (PS-CC).

3.2.1 Création d'un jeu de données simulées

Pour créer un jeu de données simulées réaliste, nous utilisons

les données issues du SNDS et en particulier de l'EGB pris dans l'intervalle de temps entre 2010-01-01 et 2015-12-15 et composé de 3000 personnes ayant souffert d'un infarctus du myocarde sur cette période. Nous conservons les données d'expositions médicamenteuses de ces individus que nous notons $\boldsymbol{X}$. La variable réponse Y est entièrement simulée.

Pour mettre en forme $\boldsymbol{X}$, nous considérons 23 périodes de temps de 30 jours pour chaque individu, toutes séparées par des périodes de washout. L'exposition médicamenteuse est considérée binaire et vaut 1 s'il y a eu remboursement du médicament associé durant la période et 0 sinon. Pour limiter le temps de calcul nous considérons seulement les médicaments prescrits à au moins 50 patients ($p = 108$).

Une fois la mise en forme de $\boldsymbol{X}$ établie, Y est simulé comme une variable binaire dépendant à la fois de $\boldsymbol{X}$ et d'un effet individuel propre à chaque patient noté α_i, avec $i = 1, \dots, n$, identifiant les patients. Pour la dépendance à *Y*, un nombre de médicaments p_y pouvant valoir 0, 5 ou 20 sont tirés aléatoirement et on note β_{select} l'indicatrice de cet ensemble. Ces médicaments sont par la suite associés à Y avec une force d'association β égale à 0, 1 ou 1.5. Les médicaments absents de cet ensemble ne sont pas associés à Y. Ainsi pour chaque période de $\boldsymbol{X}$ une valeur de Y est associée par les formules suivantes :

$$\alpha_i \sim U(-1.5, 1.5)$$

$$\Pr(Y_{it} = 1|\boldsymbol{X}_{it}) = \frac{\exp(\alpha_i + X_{it}(\beta_{select} * \beta))}{1 + \exp(\alpha_i + X_{it}(\beta_{select} * \beta))}.$$

Dans ces formules i indique le patient et t la période associée. L'objectif des simulations est de comparer la capacité des différentes méthodes à trouver les médicaments associés à Y.

3.2.2 Critères de comparaison

Chaque scénario est répété 500 fois et les résultats moyens sont fournis. Nous appliquons deux critères d'évaluation pour comparer les méthodes. Le premier est le nombre de détections de médicaments réellement associés à la variable réponse Y à un nombre fixé de signaux générés. Ce critère mesure la capacité des méthodes à fournir un classement correct des médicaments.

Le deuxième critère évalue la capacité des méthodes permettant un contrôle du FDR c'est-à-dire les méthodes PS-CC et U-CC à contrôler effectivement le FDR au niveau $\alpha = 5\%$ et $\alpha = 15\%$ et la sensibilité associée en appliquant cette correction FDR. Le FDR est estimé en moyennant le résultat des 500 réplications de la proportion de fausses découvertes (FDP) observée. Le FDP ainsi reporté vaut 0 si aucun signal n'est détecté. Ces valeurs sont également fournies à titre indicatif pour la méthode LASSO-CC même si aucune correction n'est appliquée.

Tous les calculs ont été effectués avec le logiciel R en utilisant les paquets Cyclops pour LASSO-CC et xgboost pour les arbres de

régression boostés en utilisant eta = 0,1, max_depth = 4. Le nombre maximum de covariables utilisées *U* a été fixé à 50.

3.2.3 Résultats

La Figure 4 illustre les résultats comparés des trois méthodes pour le premier critère d'évaluation. On observe que LASSO-CC et PS-CC sont au moins aussi efficaces que la méthode univariée U-CC. Pour $p_y = 5$ et pour les grandes valeurs de β (1 et 1.5), on observe que toutes les méthodes trient parfaitement les variables selon leur association à y.

Le Tableau 2, met en avant les performances des algorithmes pour le deuxième critère d'évaluation à savoir le contrôle potentiel du FDR et la sensibilité obtenue en appliquant une correction pour ce contrôle aux niveaux 5% et 15%. La méthode PS-CC est capable de contrôler le FDR dans tous les scénarios, mais la sensibilité obtenue peut être faible. Ainsi cette approche apparaît comme plutôt conservative. En revanche, pour U-CC le FDR n'est pas contrôlé pour la majorité des scénarios, en particulier pour les valeurs élevées de β. La méthode LASSO-CC est aussi associée avec des grandes valeurs FDR estimés. En termes de nombre de signaux, U-CC et LASSO-CC génèrent un grand nombre de signaux en comparaison avec PS-CC.

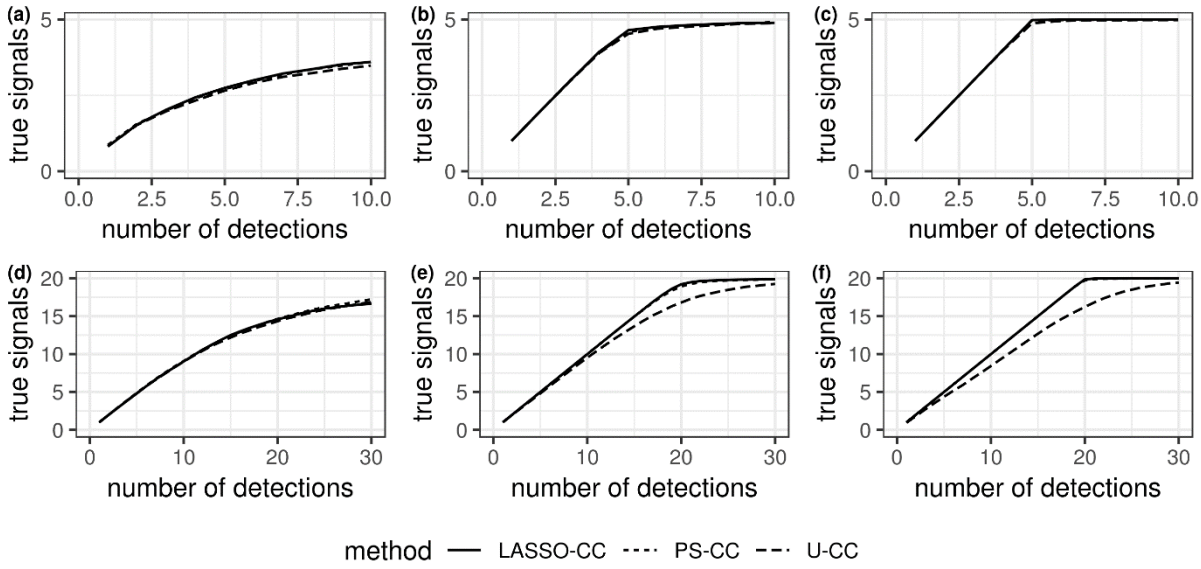

Figure 4 : Résultats des simulations pour le critère de nombre de vraies détections à un nombre total de détections fixé. En ligne, les valeurs de m_y égales à 5 en haut et 20 en bas. En colonne la magnitude de β, 0.5 à gauche, 1 au milieu et 1.5 à droite.

	β	0	0.5	0.5	1	1	1.5	1.5
	p_y	0	5	20	5	20	5	20
U-CC								
α =0.05	$\widehat{\mathrm{FDR}}$	0.22	0.05	0.09	0.04	0.58	0.15	0.77
	Sensibil-ité	-	0.14	0.40	0.80	0.99	0.99	1.00
	# signals	0.26	0.80	13.3	4.25	53.23	6.54	88.75
α =0.15	$\widehat{\mathrm{FDR}}$	0.82	0.21	0.17	0.12	0.68	0.23	0.80
	Sensitiv-ity	-	0.33	0.53	0.89	1.00	0.99	1.00
	# signals	2.14	2.46	14.26	5.40	70.41	8.11	99.52
LASSO-CC								
	$\widehat{\mathrm{FDR}}$	1	0.72	0.45	0.69	0.45	0.69	0.44
	Sensibil-ité	-	0.80	0.85	0.99	1.00	1.00	1.00
	# signals	9.29	15.98	32.17	18.68	37.2	18.55	37.06
PS-CC								
α =0.05	$\widehat{\mathrm{FDR}}$	0.01	0.04	0.03	0.01	0.01	0.01	0.02
	Sensibil-ité	-	0.17	0.24	0.61	0.83	0.96	0.99
	# signals	0.02	0.93	5.04	3.06	16.66	4.83	20.11
α =0.15	$\widehat{\mathrm{FDR}}$	0.05	0.08	0.07	0.01	0.04	0.01	0.07
	Sensibil-ité	-	0.24	0.39	0.75	0.91	0.99	0.99
	# signals	0.05	1.43	8.48	3.82	19.07	4.99	21.35

Tableau 2 : Résultats des simulations pour le deuxième critère évaluant la capacité des méthodes à effectivement contrôler le FDR ainsi que la sensibilité associée en cherchant à contrôler le FDR. Les résultats concernant le FDR pour le LASSO-CC sont donnés à titre indicatif.

3.3 ETUDE EN CAS REEL APPLIQUEE A L'INFARCTUS DU MYOCARDE

3.3.1 Matériels et méthodes

Le jeu de données étudié consiste en un ensemble de patients issu de l'EGB qui ont contracté un infarctus du myocarde pendant la période allant du 1er janvier 2010 jusqu'au 31 décembre 2016, pour lesquels nous avons un historique de remboursement de 13 mois minimum avant l'occurrence de l'infarctus du myocarde. Pour les individus ayant contracté plusieurs infarctus durant la période de suivi, seul le premier est considéré. Nous obtenons un total de 4099 individus cas et des informations relatives aux remboursements pour 741 médicaments durant la période d'étude.

Comme pour l'étude de simulations, seuls les médicaments sont considérés comme covariables. Certains sont évalués tandis que d'autres servent seulement comme facteurs de confusion potentiels. Comme certains médicaments protecteurs de l'infarctus du myocarde peuvent être prescrits juste avant l'occurrence de l'effet indésirable, la sélection de médicaments protecteurs est possible. Il s'agit d'un biais d'indication et pour essayer de le contrôler nous excluons les médicaments relatifs au système cardiovasculaire, ayant la lettre C comme première lettre de la classification ATC (drugs (Anatomical, therapeutic and chemical WHO class (WHO Collaborating Centre for Drug Statistics Methodology 2019)). En se basant sur la base de connaissance SIDER (Kuhn et al. 2016), nous excluons également les

médicaments utilisés pour traiter l'angine de poitrine ou pour gérer les événements cardiovasculaires aigus. Ainsi 132 médicaments sont exclus de l'évaluation mais peuvent être pris en compte par les méthode PS-CC et LASSO-CC.

Les trois méthodes ont été appliquées au jeu de données pour le schéma d'étude cas-croisé défini avec 6 périodes témoins de 30 jours séparées par 30 jours de washout, c'est-à-dire une période non utilisée dans l'étude pour éviter l'autocorrélation entre périodes. Pour les méthodes U-CC et PS-CC le seuil de détection pour la procédure de contrôle du FDR est fixé à 0.15.

Deux experts pharmacologues ont étudié tous les signaux générés par au moins une méthode au regard de leur pertinence sur le plan pharmacologique à savoir s'il existe un mécanisme biologique permettant de justifier leurs statuts de signaux. Ils ont également indiqué si les associations détectées pouvaient être assimilées à un biais d'indication. Cette évaluation est effectuée à l'aveugle c'est-à-dire que les experts ne savaient pas quelles méthodes avaient généré les signaux. En premier lieu les experts ont travaillé chacun seul puis se sont réunis afin de discuter les éventuels cas de désaccord. Chaque signal potentiel est ainsi considéré comme pharmacologiquement plausible ou non et potentiellement sujet à un biais d'indication ou non. Un jeu de référence a ainsi été créé pour évaluer l'efficacité des méthodes proposées.

3.3.2 Résultats

Les trois méthodes ont généré un total de 68 signaux. La Figure 5 montre le nombre de signaux généré par chaque méthode : 66 par la méthode U-CC, 43 par PS-CC et 33 par LASSO-CC et 30 signaux communs aux trois méthodes. Tous les signaux générés par PS-CC sont aussi détectés par U-CC. Deux signaux sont détectés uniquement LASSO-CC et 22 signaux uniquement par U-CC. Comme dans le cas de l'étude de simulations, le nombre de signaux retenus par U-CC est plus important que pour PS-CC. Au niveau de la pertinence pharmacologique, 15 signaux communs aux trois méthodes sont jugés pertinents. Le LASSO-CC détecte un signal pharmacologiquement pertinent supplémentaire tandis que U-CC en ajoute 5. En proportion, 51% de signaux sont jugés pertinents pour PS-CC, 41% par U-CC, et 48% par LASSO-CC.

La Figure 6 montre les signaux jugés potentiellement générés par un biais d'indication. Cette proportion est comparable pour les trois méthodes (76%, 73% et 77% pour U-CC, LASSO-CC et and PS-CC respectivement).

Un total de 8 signaux (soit 12% des signaux générés) sont considérés comme pertinents pharmacologiquement et potentiellement générés sans soupçon de biais d'indication (voir Tableau 3). En se focalisant sur ces signaux considérés comme prioritaires, la proportion de détection de signaux prioritaires est de 16% pour PS-CC, 12% pour U-CC et 12% pour LASSO-CC.

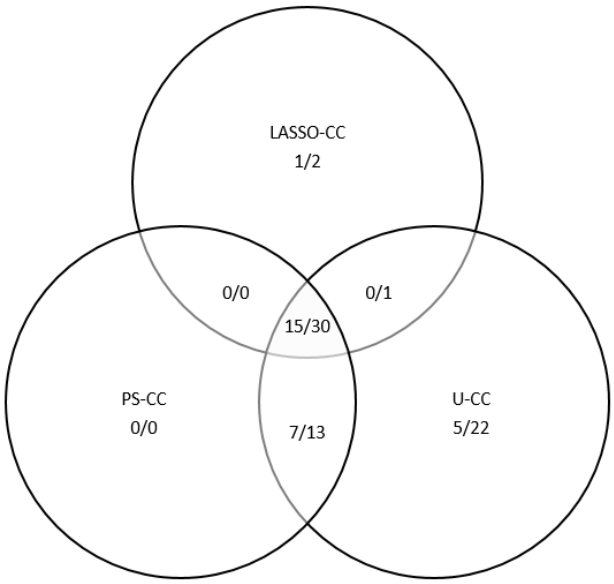

Figure 5 : Diagramme de Venn représentant l'ensemble des signaux jugés pharmacologiquement pertinents (à gauche) et le nombre total de signaux généré par les 3 méthodes.

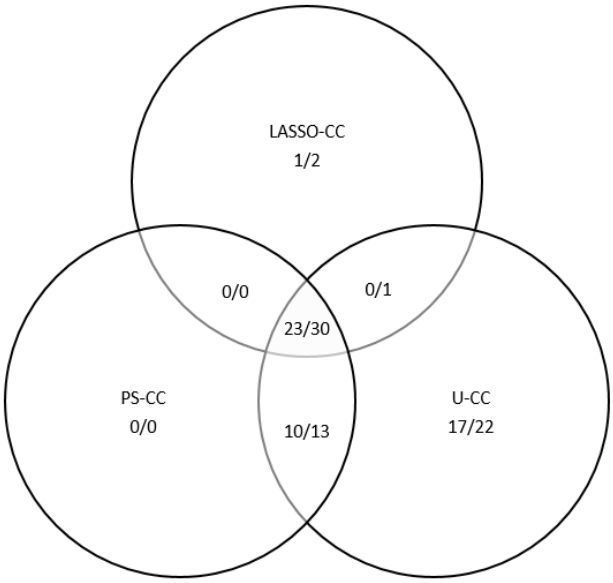

Figure 6 : Nombre de signaux générés avec suspicion de bais d'indication (à gauche) et nombre total de signaux (à droite) générés par les trois méthodes.

Drug	Pharmacological relevance	U-CC	PS-CC	LASSO-CC
Alimemazine	Known to be cardiotoxic and arrhythmogenic	✓		
Domperidone	Known to be cardiotoxic and arrhythmogenic	✓	✓	✓
Fexofenadine	Can cause tachycardia and palpitations. In patients at risk, it could be trigger AMI	✓	✓	
Hydroxyzine	Known to be cardiotoxic and arrhythmogenic	✓	✓	✓
Ibuprofen	Risk of AMI confirmed at high doses	✓	✓	✓
Metoclopramide	Antidopaminergic and serotoninergic on 5-HT4 (like cisapride)	✓	✓	
Metopimazine	Known to be arrhythmogenic	✓	✓	✓
Oxymetazoline	Alpha 1 receptor agonist, vasoconstrictor	✓	✓	

Tableau 3 : Dans le cadre de l'étude sur l'infarctus du myocarde, signaux jugés pertinents pharmacologiquement et potentiellement sans soupçon de biais d'indication

3.4 DISCUSSION

3.4.1 Discussion méthodologique

Dans cet axe de travail, nous avons proposé une approche basé sur l'intégration d'un score de propension dans un schéma d'étude cas-croisé, la construction du score s'appuyant sur l'utilisation de l'algorithme HDPS pour une selection automatisée de facteurs de confusion potentiels. Pour ajuster sur le score de propension nous avons assimilé le cas-croisé au schéma d'étude cas-témoins et éliminé la période cas de l'estimation du score de propension. Notre étude de simulations démontre que cette approche constitue une alternative solide aux approches de régression logistique conditionnelle avec pénalité LASSO pour le critère de sensibilité à nombre de signaux fixé. De plus nos simulations montrent qu'il est potentiellement possible de contrôler le FDR même si ce contrôle semble rendre la méthode peu sensible. L'absence de contrôle du FDR pour la méthode univariée U-CC montre l'importance de prendre en compte les corrélations entre expositions et donc justifie l'emploi de méthodes LASSO ou basées sur le score de propension. Dans nos simulations, plus la magnitude de l'effet traitement est forte plus il est difficile d'empêcher l'approche U-CC de sélectionner des médicaments non associés mais corrélés avec les vrais prédicteurs de l'effet indésirable. Dans les scénarios les plus extrêmes ($\beta = 1.5$), à la fois les résultats à nombre de détections fixé et le contrôle du FDR par U-CC sont affectés par l'absence de contrôle sur les facteurs de confusion potentiels. Dans les scenarios où la vraie

valeur de β est plus faible, les résultats à nombre de détections fixé restent similaires aux autres approches mais le contrôle du FDR est affecté.

Dans l'étude de simulations, l'approche basée sur le score de propension en grande dimension (PS-CC) réduit le nombre total de signaux générés en comparaison avec l'approche U-CC. L'étude sur données réelles confirme cette tendance. Ainsi, U-CC génère un plus grand nombre de signaux au total ; le nombre de signaux jugés pertinents par les experts est plus important mais cela se fait au prix d'un grand nombre de faux positifs supplémentaires. En routine, cela correspond à un moins bon niveau d'efficacité et à plus de temps perdu par les experts pour réaliser le travail d'évaluation.

L'approche LASSO-CC en revanche ne présente pas le même comportement sur données simulées et sur l'étude réelle. En effet, on observe que cette approche sélectionne plus de signaux sur l'étude de simulations par rapport à PS-CC alors que dans l'étude sur l'infarctus du myocarde elle sélectionne moins de signaux. Cette inconsistance peut être liée au fait que dans l'étude de simulation pour des raisons de temps de calcul seuls les médicaments les plus prévalents sont inclus. Dans l'étude en cas réel PS-CC et LASSO-CC ont fourni les meilleures performances en termes de détection des signaux jugés prioritaires (pertinents pharmacologiquement et potentiellement sans biais d'indication) et en termes de valeur positive prédictive, ce qui pouvait être attendu compte tenu de la prise en compte de la

polyexposition médicamenteuse par ces méthodes. En revanche un résultat plus surprenant est que les approches PS-CC et LASSO-CC présentent des résultats similaires alors que les mécanismes utilisés pour gérer la grande dimension sont très différents.

Dans ce travail, nous avons effectué plusieurs choix pour la méthodologie d'application du score de propension au schéma d'étude cas-croisé. Dans notre étude sur données réelles nous avons utilisé seulement les expositions médicamenteuses mais, à terme, d'autres variables pourraient être utilisées comme les consultations chez les médecins (actes médicaux). Nous avons également utilisé le score de propension en grande dimension pour filtrer les variables instrumentales et nous avons fixé le nombre de facteurs de confusion inclus dans le modèle de score de propension à la valeur $U = 50$. En pratique, le nombre de variables sélectionnées par le modèle d'arbres n'a jamais atteint ce maximum (données non montrées). D'autres algorithmes pour calculer le score de propension pourraient également être utilisés comme les forêts aléatoires (Cutler, Cutler, et Stevens 2012) ou le perceptron multicouche. Une limite potentielle de ces algorithmes d'apprentissage statistique est qu'ils peuvent nécessiter plus de données pour l'apprentissage que les approches des modèles linéaires généralisés avec pénalisation LASSO. Par ailleurs, nous avons choisi d'ajuster sur le score de propension alors que d'autres approches basées sur l'appariement et la pondération sont possibles et pourraient être étudiées dans le cadre du schéma d'étude cas-croisé dans de prochains travaux.

Nous avons identifié plusieurs limites théoriques et des questions méthodologiques pour l'approche PS-CC. La première est l'utilisation ou non de la période cas pour construire le score de propension. Dans notre travail nous avons choisi d'inclure cette période pour la sélection de variables à inclure dans le score de propension via l'algorithme HDPS puis nous l'avons exclu pour l'estimation du score afin de corriger un biais dû à la surreprésentation de la période cas (typiquement observé dans les schémas d'étude cas-témoins). Néanmoins cette approche d'exclusion de la période cas peut ne pas tout à fait éliminer le biais de surreprésentation comme le montre les travaux de Månsson *et al.* (Månsson et al. 2007). Par ailleurs l'utilisation du cas-croisé ne semble pas adaptée pour l'estimation d'effets causaux. Ainsi notre approche est limitée à la détection de signal et la recherche d'expositions prioritaires qui devraient par la suite être confirmées par d'autres études pharmaco épidémiologiques.

3.4.2 Evaluation pharmacologique

Cette partie résume l'évaluation pharmacologique effectuée par les deux experts pharmacologues.

Concernant la pertinence pharmacologique des signaux détectés, il est d'abord important de mettre en avant le fait que toutes les méthodes en détection de signal pour la pharmacovigilance sont affectées par le grand nombre de signaux faux positifs. Les signaux issus des bases médico administratives sont souvent sujets au biais d'indication. Un exemple dans ce travail est donné par les signaux

générés par les traitements prescrits pour la broncho-pneumopathie chronique obstructive (BPCO). Même si des mécanismes pourraient être identifiés, les patients souffrant de cette maladie présentent un risque d'infarctus du myocarde accru indépendamment du traitement. D'autres signaux détectés notamment par U-CC sont considérés comme détectés à cause du biais d'indication, comme les inhibiteurs de la pompe à protons (IPP) qui peuvent être prescrit dans le cadre de troubles dyspeptiques. Ces derniers peuvent constituer des signes avant-coureurs de l'infarctus du myocarde, de plus les IPP peuvent être coprescrits avec des médicaments à risque comme les anti-inflammatoires non-stéroïdiens pour prendre en charge la toxicité gastrointestinal. Quelques autres médicaments comme l'oxazepam et l'alprazolam sont détectés mais ne semblent ni liés à l'infarctus du myocarde ni à un biais d'indication.

Plus de 10% des signaux détectés sont considérés comme très pertinents. Ces signaux incluent trois antihistaminiques (alimemazine, hydroxyzine, and fexofenadine) connus pour leur potentiel arythmogène et qui peuvent potentiellement déclencher un infarctus du myocarde chez des patients à risque. Il y aussi trois médicaments utilisés pour contrôler les nausées qui sont connus également pour leur potentiel arythmogène : metopimazine, domperidone, et metoclopramide. Des études ont déjà suggéré des risques induits par domperidone et metoclopramide tandis que metopimazine pourrait être sujet d'un effet classe potentiel même s'il existe moins d'éléments à ce sujet. L'oxymetazoline est aussi considéré comme signal très

pertinent parce que c'est un agoniste des récepteurs alpha-adrénergiques utilisé comme vasoconstricteur, et pourrait ainsi augmenter le risque d'infarctus du myocarde. Enfin l'ibuprofène est un anti inflammatoire non-stéroïdien pour lequel il existe de nombreux éléments suggérant un risque d'infarctus du myocarde, surtout en cas d'utilisation à haute dose.

Notre étude démontre qu'il est difficile de prendre en compte le biais d'indication qui semble avoir une grande influence sur les résultats. De notre point de vue, la prise en compte de ce biais et le développement de méthodologies pouvant limiter son influence constituerait un axe de recherche principal pour l'utilisation des bases de données médico-administratives.

Dans notre étude nous avons utilisé une base de connaissance (SIDER) afin de pouvoir opérer un premier filtre. Néanmoins, s'il est faisable de constituer une liste de médicaments comprenant l'ensemble des traitements pour l'infarctus du myocarde, il est difficile de constituer une liste de médicaments pour l'ensemble des traitements de tous les symptômes et événements précurseurs de l'infarctus du myocarde.

Au final, il est fortement probable que les bases médico-administratives seront de plus en plus utilisées pour la détection de signal en pharmacovigilance. Néanmoins, l'utilisation de telles bases nécessite également des développements méthodologiques importants pour profiter de la richesse des données disponibles mais

également du manque de certaines informations pouvant être cruciales.

3.5 CONCLUSION ET PERSPECTIVES

En conclusion, l'approche proposée basée autour du score de propension et du schéma d'étude case-croisé obtient des résultats intéressants en détection de signal. Cette approche PS-CC se situe au niveau des approches cas croisé avec pénalité LASSO proposées précédemment pour la détection de signal. L'emploi du score de propension permet également de réduire sensiblement les temps de calcul.

Ainsi, cette nouvelle approche peut s'ajouter aux méthodes utilisées dans la détection de signal sur les bases médico-administratives. Une extension à la méthode de la série de cas est également possible, toujours dans le cadre de la détection de signal. En effet, la structure de données du cas-croisés n'entre pas en compte dans l'estimation du score de propension. Comme PS-CC est basée sur une analogie avec l'utilisation d'un schéma d'étude cas-témoins, cette approche peut également être utilisée dans le contexte d'un schéma cas-témoins pour la détection de signal en dehors du cadre du cas-croisé.

En revanche, cette approche ne peut pas s'appliquer à des études pharmacologiques ou plus généralement à des études

cherchant à mesurer un effet causal. En effet, dans ce nouveau cadre, l'estimation de la magnitude de l'effet est importante et l'approche PS-CC proposée dans ce travail est probablement biaisée et ne conviendrait donc pas.

Par ailleurs, notre étude se base sur une forme paramétrique fixée qui est celle de régression logistique conditionnelle et qui permet uniquement l'estimation d'un odds-ratio sous la prise en compte d'hypothèses paramétriques fortes. Une estimation d'un risque absolu n'est donc pas possible. Ainsi, cette approche ne se situe donc pas dans le cadre des approches d'apprentissage ciblé ou de g-computation qui sont construites pour relâcher les hypothèses paramétriques et utiliser des modèles non-paramétriques. Les approches d'apprentissage ciblé et de g-computation seront étudiées par la suite dans un autre contexte dans ce manuscrit.

Néanmoins malgré ces limitations, l'utilisation de cette forme paramétrique permet l'estimation et la prise en compte de variables de confusion non mesurées et statiques, situation dans laquelle les approches causales ne peuvent pas correctement fonctionner car une hypothèse de l'inférence causale est l'absence de facteurs de confusion non mesurés.

4 APPROCHES NON PARAMETRIQUES CAUSALES POUR LA DETECTION DE SIGNAL EN GRANDE DIMENSION APPLIQUEES A LA PHARMACOVIGILANCE SUR BASES DE DONNEES DE NOTIFICATIONS SPONTANEES

Les bases de notifications spontanées agrègent l'ensemble des notifications émises essentiellement par des professionnels de santé. En termes de structure de données, ces bases étaient initialement plutôt représentées par des tables de contingence. Ces tables de contingence sont constituées des expositions médicamenteuses en ligne et des différents effets indésirables en colonnes. Ainsi cette structure de données ne prend pas en compte les notifications spontanées présentant plusieurs expositions. Dès lors, une structure de données alternative a été proposée basée sur les matrices creuses.

Les matrices creuses sont des structures de données adaptées aux matrices comprenant essentiellement des valeurs nulles. Ainsi seuls les éléments non nuls de ces matrices sont représentés en mémoire ainsi que leurs positions dans ces matrices.

Dès lors, la structure de données basée sur les matrices creuses est composée de deux matrices, une matrice d'expositions médicamenteuses et une matrice des effets indésirables. La première matrice comporte en ligne les identifiants propres à chaque notification et en colonne les expositions médicamenteuses. La deuxième matrice comporte en ligne les identifiants propres à chaque notification et en colonne les événements indésirables observés pour ces notifications.

A partir de cette nouvelle structure de données, plusieurs modèles ont été proposés notamment basés sur la régression logistique avec emploi d'un *a priori* dans un cadre bayésien ou d'une

régularisation pour gérer la grande dimension. Ces méthodes ont pour objectif de sélectionner les variables d'expositions médicamenteuses statistiquement associées avec un événement indésirable unique en considérant les cas d'autres événements indésirables comme témoins. Dans ce cadre, un premier travail important a été proposé par Caster *et al.* (Caster, O. et al. 2010) dans un cadre bayésien. Plus récemment, des adaptations des méthodes de régressions pénalisées ont été proposées (Ahmed, Pariente, et Tubert-Bitter 2018; Courtois et al. 2018; Courtois, Tubert-Bitter, et Ahmed 2021). Ces approches, dans un cadre fréquentiste, ont notamment été développées pour améliorer la détection dans le cadre d'expositions rares et d'événements indésirables rares mais également pour permettre de choisir un seuil de détection plus adapté que le seuil induit par la validation croisée. En effet, l'emploi de la validation croisée pour choisir un seuil de détection ajustant au mieux le modèle avait tendance à détecter trop de faux positifs.

A notre connaissance, l'utilisation de méthodes non paramétriques associées à des mesures d'importance de variables n'a pas été explorée dans le cadre de la détection de signal en pharmacovigilance sur notifications spontanées. Ces approches pourraient pourtant s'avérer intéressantes car plus flexibles et pouvant prendre en compte des interactions médicamenteuses ou des effets spécifiques à une sous-population.

L'approche paramétrique se base sur la définition de

paramètres servant à la fois à l'ajustement et dans notre cas à la génération de signal, l'interprétation n'étant pas l'objectif. En revanche, dans le cadre non paramétrique, le paramètre générant (ou non) le signal est dissocié des paramètres utilisés pour ajuster le modèle d'apprentissage statistique aux données.

Dans le cadre de la détection de signal non paramétrique l'objectif est donc d'établir à la fois un modèle f pour ajuster les données ainsi qu'une manière d'obtenir des paramètres grâce à une fonction prenant f comme entrée et donnant une valeur dans un espace de $\mathbb{R}^p$ où p est le nombre de paramètres d'intérêt c'est-à-dire dans notre travail le nombre de médicaments étudiés.

Pour tester l'intérêt potentiel de méthodes non paramétriques pour la détection de signal sur bases de notifications spontanées, nous proposons d'utiliser comme modèle statistique non paramétrique f une approche dérivée des arbres de régression boostés. Cette approche permet la prédiction de l'occurrence d'un effet indésirable défini précédemment comme étant d'intérêt et noté Y.

Les méthodes étudiées conjointement avec les arbres de régression boostés pour l'estimation dans ce travail sont la g-computation et l'apprentissage ciblé en incluant la validation croisée ainsi qu'un éventuel enrichissement du modèle. Aucune hypothèse causale n'enrichit le modèle qui reste ainsi un pur modèle de détection de signal sans interprétation causale. Une telle utilisation des méthodes d'inférence causale pour établir des mesures d'importance

des variables non paramétriques est proposée par Van der Laan et Rose (Laan et Rose 2011). Cette approche n'a en revanche pas encore été étudiée en détection de signal pour la pharmacovigilance.

Dans cette partie, nous développerons une approche non paramétrique pour la pharmacovigilance sur bases de notifications spontanées intégrant des outils de la causalité. Une étude de simulation sera menée puis une application sur données réelles suivies d'une discussion.

4.1 PROPOSITION D'UNE APPROCHE NON PARAMETRIQUE

Les approches non-paramétriques flexibles utilisées sur des bases de données en grande dimension, parcimonieuses et fortement déséquilibrées pour la variable réponse et certaines expositions (beaucoup plus de non-traités que de traités et beaucoup plus de cas que de témoins) peuvent se heurter à des difficultés. Par exemple, certaines variables d'exposition pourtant fortement associées à la variable réponse risquent d'être totalement absentes du modèle, celui-ci favorisant les expositions les plus présentes dans la base.

L'inclusion systématique de la variable traitement dans le modèle est une propriété que nous jugeons importante, même si l'exposition n'est pas associée à Y, pour pouvoir construire des intervalles de confiance en utilisant le bootstrap. En effet, si dans certains cas de rééchantillonnage, la variable traitement est exclue du

modèle d'ensemble d'arbres on obtient une couverture de l'intervalle de confiance faussée car les estimations de l'odds-ratio correspondent alors à un mélange d'une loi continue et d'une loi de Dirac en 1. Cela se traduit en pratique par des intervalles de confiance trop petits.

Bien qu'en théorie, la prise en compte du score de propension telle qu'effectuée avec le TMLE permet d'apporter une solution à ce problème, nous avons choisi d'inclure systématiquement la variable traitement d'intérêt dans nos modèles d'apprentissage statistique.

Le modèle linéaire généralisé sans pénalisation permet l'inclusion de toutes les variables mais ce modèle n'est pas adapté à la grande dimension. Nous proposons dans ce chapitre une adaptation basée sur l'utilisation conjointe d'un modèle linéaire et d'un modèle non paramétrique par arbres de régression afin de « forcer » l'inclusion de la variable traintement (grâce à la partie linéaire) et ce tout en gardant un ajustement non paramétrique sur les éventuelles coexpositions.

On considère n réalisations indépendantes et identiquement distribuées notées $O_i = (X_{ij}, Y_i, X_{i(-j)})$, $i = 1, \dots, n$. On considère un modèle statistique non paramétrique $\mathcal{M}$ liant Y_i à $(X_{ij}, X_{i(-j)})$,

Nous considérons le modèle statistique semi-paramétrique suivant :

$$\mathcal{M}^* = \{M_0(X_j) + M_1(\boldsymbol{X}_{-j}),\ M_0 \in L, M_1\}.$$

L'ensemble des modèles $\mathcal{M}^*$ est construit de manières à ce que M_0 appartient à l'ensemble des modèles linéaires L et que M_1 soit non paramétrique. Nous proposons ainsi de considérer l'ensemble des modèles statistiques construits en deux parties, une partie non paramétrique ne dépendant pas du traitement d'intérêt et une partie linéaire dépendant uniquement du traitement d'intérêt.

Le but initial de cette décomposition est d'assurer la sélection de la variable d'intérêt pour le calcul des intervalles de confiance de la g-computation par bootstrap. Néanmoins, une telle décomposition du modèle reste flexible pour prédire Y et peut donc être utilisée pour estimer la variable réponse dans le cadre d'une approche TMLE. Cette approche que nous nommons NPCPV-TMLE peut aussi être considérée comme une forme de Super Learner associé au TMLE.

Ainsi, la procédure du TMLE décrite au chapitre 2 peut être étendue au cas où le modèle statistique utilisé pour l'ajustement au jeu de données est construit de manière semi-paramétrique et non totalement non paramétrique. L'intérêt d'une telle approche est dans un cadre de grande dimension avec des expositions rares et des événements rares de miser sur la partie paramétrique pour « forcer » l'inclusion de la variable d'intérêt dans le modèle. De plus, cette partie linéaire devrait permettre une efficacité plus importante à taille d'échantillon fixé dans le cadre de variables d'expositions rares. En revanche avec cette approche, les éventuelles interactions modifiant la survenue de l'effet indésirable entre le traitement d'intérêt et les

covariables ne sont pas prises en compte sauf si elles sont définies explicitement par ajout d'un terme supplémentaire dans la partie linéaire. Les interactions entre les covariables (traitement d'intérêt exclu) sont elles automatiquement prises en compte par le modèle non paramétrique.

Ainsi le modèle dont le paramètre est estimé de cette manière ne correspond pas tout à fait au modèle du TMLE standard et est un peu moins flexible. Il est néanmoins plus flexible que les approches purement linéaires.

On note que dans le cas où le paramètre d'intérêt est additif seule la partie non paramétrique nécessite un réajustement lors de la « targeted step ». En effet, s'il existe une décomposition de la projection Ψ de M vers l'espace du paramètre d'intérêt dépendant de (M_0, M_1), telle que $M_0 \rightarrow \Psi(M_0, M_1)$ est linéaire et $M_0(X_j)$ est linéaire en X_j comme décrit plus haut alors la partie linéaire peut être estimée après la partie non linéaire par un estimateur non paramétrique sans biais utilisant la distribution empirique. Cet estimateur étant sans biais, il n'y a donc pas d'intérêt à « cibler » la partie dépendant de M_0.

Concrètement, pour le calcul d'un ATT (« Average Treatement effect on the Treated »), par exemple, cette approche consiste à sommer les différences entre les valeurs prises par Y chez les traités et des prédictions estimées de manière non paramétrique sans prendre en compte la variable d'intérêt pour ces traités. Les prédictions estimées sans prendre en compte la variable d'intérêt étant

préalablement réajustées par une étape ciblée afin de constituer un « risque de base » correspondant au risque associé aux observations en excluant le traitement.

En détection de signal sur bases médico-administratives les paramètres utilisés pour la prise de décision sont souvent assimilables à des risques relatifs ou des odds-ratio. Ainsi, on restera dans cet axe sur l'estimation de ce type de paramètres et non sur l'estimation d'un paramètre additif. Cepdendant, dans le cadre d'un paramètre d'intérêt basé sur un odds-ratio, l'approche présentée au paragraphe précédent ne s'applique pas. Il faut donc réajuster l'estimation initiale en opérant un réajustement, c'est-à-dire une étape ciblée en considérant la somme des deux composantes comme modèle.

Lors de l'évaluation d'un grand nombre de médicaments, l'implémentation de cette approche consiste à effectuer une boucle sur l'ensemble des expositions d'intérêt. Pour chaque exposition d'intérêt, il faut estimer un modèle non paramétrique comportant toutes les variables autres que l'exposition d'intérêt. Puis il faut estimer la composante linéaire dépendant du traitement. Puis enfin, effectuer une étape ciblée qui consiste à estimer un score de propension puis à effectuer un réajustement.

Pour chaque traitement, l'ajustement de deux modèles non paramétriques, l'un pour le modèle issu de l'ensemble M_1, l'autre pour le score de propension pour chaque exposition d'intérêt est potentiellement lourd en temps de calcul. Nous proposons une

heuristique pour le passage à la grande dimension basée sur un algorithme glouton.

L'algorithme glouton que nous proposons se base sur la construction d'un unique modèle de gradient boosting prédisant la variable réponse. Ce modèle est par la suite modifié de manière légère pour chacune des taches nécessaires à l'estimation de chaque effet traitement. Deux types de modifications sont effectuées. La première consiste à éliminer la variable traitement d'intérêt du modèle en effectuant un élagage des branches basées sur cette variable. Une fois cette première modification effectuée, les poids associés à chacune des feuilles du modèle élagué sont réajustés afin de prédire au mieux soit la variable réponse en excluant la variable d'intérêt, soit l'exposition à la variable d'intérêt (pour calculer le score de propension) toujours en excluant la variable d'intérêt.

Ainsi la structure des arbres du modèle élagué, c'est-à-dire l'ensemble des choix de variables autres que la variable d'intérêt partitionnant le jeu de données, est conservée. Pour effectuer le réajustement de chaque feuille pour le calcul du score de propension, la valeur minimisant l'entropie croisée au sein de chaque feuille en considérant comme variable réponse la variable exposition d'intérêt est calculée puis est assigné à l'ensemble des individus instanciant la feuille.

La même procédure est appliquée pour estimer la variable réponse en éliminant la variable d'intérêt. L'approche consistant à

diviser le problème de construction des arbres en deux parties (structure de l'arbre puis estimation du poids des feuilles) a été proposée dans le cadre des forêts aléatoires par Athey et al. (Athey et Imbens 2016) sous le nom « d'inférence honnête ». La Figure 7 illustre la mise en place du TMLE en réajustant les poids du modèle f élagué de toutes les parties dépendantes de X_j.

En revanche la décomposition du modèle ne correspondant pas à une décomposition sur une famille libre de fonctions, l'unicité de la décomposition n'est donc pas assurée. La méthode proposée consiste juste à permettre l'inclusion de la variable traitement dans le modèle. Un autre avantage est de permettre l'exclusion de variables instrumentales dans le modèle du score de propension.

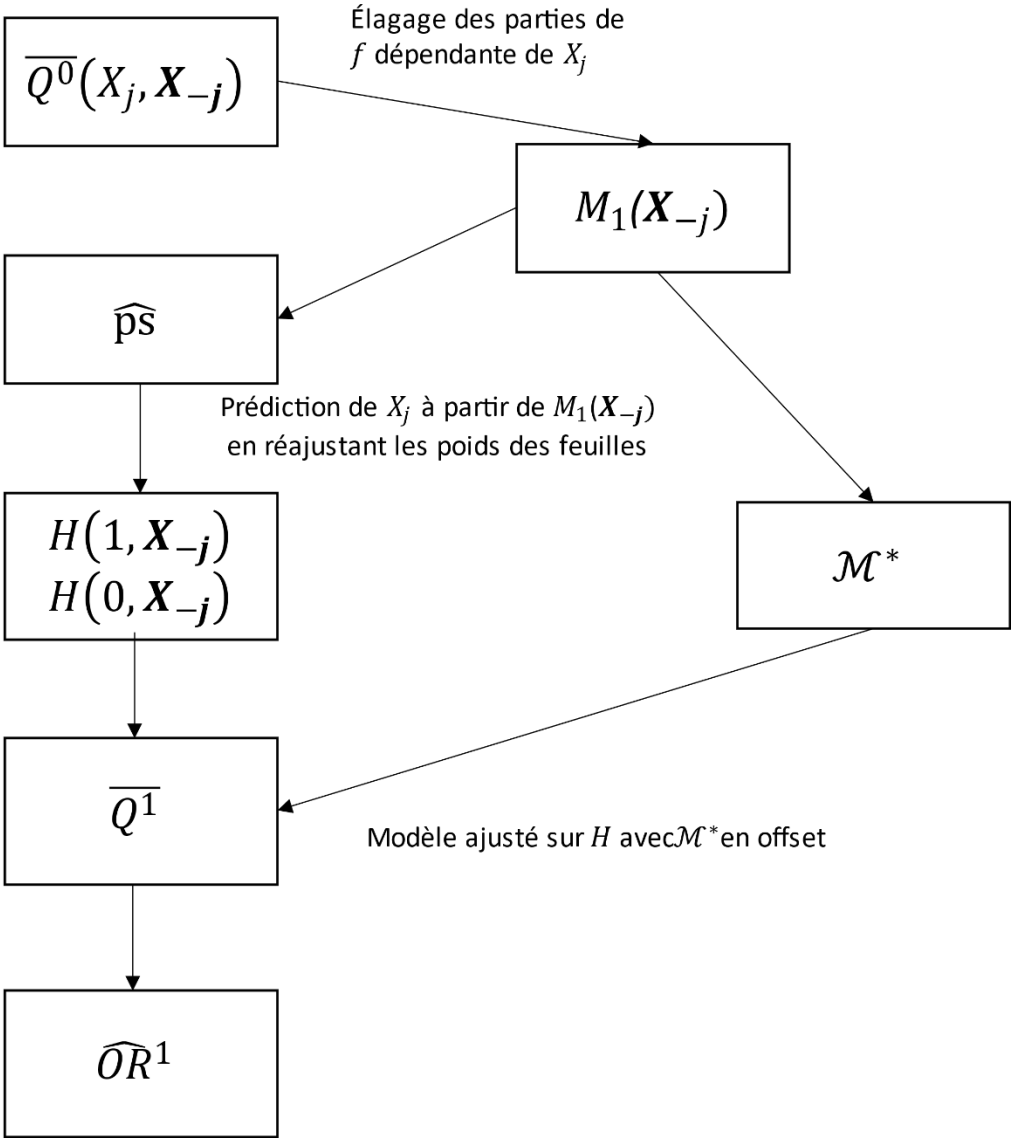

Figure 7 : Schéma illustrant les différentes étapes de la procédure d'estimation du TMLE avec score de propension adaptif.

4.1.1 Elimination des variables instrumentales dans le score de propension

Dans cette sous partie nous décrivons comment la construction du score de propension à partir d'une modification du modèle prédisant la variable réponse telle que présentée dans la partie précédente peut filtrer les variables instrumentales. L'enjeu de filtrer les variables instrumentales (variables liées à l'exposition d'intérêt mais pas à la variable réponse) est considérée comme importante quand un grand nombre de facteurs de confusion ou de covariables sont présentes dans la base de données. L'emploi de ces variables dans l'estimation du score de propension doit être évité (Austin 2011).

Récemment, plusieurs approches dont l'algorithme HDPS utilisé dans chapitre 3 ont été proposées pour effectuer une pré-sélection de variables pour le score de propension dans le cadre de la grande dimension (S. Schneeweiss et al. 2009; Franklin et al. 2015) y compris une heuristique analogue à celle proposée basée sur l'utilisation de l'adaptive LASSO (Shortreed et Ertefaie 2017). La principale idée mise en avant par ces approches est de retirer des variables explicatives du score de propension les variables non sélectionnées pour prédire Y. Dans notre cas, ayant entraîné un modèle pour prédire Y par les arbres de régressions boostés, nous pouvons réutiliser les structures d'arbres utilisées pour prédire Y précédant l'apparition de la variable d'intérêt dans la forêt pour prédire le score de propension.

L'heuristique suivante est développée :

1. Repérer le premier arbre pour lequel la variable traitement d'intérêt X_j intervient dans la prédiction de Y.
2. Reprendre la structure des arbres précédents et recalculer les valeurs associées aux feuilles pour prédire X_j. Obtenir ainsi le score de propension.

Cette heuristique est également rapide à mettre en œuvre. On peut en outre élaguer les arbres sélectionnés afin de gagner en robustesse.

Cette approche constitue une forme de « gradient boosting » adaptatif et permet de réutiliser les structures d'arbres générées pour le modèle entraîné sur la variable réponse. En effet, comme la structure des arbres est construite pour estimer la variable réponse, seules les combinaisons de variables associées à un changement significatif pour la prédiction de Y sont sélectionnées pour pouvoir ensuite construire le score de propension. Au regard du temps de calcul pour tous les scores de propension, le screening nécessaire pour construire les arbres à chaque nœud devient inutile et seules les feuilles sont modifiées ce qui est très peu coûteux.

4.1.2 Heuristiques pour la sélection de modèles par pondération des individus

On peut considérer que plusieurs modèles concurrents au

modèle *f* initial sont acceptables et performent de manière quasiment équivalente en termes de prédiction de la variable Y. Nous proposons dans l'étude en cas réel d'appliquer des contraintes supplémentaires à f afin d'opérer un choix de modèle à la fois performant pour la prédiction et faisant sens d'un point de vue pharmacologique.

Nous proposons d'introduire lors de l'estimation de f une pondération par l'inverse de la norme du vecteur d'exposition afin de favoriser les modèles les plus simples (basés principalement sur les notifications comportant peu de médicaments suspectés). Par ailleurs dans l'étape ciblée nous proposons de nous limiter aux notifications comportant moins de trois médicaments notifiés afin de ne pas dépendre de notifications rares et ambiguës (donc avec un score de propension très faible) mais comportant un nombre important de médicaments et donc une influence sur l'ensemble du vecteur des paramètres extrêmement importante.

4.2 ETUDE DE SIMULATIONS

4.2.1 Modèle de simulations

Les simulations sont effectuées en prenant la matrice X des expositions issue de données réelles constituant un sous-ensemble de la BNPV (cf partie application). Cette matrice X est réduite à un nombre n d'individus tiré aléatoirement parmi l'ensemble des notifications présentes dans la BNPV et un nombre p de traitements également tirés

aléatoirement parmi les traitements ayant au moins un nombre d'occurrences de 1000 dans la base source. Parmi les p traitements, un nombre de traitements aléatoirement sélectionnés k sont associés à la survenue de l'effet indésirable avec une magnitude de β. Une valeur d'intercept du modèle est également définie et est notée α. Ce paramètre peut être interprété comme la prévalence sans tenir compte des associations avec les variables d'expositions médicamenteuses de l'événement indésirable dans la base.

L'occurrence de l'événement indésirable Y est ensuite simulée à partir de la formule suivante :

$$q = \frac{1}{1 + e^{-\alpha - \boldsymbol{\beta X}'}},$$

$$Y \sim Bernoulli(q).$$

Les paramètres de simulation peuvent différer selon plusieurs scénarios : n le nombre d'individus pouvant être de 10 000 ou 50 000 et α peut être fixé à -3, -4, -5 ou -6. Un coefficient α faible induit une prévalence moins forte.

Les paramètres invariants sont le nombre de traitements p fixé à 100. Le nombre $k = 15$ traitements associés à l'événement d'intérêt et la force d'association pour ces k médicaments est égale à $\beta = 1$.

Pour chaque scénario, 500 répétitions sont effectuées et les résultats sont moyennés sur l'ensemble des simulations effectuées.

Le processus de simulations complet est résumé en Figure 8 et est répété pour chaque simulation.

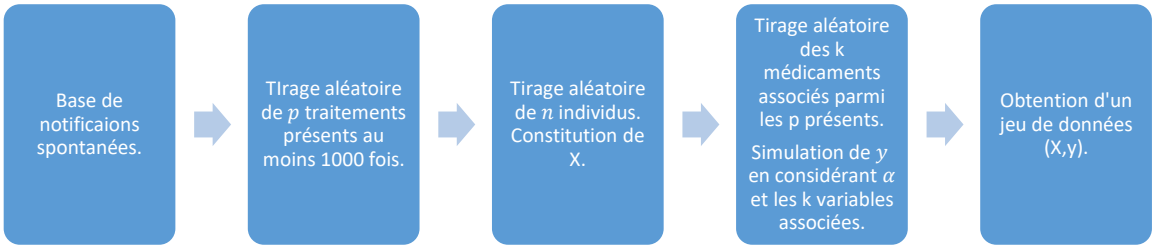

Figure 8: Processus de construction des jeux de données simulées.

4.2.2 Algorithmes évalués

Après la construction des simulations, le but est de mesurer la capacité des algorithmes de détection de signal à sélectionner les k variables associées avec Y parmi les p variables présentes dans X. Les algorithmes de détection de signal étudiés sont :

La méthode Class Imbalanced Subsampling LASSO (CISL) introduite par Ahmed *et al.* (Ahmed, Pariente, et Tubert-Bitter 2018).

Le LASSO avec validation croisée pour déterminer la magnitude de la pénalité avec 10 sous-partitions des données tel que présenté dans la partie 2.2. (LASSO-CV)

La g-computation en utilisant l'approche NPCPV comme modèle d'apprentissage statistique (basé sur les arbres de régression boostés avec intégration systématique de la variable traitement) et avec l'approche de Benjamini-Hochberg comme correction des tests multiples (NPCPV-GCOMP).

L'apprentissage ciblé en utilisant l'approche NPCPV comme modèle d'apprentissage statistique (basé sur les arbres de régression boostés avec intégration systématique de la variable traitement) et avec l'approche de Benjamini-Hochberg comme correction des tests multiples (NPCPV-TMLE).

4.2.3 Critères d'évaluation

Les résultats sont présentés

- à seuil de détection fixé
- en termes de classement (sans chercher à fixer un seuil de détection).

Le premier critère intégrant un seuil de détection est le critère prioritaire.

Pour les méthodes générant des p-values, à savoir la g-computation et le TMLE, le seuil de détection est fixé en termes de valeurs cibles du FDR. Pour les méthodes LASSO, celles-ci ne générant pas de p-values le seuil est défini en considérant la magnitude du coefficient, un coefficient strictement supérieur à 0 indiquant la

présence d'un signal. Pour CISL, le quantile à 20% est utilisé (Ahmed, Pariente, et Tubert-Bitter 2018).

En plus du taux observé de faux découvertes, la comparaison des approches à seuil de détection fixé se fait à partir d'un taux de détection des $k = 15$ traitements positivement associés à Y.

Pour le classement, on observe les courbes indiquant le nombre de faux positifs en fonction du nombre total de détections. Une telle courbe permet de distinguer la capacité de chacune des méthodes à trier les expositions médicamenteuses pour toutes valeurs de nombre de signaux générés totaux.

On effectue 500 répétitions pour chaque scénario et les valeurs présentées en section résultats sont moyennées sur l'ensemble de ces répétitions.

4.3 RESULTATS DE L'ETUDE DE SIMULATIONS

4.3.1 Résultats à seuil de détection fixé

Le Tableau 4 recense les résultats à seuil de détection fixé pour chaque scénario défini par le nombre d'individus n et la valeur de l'intercept α. Les méthodes utilisées sont CISL, LASSO NPCPV-GCOMP et NPCPV-TMLE avec seuil de FDR fixé à 25%.

	#vrais positifs	#faux positifs	$\widehat{FDR}$ %	Sensibilité %
n=10000, α=-3				
CISL	10.30	1.65	14	69
LASSO CV	13.37	10.36	44	89
NPCPV-GCOMP	10.42	1.54	13	69
NPCPV-TMLE	7.88	0.62	7	53
n=10000, α=-4				
CISL	6.18	1.38	18	41
LASSO- CV	10.24	8.79	46	68
NPCPV-GCOMP	6.38	1.30	17	43
NPCPV-TMLE	4.84	0.92	16	32
n=10000, α=-5				
CISL	3.53	1.60	31	24
LASSO-CV	5.83	6.35	52	39
NPCPV-GCOMP	3.98	2.30	37	27
NPCPV-TMLE	2.38	2.02	46	16
n=10000, α=-6				
CISL	1.87	1.57	46	12
LASSO-CV	2.21	3.12	59	15
NPCPV-GCOMP	3.28	3.66	53	22
NPCPV-TMLE	1.37	3.19	70	9
n=50000, α=-3				
CISL	14.90	1.60	10	99
LASSO-CV	14.98	10.48	41	100
NPCPV-GCOMP	14.88	1.60	10	99
NPCPV-TMLE	14.62	0.34	2	97
n=50000, α=-4				
CISL	13.38	1.26	9	89
LASSO-CV	14.64	10.18	41	98
NPCPV-GCOMP	14.06	2.08	13	94
NPCPV-TMLE	12.02	0.54	4	80
n=50000, α=-5				
CISL	9.80	1.20	11	65
LASSO-CV	13.14	9.96	43	88
NPCPV-GCOMP	10.46	1.28	11	70
NPCPV-TMLE	8.32	0.92	10	55
n=50000, α=-6				
CISL	4.96	1.06	18	33
LASSO-CV	8.78	8.46	49	59
NPCPV-GCOMP	6.22	2.08	25	41
NPCPV-TMLE	4.32	2.00	32	29

Tableau 4 : Résultats de l'étude de simulations en termes de nombres (#) de vrais et faux positifs, FDR et sensibilité. Pour les méthodes NPCPV, le seuil est fixé pour cibler un FDR de 0.25. Pour CISL, la quantité le quantile à 20%. Pour le LASSO, la validation croisée fournit le seuil de détection.

On observe sur le Tableau 4, que les résultats du LASSO avec validation croisée sont affectés par un nombre important de faux positifs dans tous les scénarios. La baisse du paramètre α se traduit par une baisse de performances de toutes les méthodes. Les scénarios $n = 10000, \alpha = -5$ et $n = 10000, \alpha = -6$, sont particulièrement difficiles pour toutes les méthodes.

En excluant ces scénarios, les méthodes NPCPV parviennent à un taux de fausses découvertes inférieures à la valeur ciblée de 0.25 (à l 'exception du scénario $n = 50000, \alpha = -6$ pour l'approche NPCPV-TMLE). CISL bien que n'intégrant pas une correction pour cibler un FDR obtient un taux de fausses découvertes inférieur à la valeur de 0.25 souhaité sur ces scénarios également.

En termes de pourcentage de découvertes le LASSO avec validation croisée est l'approche détectant le plus de signaux positivement associés mais avec un taux de fausses découvertes important. L'approche NPCPV avec g-computation sélectionne légèrement plus de signaux positivement associés que l'approche CISL qui elle-même sélectionne plus de signaux positivement associés que l'approche NPCPV-TMLE. L'approche NPCPV-TMLE est souvent très conservative c'est-à-dire que son taux de fausses découvertes est très

faible par rapport à la valeur ciblée de 0.25 et son taux de signaux positivement associés plus faible comparativement aux autres approches.

4.3.2 Résultats en termes de classement

Les Figure 9 et Figure 10 montrent le nombre de vrais positifs en ordonnée par rapport à un nombre de signaux fixés en abscisse. Pour les approches NPCPV les courbes représentent le nombre de vrais positifs obtenu en faisant varier le FDR ciblé de 0.01 à 0.25. A titre de comparaison les approches CISL et LASSO-CV qui ne peuvent pas produire une continuité du taux de vrais positifs en fonction d'un critère FDR cible sont représentés par des points. Ces figures montrent à la fois la capacité de classement des signaux prioritaires et les performances respectives des algorithmes quand l'expérimentateur choisit lui-même le seuil de détection pour les méthodes NPCPV. Plus le ratio nombre de signaux vrais positifs sur nombre total de signaux générés est important, plus la méthode de détection est efficace.

On observe dans la Figure 9 et la Figure 10 que le LASSO avec validation croisée détecte beaucoup plus de signaux que les autres approches. Cela se traduit par un nombre plus important de vrais positifs mais au prix d'un nombre également important de faux positifs.

Les approches CISL et NPCPV-GCOMP sont les approches qui semblent les plus efficaces dans tous les scénarios. La méthode NPCPV-TMLE est aussi efficace dans les scénarios avec une prévalence

de l'événement indésirable plus importante et avec $n = 50000$.

Pour les scénarios avec $n = 10000$, NPCPV-TMLE est conservative quand $\alpha = -3$ et $\alpha = -4$ avec un nombre de détections plus faible quand le FDR cible atteint 0.25. Dans les scénarios avec $\alpha = -5$ et $\alpha = -6$, cette approche apparait inférieure à CISL et NPCPV-GCOMP. Par ailleurs, dans ces scénarios, NPCPV-GCOMP apparaît plus efficace.

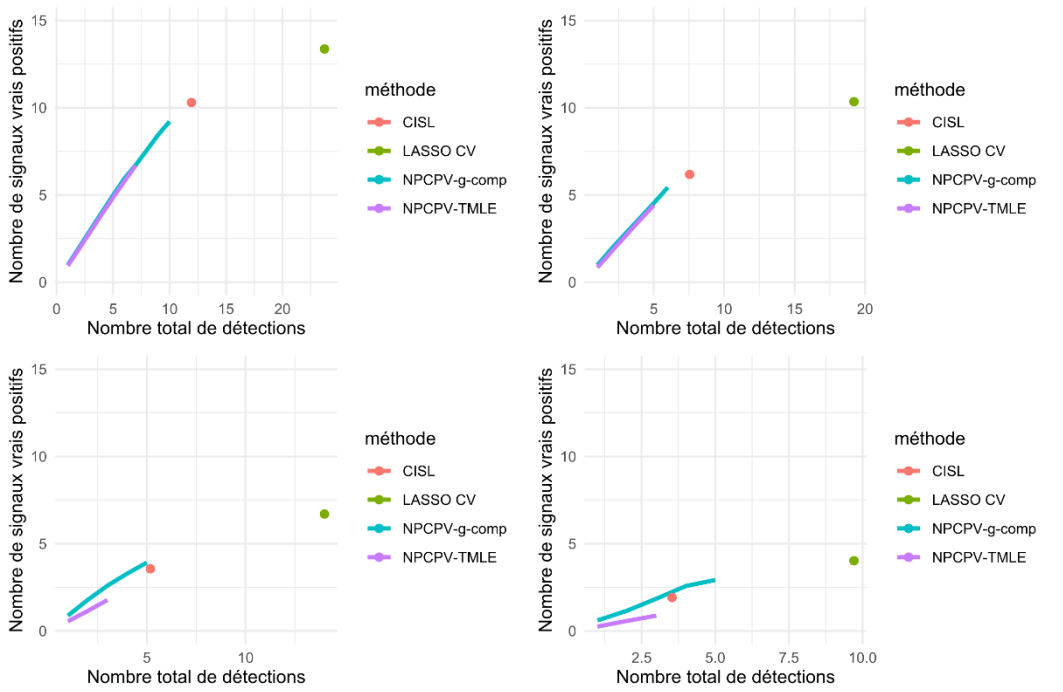

Figure 9 : Etude de simulations. Nombre de vrais positifs à nombre de détections fixé. Scénarios où $n = 10000$. *En haut à gauche (a)* $\alpha = -3$, *en haut à droite(b)* $\alpha = -4$, *en bas à gauche (c)* $\alpha = -5$ *et en bas à droite (d)* $\alpha = -6$. *Le FDR ciblé maximum est fixé à 0.25 pour les méthodes NPCPV et l'arrêt de la courbe est fixé quand la moitié des répétitions atteint cette valeur.*

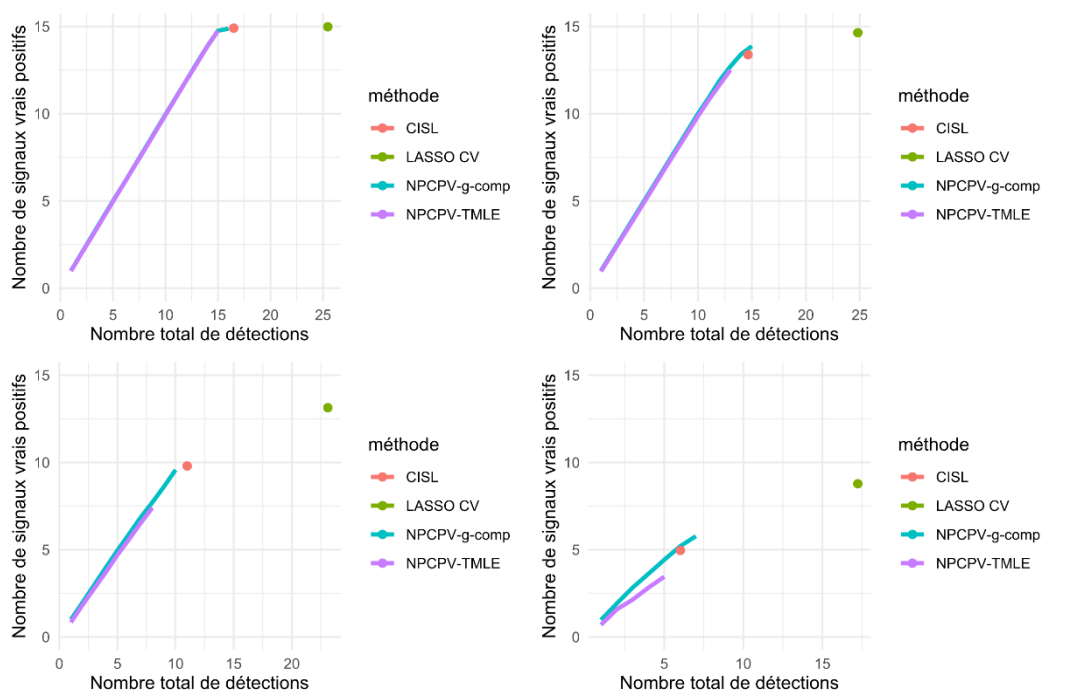

Figure 10 : Etude de simulations. Nombre de vrais positifs à nombre de détections fixé. Scénarios où $n = 50000$. En haut à gauche (a) $\alpha = -3$, en haut à droite(b) $\alpha = -4$, en bas à gauche (c) $\alpha = -5$ et en bas à droite (d) $\alpha = -6$. Le FDR ciblé maximum est fixé à 0.25 pour les méthodes NPCPV et l'arrêt de la courbe est fixé quand la moitié des répétitions atteint cette valeur.

4.3.3 Conclusion de l'étude de simulations

En conclusion, il apparaît que les approches non paramétriques issues des outils de la causalité sont capables de produire des mesures d'importance des variables utiles pour la détection de signal. L'emploi du score de propension pour les approches TMLE ne semble pas forcément utile dans un cadre de détection de signal en grande dimension au regard du temps de calcul important que demande cette méthode.

L'emploi d'un modèle comportant une part linéaire et une part non-linéaire semble, en revanche, un compromis efficace avec des données binaires fortement déséquilibrées (beaucoup de témoins peu de cas et beaucoup de non exposés peu d'exposés) comme en témoignent les résultats de NPCPV-GCOMP dans tous les scénarios.

4.4 APPLICATION SUR DONNEES REELLES

Nous avons mis en oeuvre les méthodes NPCPV-GCOMP, NPCPV-TMLE, CISL, et LASSO-CV sur une extraction de la base française de notifications spontanées (BNPV) couvrant la période 01/01/2000 – 29/12/2017. Cette extraction comprenait 452914 notifications impliquant 6617 types d'événements indésirables et 2378 types de médicaments.

Pour illustrer notre approche, nous nous sommes intéressés à

un événement indésirable particulier : les lésions hépatiques aiguës. Les lésions hépatiques aiguës dues à l'usage de médicaments (Drug induced liver injury (DILI)) constituent un effet indésirable potentiel de nombreux médicaments. Ces lésions peuvent entraîner des complications sérieuses chez les patients, une hospitalisation et dans les cas les plus graves un décès. Il est donc important de pouvoir effectuer un suivi de cet événement indésirable et trouver les médicaments associés dans une base de données de notifications spontanées. Il est à noter que cet événement indésirable correspond à une combinaison de plusieurs types d'événements indésirables établis par Antoine Pariente et Franscesco Salvo de l'unité Inserm U1219 (Bordeaux Population Health)(Courtois et al. 2018). Au total, 25187 notifications spontanées impliquaient un événement indésirable de type DILI, ce qui en fait un événement indésirable très fréquent.

On codera par, $Y_i = 1$ la présence d'au moins un événement indésirable DILI parmi les événements indésirables pour la notification i avec $i = 1, \dots, n$. On notera également $Y_i = 0$ l'absence de tout événement indésirable DILI dans les événements indésirables. On dénote X la matrice des expositions médicamenteuses associées à chaque notification.

On applique plusieurs prétraitements à l'ensemble des données (X, Y). Seuls les médicaments étant reportés au moins 10 fois, potentiellement conjointement avec d'autres médicaments, et 3 fois pour un événement de DILI peuvent être évalués. Ce prétraitement est

courant pour l'emploi de la régression logistique et limite notre étude à $p = 1136$.

4.4.1 Evaluation des méthodes

Nous cherchons à évaluer l'efficacité des différentes méthodes. Néanmoins, le statut réel de chaque association est généralement inconnu rendant l'évaluation difficile sans expertise médicale. Pour annoter les signaux comme intéressants ou non avérés, nous effectuons une comparaison des résultats obtenus par notre étude avec un jeu de données d'annotation produit par Chen (M. Chen et al. 2016). Le travail d'annotation effectué par Chen est effectué sans l'aide des notifications spontanées et est considéré comme référence pour évaluer les résultats fournis par les différents algorithmes. L'indépendance de ce travail avec la base de notifications spontanées de la BNPV permet d'écarter le risque que des données de notifications spontanées françaises aient pu être utilisées pour l'annotation. Appliqué à notre base de données, cet ensemble de référence contient 203 témoins négatifs (absence d'association avérée entre le médicament et DILI) et 133 témoins positifs (lien délétère avéré entre DILI et le médicament). Sur les 1136 médicaments finalement considérés, cet ensemble de référence contient 123 témoins négatifs et 119 témoins positifs.

Les différentes méthodes sont appliquées en considérant l'ensemble des 1136 médicaments. Les performances sont évaluées en termes de

- nombre de signaux générés en considérant l'ensemble des médicaments,
- du nombre de signaux dont le statut est connu (témoin positif ou négatif),
- du nombre de faux positifs parmi les signaux dont le statut est connu et
- de la proportion de faux positifs (FDP pour *false discovery proportion*) en se restreignant à l'ensemble de référence

4.4.2 Résultats

Les résultats obtenus par les méthodes NPCPV pour l'étude des lésions hépatiques aiguës sur la base de la BNPV sont données en Tableau 5. Les résultats peuvent être comparés avec l'étude basée sur les approches LASSO adaptives effectuées par Courtois *et al.* (Courtois, Tubert-Bitter, et Ahmed 2021) et reproduites dans ce même tableau.

On observe que le nombre de signaux générés par ces nouvelles approches est plutôt faible comparé aux approches LASSO, en particulier par rapport à l'approche LASSO-CV. Les taux de signaux faux positifs des méthode NPCPV sont proches de ceux de l'approche CISL. Les différences entre les résultats aux deux seuils de FDR ciblé sont faibles pour le TMLE mais plus important pour la g-computation.

Méthode	FDR ciblé	Nombre de signaux générés	Nombre de signaux connus	Signaux faux positifs	FDP
NPCPV-GCOMP	0.05	132	49	1	0.020
NPCPV-GCOMP	0.15	153	58	2	0.033
NPCPV-TMLE	0.05	131	50	1	0.020
NPCPV-TMLE	0.15	135	51	1	0.019
LASSO-CV		220	79	5	0.159
CISL		109	48	2	0.042

Tableau 5 : Résultats sur données réelles (BNPV, DILI) pour les méthodes NPCPV avec g-computation et TMLE. Les résultats sont présentés à deux seuils de FDR ciblé. Les résultats concernant CISL et LASSO-CV sont reproduits d'après l'article de Courtois et al (2021)

4.5 DISCUSSION

Au regard de l'étude de simulation, il semble difficile de déterminer une méthode dominante par rapport à toutes les autres pour l'ensemble des effets indésirables. L'approche CISL ainsi que les approches développées en se basant sur les outils de la causalité semblent néanmoins assez robustes car adéquates dans la plupart des scénarios étudiés. Ces approches sont néanmoins conservatives dans l'étude en cas réel pour les approches NPCPV et en cas réel et dans les simulations pour CISL. Au vu des résultats obtenus, l'utilisation du score de propension ne semble pas toujours nécessaire.

Notre étude de simulation semble indiquer que la prévalence de l'événement indésirable dans la base de données est un critère important pour choisir une méthode adéquate. En particulier, l'approche adaptive que nous proposons pour le score de propension semble inefficace comparativement à la g-computation dans les scénarios à faible prévalence de l'effet indésirable dans la base. En revanche, l'approche NPCPV-TMLE pourrait être efficace pour les effets indésirables fortement représentés dans la base de notifications spontanées, comme c'est le cas pour l'effet indésirable DILI étudié dans cette étude. Ces résultats illustrent le problème dû à l'écart à l'hypothèse de positivité qui affecte négativement les résultats des approches issues de l'inférence causale.

Une limite de notre étude de simulations est que le lien entre événement indésirable et expositions est basé sur un modèle linéaire avec lien logit. Cela avantage les méthodes présentant elles-mêmes au moins une composante linéaire. Les quatre approches exposées dans cette partie présentent cette caractéristique et donc il n'y a pas de comparaison faussée à cause de cette limitation dans nos simulations. Par ailleurs, le caractère binaire des expositions dans le cadre de la base de notifications spontanées devrait rendre les approches dérivées du modèle linéaire généralisé relativement efficaces comparativement à des approches non linéaires. Cela s'explique parce que dans le cadre de données binaires l'hypothèse de linéarité s'applique, ce qui justifie le choix de modèles linéaires avec lien logit.

La méthode NPCPV, pourrait être étendue à d'autres modèles que la régression logistique. Par exemple, le modèle linéaire pourrait être employé dans le cadre d'une réponse réelle. Pour mesurer un effet non linéaire, on pourrait employer une fonction spline. En revanche, la décomposition effectuée, ne reposant pas sur une famille libre de fonctions, l'identifiabilité de cette décomposition n'est pas garantie. Cela limite ce modèle à la prédiction, car il peut être comparé à d'autres modèles par l'emploi d'un jeu de données test, mais n'offre pas de garantie d'unicité pour l'interprétation de cette décomposition.

L'emploi de l'apprentissage ciblé à la suite de ce modèle pour pouvoir estimer l'effet traitement est ainsi justifié dans les études où les variables ne sont pas trop déséquilibrées comme le montrent nos simulations. Dans le cas de variables d'expositions binaires et déséquilibrées, il est connu que les méthodes basées sur le score de propension ont tendance à ne pas être efficaces, car l'inclusion de l'inverse du score de propension amène une grande variance pour l'estimateur. Quand un jeu de données présente peu d'occurrences d'individus exposés et présentant l'effet indésirable, l'estimation devient difficile comme le montre nos simulations. De plus nous nous sommes limités à des médicaments notifiés de manière assez importante et ainsi les occurrences rares d'exposition n'ont pas été étudiés. La méthode CISL, développée essentiellement pour traiter ces situations, semble alors assez adéquate.

4.6 CONCLUSION

En conclusion, les travaux conduits dans ce chapitre démontrent, par une étude de simulations, la faisabilité et l'utilité de l'emploi de méthodes d'inférence causale pour la détection de signal en pharmacovigilance sur bases de données spontanées.

Ces approches ont pour avantage de permettre l'utilisation de tests statistiques et ainsi d'utiliser un critère statistique de contrôle d'erreur comme le FDR pour choisir un seuil de détection. Néanmoins, la nature souvent binaire et déséquilibrée (beaucoup plus de témoins que de cas et beaucoup plus de non exposés que d'exposés) du jeu de données, ne semble pas permettre une efficacité supérieure des méthodes basées sur l'utilisation du score de propension telles que le TMLE.

Même si elle n'intègre pas de tests statistiques, la méthode CISL permet d'obtenir un taux de faux positifs souvent plus faible ou similaire à ceux des méthodes intégrant un seuil basé sur les tests et le FDR. Les méthodes NPCPV intègrent le choix d'un seuil de détection fixé par un FDR ciblé, et semblent en mesure de maintenir un taux de faux positifs inférieur à ce seuil, ceci avec les limites du modèle de simulations. En revanche, la validation croisée dans le cadre du LASSO a tendance à sur-détecter confirmant ce qui a déjà pu être observé (Ahmed, Pariente, et Tubert-Bitter 2018; Courtois, Tubert-Bitter, et Ahmed 2021), ce qui implique un large nombre de faux positifs. Cette

méthode utilisée en routine pourrait ainsi engendrer un travail très important d'évaluation de signaux par les experts en pharmacovigilance.

L'ensemble des méthodes proposées permet également de fournir de nouvelles approches flexibles pour la détection de signal, approches qui peuvent être étendues sur d'autres sources de données, notamment les données médico-administratives.

5 DISCUSSION GENERALE

Les travaux présentés dans ce manuscrit avaient pour objectifs d'explorer la combinaison de méthodes d'inférence causale et de méthodes d'apprentissage statistique avec l'utilisation de bases de données de grande dimension pour la détection de signal en pharmacovigilance. Globalement, les études conduites ont montré que cette combinaison d'approches avait un potentiel pour la détection de signal même si méthodologiquement il est difficile de corriger les nombreux biais induits par la nature des données disponibles. Le fait d'avoir basé nos études de simulation sur la structure réelle de corrélation des médicaments a permis une comparaison relativement réaliste des méthodes, que ce soit pour le chapitre 3 ou le chapitre 4.

Sur les bases médico-administratives le manque d'information sur les caractéristiques des patients telles que le poids, le statut tabagique ou la consommation d'alcool rend difficile la détection de signal. Par ailleurs, en utilisant ces bases, il y a un risque de causalité inverse ou de biais d'indication qui est important. Le biais d'indication semble être une limite importante lors de l'utilisation de ces bases pour des objectifs de détection de signal comme illustré par notre étude sur l'infarctus du myocarde.

Sur les bases de notifications spontanées la sous-déclaration, d'ampleur inconnue, variable dans le temps et en fonction des événements indésirables et/ou des médicaments considérés, peut affecter la qualité de la détection de signal. Notre étude sur les lésions hépatiques aiguës a donné des résultats corrects. Néanmoins,

s'agissant d'un ensemble de référence avec nécessairement des associations déjà connues, cette étude ne permet pas d'évaluer le potentiel de détection de nouveaux signaux.

Pour prendre en compte et tenter de surpasser ces difficultés, nous avons effectué des développements méthodologiques à la fois pour le traitement des bases de données médico-administratives et sur bases de données de notifications spontanées. Nous avons, dans ce cadre, également fait un grand usage de méthodes d'apprentissage statistique basées sur des ensembles d'arbres de décision. En effet, nous avons trouvé que ces méthodes sont assez complémentaires avec les méthodes d'inférence causale fournissant des modèles flexibles qui peuvent être par la suite utilisés par une étape *g*-computation pour l'estimation des paramètres d'intérêt.

Par ailleurs, l'utilisation de modèles d'apprentissage statistique permet également l'estimation du score de propension même quand les interactions entre variables impliquées sont complexes. En revanche, l'utilisation de ces modèles seuls rendrait difficile l'interprétation et donc la prise de décision qui dans le cadre de la détection de signal consiste en une décision binaire (signal ou absence de signal). Nous avons choisi de ne pas utiliser les mesures d'importance des variables classiquement construites pour ce type de modèles mais d'utiliser plutôt des méthodes d'inférence causale pour les estimations de l'importance des variables.

Une approche d'inférence causale non-étudiée dans cette thèse

est la méthode des variables instrumentales qui pourrait potentiellement être utile dans la détection de signal sur données médico administrative en prenant par exemple la préférence du médecin comme instrument (Brookhart et al. 2006). Néanmoins cette mise en œuvre déjà complexe pour les études ciblées le seraient encore plus pour une détection de signal automatisée. Nous n'avons pas également utilisé des algorithmes de découverte causale comme l'algorithme PC car la complexité de la mise en œuvre de cet algorithme est lourde et d'autre part nous ne sommes intéressés que par les associations entre la variable cible et les expositions médicamenteuses.

L'utilisation croissante de bases de données regroupant des données de santé individuelles pour la pharmacovigilance, et plus généralement pour la santé publique, appelle à la construction de nouveaux algorithmes et de nouvelles méthodes pour tenir compte de la richesse potentielle de ces bases mais également du manque parfois d'informations importantes. L'utilisation de ces bases pour la détection de signal en routine de manière systématique présente de nombreuses difficultés, car il semble qu'il n'existe pas de solution algorithmique et statistique permettant de s'appliquer à l'ensemble des problématiques de détection de signal en pharmacovigilance sur l'ensemble des bases de données. Cela se traduit dans ce travail de thèse, par l'utilisation de deux méthodes issues certes des outils statistiques de l'inférence causale mais très différentes conceptuellement, à savoir les méthodes autocontrôlées sur les bases médico-administratives et l'approche

d'apprentissage ciblé à composante non paramétrique pour les bases de notifications spontanées.

Néanmoins, il faut noter que les concepts clés de l'inférence causale peuvent s'appliquer sur l'utilisation aussi bien des bases médico-administratives et des bases de notifications spontanées. Ainsi l'estimation et l'incorporation du score de propension, même dans un cadre biaisé comme dans l'étude effectuée dans le chapitre 3 de cette thèse, peut améliorer les résultats en termes de détection de signal. Une limite de cette approche est qu'elle ne peut pas être appliquée dans un cadre d'études pharmaco-épidémiologiques ni dans un cadre d'estimation d'effets causaux. L'approche issue de l'apprentissage ciblé abordée dans le chapitre 4 dépend également de l'estimation du score de propension et semble propice à une utilisation plus large.

En ce qui concerne l'estimation du score de propension, l'utilisation de méthodes automatiques pour éliminer les variables instrumentales telles qu'initialement proposée par Schneeweiss *et al.* (Sebastian Schneeweiss et al. 2009) semble importante pour l'utilisation de bases de données en grande dimension à des fin de détection car il est impossible d'effectuer un filtrage « à la main » pour l'ensemble des médicaments criblés. Construire un score de propension pour chaque exposition d'intérêt est, en revanche, très coûteux.

Une approche non-explorée dans cette thèse serait d'utiliser un algorithme modélisant la loi jointe de l'ensemble des expositions. Cela

permettrait par la suite de reconstruire les lois de probabilités conditionnelles en prenant en compte cette estimation de la loi jointe et des estimations des lois marginales. Une difficulté de cette approche est de parvenir à trouver un algorithme, plus complexe que le simple classifieur bayésien naïf, pouvant s'appliquer à cette tâche en grande dimension. Des méthodes basées sur les réseaux de neurones telles que l'approche dite de « l'autoencoder variationel » pourraient être testées.

On peut aussi imaginer pouvoir utiliser un algorithme construit pour prédire plusieurs effets indésirables conjointement. Une difficulté serait alors de regrouper les variables indiquant des effets indésirables similaires en une seule variable pour pouvoir travailler avec des variables réponses moins déséquilibrées en termes de ratio nombre de cas-nombre de témoins. Ainsi l'utilisation de pré traitement des bases de notifications spontanées ou médico-administrative en utilisant des bases de connaissances serait également importante pour cet objectif.

L'importance des bases de connaissance, la complexité grandissante des algorithmes proposés et l'importance des experts en pharmacovigilances pour annoter les signaux générés montrent que la pharmacovigilance pour être efficace ne peut se faire que dans un cadre hautement multi disciplinaire.

6 BIBLIOGRAPHIE

Ahmed, Ismaïl, Antoine Pariente, et Pascale Tubert-Bitter. 2018. « Class-Imbalanced Subsampling Lasso Algorithm for Discovering Adverse Drug Reactions ». *Statistical Methods in Medical Research* 27 (3): 785-97. https://doi.org/10.1177/0962280216643116.

Arnaud, Mickael, Bernard Bégaud, Frantz Thiessard, Quentin Jarrion, Julien Bezin, Antoine Pariente, et Francesco Salvo. 2018. « An Automated System Combining Safety Signal Detection and Prioritization from Healthcare Databases: A Pilot Study ». *Drug Safety* 41 (4): 377-87. https://doi.org/10.1007/s40264-017-0618-y.

Athey, S., et G. Imbens. 2016. « Recursive partitioning for heterogeneous causal effects. » *Proc Natl Acad Sci U S A* 113 (27): 7353-60. https://doi.org/10.1073/pnas.1510489113.

Austin, Peter C. 2011. « An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies ». *Multivariate Behavioral Research* 46 (3): 399-424. https://doi.org/10.1080/00273171.2011.568786.

Bate, A., M. Lindquist, I. R. Edwards, S. Olsson, R. Orre, A. Lansner, et R. M. De Freitas. 1998. « A Bayesian Neural Network Method for Adverse Drug Reaction Signal Generation ». *European Journal of Clinical Pharmacology* 54 (4): 315-21. https://doi.org/10.1007/s002280050466.

Benjamini, Yoav, et Yosef Hochberg. 1995. « Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing ». *Journal of the Royal Statistical Society. Series B (Methodological)* 57 (1): 289-300.

Breiman, Leo. 2001. « Random Forests ». *Machine Learning* 45 (1): 5-32. https://doi.org/10.1023/A:1010933404324.

Brookhart, M. A., P. S. Wang, D. H. Solomon, et S. Schneeweiss. 2006. « Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. » *Epidemiology* 17 (3): 268-75. https://doi.org/10.1097/01.ede.0000193606.58671.c5.

Caster, O., Norén, G.N., Madigan, D., et Bate, A. 2010. « Large-scale regression-based pattern discovery: The example of screening the WHO global drug safety database - - - Statistical Analysis and

Data Mining: The ASA Data Science Journal - Wiley Online Library ». 2010. https://onlinelibrary.wiley.com/doi/abs/10.1002/sam.10078.

Caster, Ola, G. Niklas Norén, David Madigan, et Andrew Bate. 2010. « Large-Scale Regression-Based Pattern Discovery: The Example of Screening the WHO Global Drug Safety Database: Large-Scale Regression-Based Pattern Discovery ». *Statistical Analysis and Data Mining: The ASA Data Science Journal* 3 (4): 197-208. https://doi.org/10.1002/sam.10078.

Chen, J., et Z. Chen. 2008. « Extended Bayesian information criteria for model selection with large model spaces ». *Biometrika* 95 (3): 759-71. https://doi.org/10.1093/biomet/asn034.

Chen, Minjun, Ayako Suzuki, Shraddha Thakkar, Ke Yu, Chuchu Hu, et Weida Tong. 2016. « DILIrank: The Largest Reference Drug List Ranked by the Risk for Developing Drug-Induced Liver Injury in Humans ». *Drug Discovery Today* 21 (4): 648-53. https://doi.org/10.1016/j.drudis.2016.02.015.

Chen, Tianqi, et Carlos Guestrin. 2016. « XGBoost: A Scalable Tree Boosting System ». In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-94. KDD '16. New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785.

Christodoulou, Evangelia, Jie Ma, Gary S. Collins, Ewout W. Steyerberg, Jan Y. Verbakel, et Ben Van Calster. 2019. « A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models ». *Journal of Clinical Epidemiology* 110 (juin): 12-22. https://doi.org/10.1016/j.jclinepi.2019.02.004.

Cole, Stephen R., et Miguel A. Hernán. 2008. « Constructing Inverse Probability Weights for Marginal Structural Models ». *American Journal of Epidemiology* 168 (6): 656-64. https://doi.org/10.1093/aje/kwn164.

Courtois, Émeline, Antoine Pariente, Francesco Salvo, Étienne Volatier, Pascale Tubert-Bitter, et Ismaïl Ahmed. 2018. « Propensity Score-Based Approaches in High Dimension for Pharmacovigilance Signal Detection: an Empirical Comparison on the French Spontaneous Reporting Database ». *Frontiers in Pharmacology* 9 (septembre). https://doi.org/10.3389/fphar.2018.01010.

Courtois, Émeline, Pascale Tubert-Bitter, et Ismaïl Ahmed. 2021. « New adaptive lasso approaches for variable selection in automated pharmacovigilance signal detection ». *BMC Medical Research Methodology* 21 (1): 271. https://doi.org/10.1186/s12874-021-01450-3.

Cowling, Thomas E., David A. Cromwell, Alexis Bellot, Linda D. Sharples, et Jan van der Meulen. 2021. « Logistic regression and machine learning predicted patient mortality from large sets of diagnosis codes comparably ». *Journal of Clinical Epidemiology* 133 (mai): 43-52. https://doi.org/10.1016/j.jclinepi.2020.12.018.

Cutler, Adele, D. Richard Cutler, et John R. Stevens. 2012. « Random Forests ». In *Ensemble Machine Learning: Methods and Applications*, édité par Cha Zhang et Yunqian Ma, 157-75. Boston, MA: Springer US. https://doi.org/10.1007/978-1-4419-9326-7_5.

Demailly, Romain, Sylvie Escolano, Françoise Haramburu, Pascale Tubert-Bitter, et Ismaïl Ahmed. 2020. « Identifying Drugs Inducing Prematurity by Mining Claims Data with High-Dimensional Confounder Score Strategies ». *Drug Safety* 43 (6): 549-59. https://doi.org/10.1007/s40264-020-00916-5.

DuMouchel, William. 1999. « Bayesian Data Mining in Large Frequency Tables, with an Application to the FDA Spontaneous Reporting System ». *The American Statistician* 53 (3): 177-90. https://doi.org/10.2307/2686093.

Džeroski, Saso, et Bernard Ženko. 2004. « Is Combining Classifiers with Stacking Better than Selecting the Best One? » *Machine Learning* 54 (3): 255-73. https://doi.org/10.1023/B:MACH.0000015881.36452.6e.

EMA. 2018. « EudraVigilance ». Text. European Medicines Agency. 17 septembre 2018. https://www.ema.europa.eu/en/human-regulatory/research-development/pharmacovigilance/eudravigilance.

Evans, S. J., P. C. Waller, et S. Davis. 2001. « Use of Proportional Reporting Ratios (PRRs) for Signal Generation from Spontaneous Adverse Drug Reaction Reports ». *Pharmacoepidemiology and Drug Safety* 10 (6): 483-86. https://doi.org/10.1002/pds.677.

Farrington, C. P. 1995. « Relative Incidence Estimation from Case Series for Vaccine Safety Evaluation ». *Biometrics* 51 (1): 228-35. https://doi.org/10.2307/2533328.

Franklin, J. M., W. Eddings, R. J. Glynn, et S. Schneeweiss. 2015. « Regularized Regression Versus the High-Dimensional Propensity Score for Confounding Adjustment in Secondary Database Analyses. » *Am J Epidemiol* 182 (7): 651-59. https://doi.org/10.1093/aje/kwv108.

Friedman, Jerome H. 2001. « Greedy Function Approximation: A Gradient Boosting Machine. » *The Annals of Statistics* 29 (5): 1189-1232. https://doi.org/10.1214/aos/1013203451.

Hansen, B. B. 2008. « The Prognostic Analogue of the Propensity Score ». *Biometrika* 95 (2): 481-88. https://doi.org/10.1093/biomet/asn004.

Hardin, James W. 2005. « Generalized Estimating Equations (GEE) ». In *Encyclopedia of Statistics in Behavioral Science*, 39-42. American Cancer Society. https://doi.org/10.1002/0470013192.bsa250.

Hastie, Trevor, Jerome Friedman, et Robert Tibshirani. 2001. « Linear Methods for Classification ». In *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, édité par Trevor Hastie, Jerome Friedman, et Robert Tibshirani, 79-113. Springer Series in Statistics. New York, NY: Springer. https://doi.org/10.1007/978-0-387-21606-5_4.

Hauben, Manfred, Jeffrey K. Aronson, et Robin E. Ferner. 2016. « Evidence of Misclassification of Drug-Event Associations Classified as Gold Standard "Negative Controls" by the Observational Medical Outcomes Partnership (OMOP) ». *Drug Safety* 39 (5): 421-32. https://doi.org/10.1007/s40264-016-0392-2.

Hazell, Lorna, et Saad A. W. Shakir. 2006. « Under-Reporting of Adverse Drug Reactions : A Systematic Review ». *Drug Safety* 29 (5): 385-96. https://doi.org/10.2165/00002018-200629050-00003.

Hill, Austin Bradford. 1965. « The Environment and Disease: Association or Causation? » *Proceedings of the Royal Society of Medicine* 58 (5): 295-300.

Kammer, Michael, Daniela Dunkler, Stefan Michiels, et Georg Heinze. 2022. « Evaluating methods for Lasso selective inference in biomedical research: a comparative simulation study ». *BMC Medical Research Methodology* 22 (1): 206. https://doi.org/10.1186/s12874-022-01681-y.

Kuhn, Michael, Ivica Letunic, Lars Juhl Jensen, et Peer Bork. 2016. « The SIDER Database of Drugs and Side Effects ». *Nucleic Acids Research* 44 (D1): D1075-1079. https://doi.org/10.1093/nar/gkv1075.

Kulldorff, M., I. Dashevsky, T. R. Avery, A. K. Chan, R. L. Davis, D. Graham, R. Platt, et al. 2013. « Drug safety data mining with a tree-based scan statistic. » *Pharmacoepidemiol Drug Saf* 22 (5): 517-23. https://doi.org/10.1002/pds.3423.

Kumamaru, Hiraku, Sebastian Schneeweiss, Robert J. Glynn, Soko Setoguchi, et Joshua J. Gagne. 2016. « Dimension Reduction and Shrinkage Methods for High Dimensional Disease Risk Scores in Historical Data ». *Emerging Themes in Epidemiology* 13: 5. https://doi.org/10.1186/s12982-016-0047-x.

« La surveillance renforcée des médicaments ». 2023. ANSM. 20 novembre 2023. https://ansm.sante.fr/page/la-surveillance-renforcee-des-medicaments.

Laan, Mark J. van der, et Sherri Rose. 2011. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Science & Business Media.

Levy, Jonathan. 2018. « An Easy Implementation of CV-TMLE ». arXiv. http://arxiv.org/abs/1811.04573.

Li, Fan, Laine E Thomas, et Fan Li. 2019. « Addressing Extreme Propensity Scores via the Overlap Weights ». *American Journal of Epidemiology* 188 (1): 250-57. https://doi.org/10.1093/aje/kwy201.

Lindquist, Marie. 2008. « VigiBase, the WHO Global ICSR Database System: Basic Facts ». *Drug Information Journal* 42 (5): 409-19. https://doi.org/10.1177/009286150804200501.

Lockhart, Richard, Jonathan Taylor, Ryan J. Tibshirani, et Robert Tibshirani. 2014. « A significance test for the lasso ». *The Annals of Statistics* 42 (2): 413-68.

Lundberg, Scott M, et Su-In Lee. 2017. « A Unified Approach to Interpreting Model Predictions ». In *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

Luu, Maxime, Eric Benzenine, Alan Barkun, Muriel Doret, Christophe Michiels, Thibault Degand, Catherine Quantin, et Marc Bardou.

2019. « Safety of First Year Vaccination in Children Born to Mothers with Inflammatory Bowel Disease and Exposed in Utero to Anti-TNFα Agents: A French Nationwide Population-Based Cohort ». *Alimentary Pharmacology & Therapeutics* 50 (11-12): 1181-88. https://doi.org/10.1111/apt.15504.

Maclure, M. 1991. « The Case-Crossover Design: A Method for Studying Transient Effects on the Risk of Acute Events ». *American Journal of Epidemiology* 133 (2): 144-53.

Maignen, Francois, Manfred Hauben, Eric Hung, Lionel Van Holle, et Jean-Michel Dogne. 2014. « Assessing the Extent and Impact of the Masking Effect of Disproportionality Analyses on Two Spontaneous Reporting Systems Databases ». *Pharmacoepidemiology and Drug Safety* 23 (2): 195-207. https://doi.org/10.1002/pds.3529.

Månsson, Roger, Marshall M. Joffe, Wenguang Sun, et Sean Hennessy. 2007. « On the Estimation and Use of Propensity Scores in Case-Control and Case-Cohort Studies ». *American Journal of Epidemiology* 166 (3): 332-39. https://doi.org/10.1093/aje/kwm069.

Meinshausen, Nicolai, et Peter Bühlmann. 2010. « Stability selection ». *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72 (4): 417-73. https://doi.org/10.1111/rssb.2010.72.issue-4.

Mittleman, Murray A., et Elizabeth Mostofsky. 2014. « Exchangeability in the Case-Crossover Design ». *International Journal of Epidemiology* 43 (5): 1645-55. https://doi.org/10.1093/ije/dyu081.

Morel, Maryan, Emmanuel Bacry, Stéphane Gaïffas, Agathe Guilloux, et Fanny Leroy. 2017. « ConvSCCS: convolutional self-controlled case series model for lagged adverse event detection ».

Moride, Y., F. Haramburu, A. A. Requejo, et B. Bégaud. 1997. « Under-Reporting of Adverse Drug Reactions in General Practice ». *British Journal of Clinical Pharmacology* 43 (2): 177-81. https://doi.org/10.1046/j.1365-2125.1997.05417.x.

Natekin, Alexey, et Alois Knoll. 2013. « Gradient boosting machines, a tutorial ». *Frontiers in Neurorobotics* 7. https://www.frontiersin.org/articles/10.3389/fnbot.2013.00021.

Niculescu-Mizil, Alexandru, et Rich Caruana. 2005. « Obtaining calibrated probabilities from boosting ». In *Proceedings of the*

Twenty-First Conference on Uncertainty in Artificial Intelligence, 413-20. UAI'05. Arlington, Virginia, USA: AUAI Press.

Pariente, Antoine, Fleur Gregoire, Annie Fourrier-Reglat, Françoise Haramburu, et Nicholas Moore. 2007. « Impact of Safety Alerts on Measures of Disproportionality in Spontaneous Reporting Databases: The Notoriety Bias ». *Drug Safety* 30 (10): 891-98. https://doi.org/10.2165/00002018-200730100-00007.

Pearl, Judea. 1993. « [Bayesian Analysis in Expert Systems]: Comment: Graphical Models, Causality and Intervention ». *Statistical Science* 8 (3): 266-69.

Puijenbroek, Eugéne P. van, Andrew Bate, Hubert G. M. Leufkens, Marie Lindquist, Roland Orre, et Antoine C. G. Egberts. 2002. « A Comparison of Measures of Disproportionality for Signal Detection in Spontaneous Reporting Systems for Adverse Drug Reactions ». *Pharmacoepidemiology and Drug Safety* 11 (1): 3-10. https://doi.org/10.1002/pds.668.

Robins, James. 1986. « A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect ». *Mathematical Modelling* 7 (9): 1393-1512. https://doi.org/10.1016/0270-0255(86)90088-6.

Rubin, Donald B. 2005. « Causal Inference Using Potential Outcomes ». *Journal of the American Statistical Association* 100 (469): 322-31. https://doi.org/10.1198/016214504000001880.

Ryan, Patrick B., Paul E. Stang, J. Marc Overhage, Marc A. Suchard, Abraham G. Hartzema, William DuMouchel, Christian G. Reich, Martijn J. Schuemie, et David Madigan. 2013. « A Comparison of the Empirical Performance of Methods for a Risk Identification System ». *Drug Safety* 36 Suppl 1 (octobre): S143-158. https://doi.org/10.1007/s40264-013-0108-9.

Sabatier, P., M. Wack, J. Pouchot, N. Danchin, et A. S. Jannot. 2022. « A Data-Driven Pipeline to Extract Potential Adverse Drug Reactions through Prescription, Procedures and Medical Diagnoses Analysis: Application to a Cohort Study of 2,010 Patients Taking Hydroxychloroquine with an 11-Year Follow-Up ». *BMC Medical Research Methodology* 22 (1): 166. https://doi.org/10.1186/s12874-022-01628-3.

Salvo, Francesco, Florent Leborgne, Frantz Thiessard, Nicholas Moore, Bernard Bégaud, et Antoine Pariente. 2013. « A Potential Event-Competition Bias in Safety Signal Detection: Results from a Spontaneous Reporting Research Database in France ». *Drug Safety* 36 (7): 565-72. https://doi.org/10.1007/s40264-013-0063-5.

Schneeweiss, S., J. A. Rassen, R. J. Glynn, J. Avorn, H. Mogun, et M. A. Brookhart. 2009. « High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. » *Epidemiology* 20 (4): 512-22. https://doi.org/10.1097/EDE.0b013e3181a663cc.

Schneeweiss, Sebastian, Jeremy A. Rassen, Robert J. Glynn, Jerry Avorn, Helen Mogun, et M. Alan Brookhart. 2009. « High-Dimensional Propensity Score Adjustment in Studies of Treatment Effects Using Health Care Claims Data ». *Epidemiology (Cambridge, Mass.)* 20 (4): 512-22. https://doi.org/10.1097/EDE.0b013e3181a663cc.

Schuler, M. S., et S. Rose. 2017. « Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies. » *Am J Epidemiol* 185 (1): 65-73. https://doi.org/10.1093/aje/kww165.

Shortreed, Susan M, et Ashkan Ertefaie. 2017. « Outcome-adaptive lasso: variable selection for causal inference ». *Biometrics* 73 (4): 1111-22. https://doi.org/10.1111/biom.12679.

Simpson, Shawn E., David Madigan, Ivan Zorych, Martijn J. Schuemie, Patrick B. Ryan, et Marc A. Suchard. 2013. « Multiple Self-Controlled Case Series for Large-Scale Longitudinal Observational Databases ». *Biometrics* 69 (4): 893-902. https://doi.org/10.1111/biom.12078.

Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis, et Torsten Hothorn. 2007. « Bias in random forest variable importance measures: Illustrations, sources and a solution ». *BMC Bioinformatics* 8 (1): 25. https://doi.org/10.1186/1471-2105-8-25.

Tatonetti, Nicholas P., Patrick P. Ye, Roxana Daneshjou, et Russ B. Altman. 2012. « Data-Driven Prediction of Drug Effects and Interactions ». *Science Translational Medicine* 4 (125): 125ra31. https://doi.org/10.1126/scitranslmed.3003377.

Tibshirani, Robert. 1996. « Regression shrinkage and selection via the lasso ». *Journal of the Royal Statistical Society. Series B, Statistical methodology* 58: 267-88.

Uppsala Monitoring Centre. 2023. « About VigiBase ». 20 novembre 2023. https://who-umc.org/vigibase/.

Wasserman, Larry. 2004. *All of Statistics: A Concise Course in Statistical Inference*. Springer Texts in Statistics. New York, NY: Springer. https://doi.org/10.1007/978-0-387-21736-9.

Wasserman, Larry, et Kathryn Roeder. 2009. « High-dimensional variable selection ». *The Annals of Statistics* 37 (5A): 2178-2201. https://doi.org/10.1214/08-aos646.

Wells, Brian J., Kevin M. Chagin, Amy S. Nowacki, et Michael W. Kattan. 2013. « Strategies for Handling Missing Data in Electronic Health Record Derived Data ». *EGEMS (Washington, DC)* 1 (3): 1035. https://doi.org/10.13063/2327-9214.1035.

WHO Collaborating Centre for Drug Statistics Methodology. 2019. « WHO Collaborating Centre for Drug Statistics Methodology, ATC classification index with DDDs, Oslo, Norway 2018. »

Wolpert, David H. 1992. « Stacked Generalization ». *Neural Networks* 5 (2): 241-59. https://doi.org/10.1016/S0893-6080(05)80023-1.

Zhang, Zheyu, Tianping Zhang, et Jian Li. 2023. « Unbiased Gradient Boosting Decision Tree with Unbiased Feature Importance ». arXiv. https://doi.org/10.48550/arXiv.2305.10696.

Zhao, Qingyuan, et Trevor Hastie. 2021. « Causal Interpretations of Black-Box Models ». *Journal of Business & Economic Statistics* 39 (1): 272-81. https://doi.org/10.1080/07350015.2019.1624293.

Zou, Hui. 2006. « The Adaptive Lasso and Its Oracle Properties ». *Journal of the American Statistical Association* 101 (476): 1418-29. https://doi.org/10.1198/016214506000000735.

7 ANNEXE

ARTICLE PUBLIE DANS LA REVUE DRUG SAFETY

Drug Safety
https://doi.org/10.1007/s40264-022-01148-5

ORIGINAL RESEARCH ARTICLE

High-Dimensional Propensity Score-Adjusted Case-Crossover for Discovering Adverse Drug Reactions from Computerized Administrative Healthcare Databases

Etienne Volatier[1] · Francesco Salvo[2] · Antoine Pariente[2] · Émeline Courtois[1] · Sylvie Escolano[1] · Pascale Tubert-Bitter[1] · Ismaïl Ahmed[1]

Accepted: 20 January 2022

Abstract

Introduction Increasing availability of medico-administrative databases has prompted the development of automated pharmacovigilance signal detection methodologies. Self-controlled approaches have recently been proposed. They account for time-independent confounding factors that may not be recorded. So far, large numbers of drugs have been screened either univariately or with LASSO penalized regressions.

Objective We propose and assess a new method that combines the case-crossover self-controlled design with propensity scores (propensity score-adjusted case-crossover) built from high-dimensional data-driven variable selection, to account for co-medications or possibly other measured confounders.

Methods Comparison with the univariate and LASSO case-crossover was performed from simulations and a real-data study. Multiple regressions (LASSO, propensity score-adjusted case-crossover) accounted for co-medications and no other covariates. For the univariate and propensity score-adjusted case-crossover methods, the detection threshold was based on a false discovery rate procedure, while for LASSO, it relied on the Akaike Information Criterion. For the real-data study, two drug safety experts evaluated the signals generated from the analysis of 4099 patients with acute myocardial infarction from the French national health database.

Results On simulations, our approach ranked the signals similarly to the LASSO and better than the univariate method while controlling the false discovery rate at the prespecified level, contrary to the univariate method. The LASSO provided the best sensitivity at the cost of larger false discovery rate estimates. On the application, our approach showed similar performances to the LASSO and better performances than the univariate method. It highlighted 43 signals out of 609 drug candidates: 22 (51%) were considered as potentially pharmacologically relevant, including seven (16%) regarded as highly relevant.

Conclusions Our findings show the interest of a propensity score combined with a case-crossover for pharmacovigilance. They also confirm that indication bias remains a challenge when mining medico-administrative databases.

1 Introduction

The primary objective of post-marketing pharmacovigilance is the early identification of the adverse effects of marketed drugs that pre-approval trials may have failed to detect because of rarity, population subgroup specificity, or long latency. A pharmacovigilance signal is the identification of a potential association between a drug exposure and the occurrence of an adverse event that needs further study. Currently, all national and transnational pharmacovigilance systems rely on spontaneous reporting databases for which several statistical tools have been developed in order to highlight adverse drug events reported more frequently

Pascale Tubert-Bitter and Ismaïl Ahmed: Co-last authors.

✉ Etienne Volatier
etienne.volatier@inserm.fr

1 Center for Research in Epidemiology and Population Health (CESP, U1018), High-Dimensional Biostatistics for Drug Safety and Genomics Team, Inserm, Université Paris-Saclay, UVSQ, Université Paris-Sud, Villejuif, France

2 Bordeaux Population Health Research Center, Pharmacoepidemiology Team (UMR 1219), Inserm, University of Bordeaux, Bordeaux, France

Published online: 18 February 2022

Key Points

Increasing availability of large medico-administrative databases prompts the development of automated signal detection methods designed to mitigate confounding biases.

Recent methodological efforts have involved, on one side, the use of self-controlled epidemiological designs that adjust for time-invariant confounding and, on the other side, the use of high-dimensional propensity scores.

Combining the case-crossover self-controlled model and high-dimensional propensity scores improved signal detection in a simulation study by controlling the false discovery rate and generated eight highly relevant signals for acute myocardial infarction as assessed by two expert pharmacologists.

than expected [1–3]. In the last decade, penalized multiple regression approaches have been investigated [4, 5] to address poly-drug exposures. Recently, attention has shifted to the exploitation of computerized administrative healthcare databases to perform signal detection. Even though these resources were not initially designed for safety surveillance purposes, which induces some limitations such as unmeasured confounding, they may overcome the reporting biases observed in spontaneous reporting databases [6–8].

The French National System of Health Data (Système National des Données de Santé) contains precise information relative to healthcare reimbursements, such as drug reimbursements, and hospitalization for nearly the whole French population. The Échantillon Généraliste des Bénéficiaires (EGB), a representative sample at 1/97th of the Système National des Données de Santé, was recently used to perform a pilot study for drug safety signal generation [9].

A significant methodological effort for exploiting computerized administrative healthcare databases for signal detection has been made by the Observational Medical Outcomes Partnership (OMOP). The OMOP published the results of a large empirical evaluation of seven automated signal generation tools based on a reference set of 399 drug–adverse effects (containing both positive and negative controls) and involving four different adverse events [10]. One noteworthy result was that the Self-Controlled Case Series [11], a method relying only on the use of cases, performed generally well. Another self-controlled method, the case-crossover, was investigated by the OMOP [12] as an univariate method and was shown to select a large number of false positives in signal detection. However, it proved to be appropriate for studying acute events in targeted epidemiological studies, notably by avoiding using information posterior to the event [13], which is a desirable property for some severe adverse events.

Computerized administrative healthcare databases generally lack direct information on the health status of the patient as well as classical risk factors such as smoking status, body mass index, and physical activity, or classical socioeconomic covariates. Thus, when exploiting such data, it is necessary to deal with the lack of important epidemiological factors (unobserved confounders) while disposing of a large number of recorded covariates at the same time. In pharmacoepidemiology studies, i.e., targeting a given risk, measured confusion is classically accounted for by using propensity scores (PSs), which are estimates of the probability that an individual will receive the treatment of interest according to measured confounding factors. Schneeweiss et al. hypothesized that the large number of measured covariates available in computerized administrative healthcare databases may collectively be proxies of unobserved confounding factors [14]. They proposed an algorithm, the high-dimensional PS (HDPS), which prioritizes covariates according to their marginal association with the outcome of interest and the treatment, and which may be used for building the PS. The HDPS has been compared to variable selection methods such as the LASSO in the case of high-dimensional logistic regression models, and has proven to be a solid competitor [15] in the context of pharmacoepidemiologic studies. Within the context of logistic regression models, it has also recently been proposed for signal detection from both spontaneous reporting data [16] and computerized administrative healthcare databases [17], where signal detection is performed as a large-scale drug screening for a given adverse event. Other recent developments using PSs for signal detection in administrative databases extended the tree-based scan statistic designed to screen large-scale adverse events for a given drug exposure of interest [18].

These HDPS-based approaches have not yet been extended to self-controlled models such as the case-crossover. By design, a case-crossover automatically accounts for time-independent confounding factors, which makes it appealing for exploiting healthcare data not primarily designed for health research, and for which such factors may not be recorded.

The two main objectives of this work were to extend the use of the case-crossover design to the context of signal detection with multiple exposures, i.e., exposures to different drugs, and to present a new method, called PS-adjusted case-crossover (PS-CC) that combines the case-crossover self-controlled design with PSs built from a high-dimensional data-driven variable selection. A potential crucial advantage of this method when screening large numbers of drug–adverse event pairs for safety signal generation is the

possibility to choose an appropriate detection threshold. It is important to select such detection threshold because it directly affects the amount of assessment work to be done by drug safety experts that cannot document all drugs in the database. With the proposed PS-CC, it is straightforward to set the detection threshold on the basis of statistical error criteria pertaining to multiple hypothesis testing, such as the false discovery rate (FDR) [19]. In comparison, setting the amount of regularization to perform variable selection with penalized regression approaches such as the LASSO remains a field of ongoing research.

The PS-CC approach was evaluated by both Monte Carlo simulations and a real-case study. It was compared to (i) a univariate case-crossover (U-CC) applied to all drugs under scrutiny and (ii) a LASSO-based case-crossover (LASSO-CC) method like the LASSO-based Self-Controlled Case Series developed by the OMOP for signal detection purposes [20]. The real-case study concerned the adverse event acute myocardial infarction (AMI) and was performed on the EGB data for the period 1 January, 2010 to 31 December, 2016. All three approaches were applied to the cohort and signals generated by at least one method were evaluated by two drug safety experts.

2 Methods

2.1 Case-Crossover Design

The case-crossover design was first introduced by Maclure [13]. As a self-controlled design, it relies only on cases to evaluate an association between an exposure of interest and the occurrence of an adverse event. Exposure during a case period, i.e., the period just before the date of occurrence of the adverse event, is compared to those of several control periods defined before the case period (see Fig. 1). The choice of relevant control periods close to the case period is important [21]. All these periods should be separated by washout periods to avoid autocorrelations.

The case-crossover design creates a matched dataset with exposure x (in our case x equals 0 or 1) and the occurrence of the event of interest y, the individual being the matching variable. Comparisons are made only between each individual case and control periods. Hence, the Cochran–Mantel–Haenszel test and more classically conditional logistic regression are used to estimate the exposure effect in the case-crossover design. As our context is that of considering subsequently several exposures, we focus in this study on conditional logistic regression.

We first recall the regression model at the basis of the conditional logistic regression where the only difference with standard logistic regression is that intercept is allowed to vary among the matched datasets, here the individuals [22, 23]. For a given time period $p = 1, ..., t$, and a given individual $i = 1, \ldots, n$, let y_{ip} be the indicator of the occurrence of the adverse event, and x_{ip} the indicator of the exposure to a drug of interest. The probability of occurrence of the event can be written as:

$$\Pr\left(y_{ip} = 1 | x_{ip}\right) = \frac{\exp\left(\alpha_i + \beta x_{ip}\right)}{1 + \exp\left(\alpha_i + \beta x_{ip}\right)}, \quad (1)$$

where the parameters are different values of intercept α_i for each individual and a unique coefficient β of interest measuring the association between the variable of interest and the outcome.

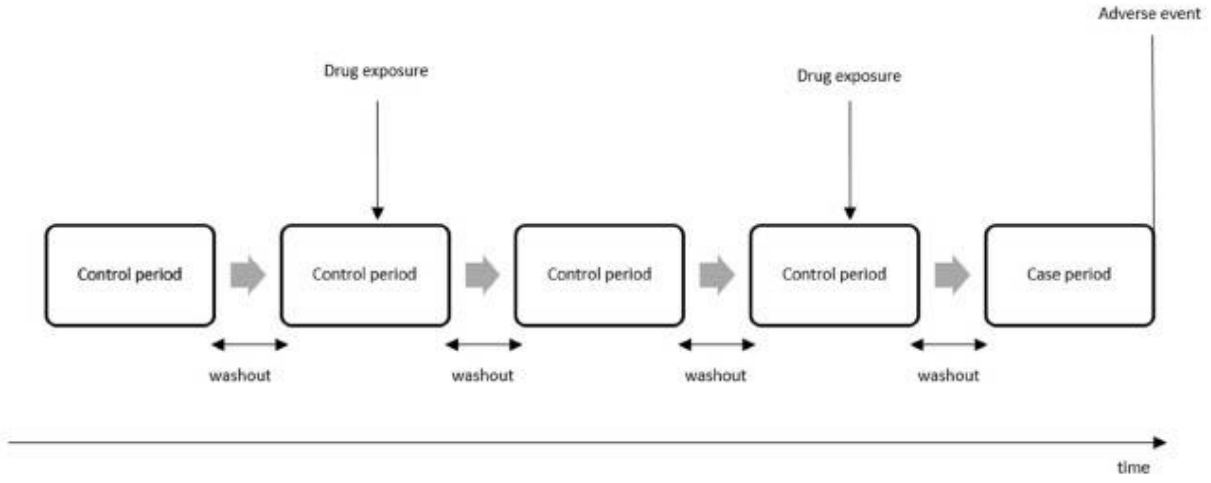

Fig. 1 Case-crossover design: case period is defined as the preceding adverse event; periods separated by washout periods

Estimating the individual baseline risks α_i is not of interest. Conditional on the fact that a unique adverse event occurred for each individual, the following conditional log-likelihood is obtained [22, 23]:

$$l\left(\beta|\sum_{p=1}^{t} y_{ip}=1\right)=\sum_{i=1}^{n}\left[\sum_{p=1}^{t} y_{ip}\beta x_{ip}-\log\left(\sum_{p=1}^{t}\exp\left(\beta x_{ip}\right)\right)\right]. \tag{2}$$

Individual baseline risks α_i vanish from the expression of the conditional log-likelihood in Eq. (2) [22, 23]. Hence, optimizing the conditional log-likelihood makes it possible to focus only on the estimation of the parameter of interest β. For univariate conditional logistic regression, the decision about the statistical significance of an association can be made with a Wald test. Note that in the context of signal detection in pharmacovigilance, the aim is to identify positive associations, thus one-sided tests are performed.

2.2 Handling Multiple Exposures: Initial Approach

In the case of m of exposures, we denote $X_{ip}=\left(x_{ip1},\ldots,x_{ipm}\right)$ the vector of exposures for a given individual i and a time period p. We denote $\beta=\left(\beta_1,\ldots,\beta_m\right)$ the vector of parameters associated with all m exposures. An initial approach to handling multiple exposures is to apply the univariate procedure described in the previous section to all drug exposures separately, and to adapt the detection thresholds to account for multiple testing with a procedure targeting the control of the FDR for one-sided hypotheses [24]. This approach is considered as naïve, as it does not account for potential confounders or concomitant exposures, but still constitutes an interesting baseline to compare the merits of more sophisticated approaches.

2.3 LASSO Case-Crossover

Alternatively, we can build multiple conditional logistic regression models including all drug exposures under surveillance. Using classical multiple conditional logistic regression can lead to numerical instabilities owing to the large number of drug exposures to be accounted for. A penalty classically used to handle a large number of covariates in regression problems is the LASSO penalty, as in the LASSO-based Self-Controlled Case Series developed by the OMOP, which corresponds to a l_1 penalty over the values of the estimated coefficients. From Eq. (2), the LASSO penalized conditional log-likelihood of the case-crossover design can be derived as [23]:

$$\begin{aligned} &l_\lambda\left(\beta|\sum_{p=1}^{t} y_{ip}=1\right)\\ &=\sum_{i=1}^{n}\left[\sum_{p=1}^{t} y_{ip}\beta^T X_{ip}-\log\left(\sum_{p=1}^{t}\exp\left(\beta^T X_{ip}\right)\right)\right]\\ &\quad-\lambda\sum_{k=1}^{m}|\beta_k|. \end{aligned} \tag{3}$$

The choice of λ governs the number of non-zero estimated coefficients and hence the number of generated signals, i.e., drug exposures associated with positive regression coefficients. Cross-validation is classically used for optimizing λ for prediction purposes and less so for variable selection. We choose instead to penalize the conditional log-likelihood by a factor of the number of selected signals, as in the Akaike Information Criterion, i.e., minimizing Eq. (4):

$$\lambda^{\text{opt}}=\text{argmin}_\lambda 2K(\lambda)-2l_\lambda\left(\hat{\beta}\right), \tag{4}$$

where K is the number of non-zero estimated β_k parameters when using λ as regularization penalty.

2.4 PS-CC Algorithm

As it is not necessarily possible in a high dimension to adjust on all covariates without penalizing the conditional log-likelihood, we propose to reduce these numerous covariates to a single variable, the PS, and to adjust the model on it. Propensity score methods are mainly used in the cohort design. The PS-CC approach is built by analogy to the use of PS in the case-control study using the control-only strategy to compute the PS [25].

For a given drug exposure of interest $k\in\{1,\ldots,m\}$, we first seek to identify a set of U potential confounders. Following the HDPS procedure by Schneeweiss et al. [14], the potential confounders are selected on the basis of the Bross bias multiplier, which is a measure of both the marginal association between (i) a potential binary confounder c and drug k and (ii) the marginal association between c and the adverse event y:

$$\text{BIAS}(k,c)=\frac{P_{C1}\left(\text{RR}_{cy}-1\right)+1}{P_{C0}\left(\text{RR}_{cy}-1\right)+1}, \tag{5}$$

where P_{C1} and P_{C0} denote the prevalence of c among exposed and non-exposed individuals to drug k, respectively, and RR_{cy} measures the relative risk between y and c. In practice, for this step, the t longitudinal measures for each individual are concatenated, and the Bross multipliers are thus calculated from $n\cdot t$ observations.

Second, a PS is built from these U potential confounders using boosted trees, as in [26] with the occurrence of the

treatment of interest as a binary outcome. Using machine learning may prove to be more efficient than multiple logistic regression, as it makes it possible to account automatically for interactions between confounders or non-linear relations with the exposure of interest. Considering the occurrence of the adverse event as very rare in our real-case study, we exclude the case period from the computation of the PS score to avoid bias due to the potential overrepresentation of the case period [25].

Third, the derived PS s is included in the regression model with the drug exposure k:

$$\Pr\left(y_{ip}=1|x_{ipk},s_{ipk}\right)=\frac{\exp\left(\alpha_i+\beta_k x_{ipk}+\gamma_k s_{ipk}\right)}{1+\exp\left(\alpha_i+\beta_k x_{ipk}+\gamma_k s_{ipk}\right)}, \quad (6)$$

and the coefficient β_k of interest is estimated by maximizing the corresponding conditional likelihood. Finally, these three steps are iterated for all m drugs of interest. As for the "naïve" univariate approach, the detection threshold for the p-values is chosen according to the FDR criterion [24].

2.5 Simulation Study

A simulation study was performed to compare the proposed method to (i) the FDR-corrected univariate case-crossover (U-CC) applied to all drugs under scrutiny and (ii) the Akaike Information Criterion-LASSO-based case-crossover (LASSO-CC).

2.6 Comparison Set-Up

Simulations were performed using the real-drug exposure data from the EGB with $n = 3000$ individuals who experienced AMI from 2010-01-01 to 2015-12-15 (see Sect. 4). This simulation study used only drug exposures as covariates. As a result, the simulated data did not include any confounder other than drugs selected to be positively associated with outcome. From these data, we selected 23 drug exposure periods of 30 days separated by washout periods of 30 days also. Drug exposures were considered binary: x_{ipk} was set to 1 as soon as drug k was delivered at least once to an individual i during period p. For computational reasons, we restricted our study to the drugs prescribed to at least 50 patients ($m = 108$).

For each simulated dataset (latter referred as to replication), m_y drugs were randomly chosen among the 108 drugs to be associated positively with response y and coefficient β. For a given individual i, we first drew the baseline risk of the individual $\alpha_i \sim U(-1.5, 1.5)$. We then generated the response y_{ip} according to a logistic multiple regression model (see Eq. 1 but with multiple exposures) for each of the 23 periods. Finally, we selected a sequence of six consecutive control periods ($y_{ip} = 0$) followed by one case period ($y_{ip} = 1$).

We considered several scenarios with m_y equal to 0, 5, or 20 and β equal to 0.5, 1, or 1.5. For each scenario, results were averaged over 500 replications. We applied two evaluation criteria. The first was the average number of true detections at a fixed number of signals. It allows a comparison of the ability of approaches to rank the signals appropriately independently of the detection threshold. The second evaluated the ability of the FDR-based approaches (PS-CC and U-CC) to control the FDR at the nominal levels $\alpha = 5\%$ and 15% and the sensitivity achieved while attempting to control for this criterion. The FDR was estimated by averaging the 500 replications of the false discovery proportions. False discovery proportions were defined as the proportion of false signals among the signals generated when the number of detections was positive. It was defined as 0 when there were no signals. False discovery rate estimates were also calculated for the LASSO-CC although no FDR control was expected. All calculations were performed with R software using packages Cyclops for LASSO-CC and xgboost for gradient-boosted trees using eta = 0.1, max_depth = 4, and the maximum number of covariates used $U = 50$.

3 Results

With regard to the ranking criterion, Fig. 2 shows that LASSO-CC and PS-CC performed as well or better than the U-CC. They outperformed U-CC for $m_y = 20$ and β equal to 1 or 1.5. For $m_y = 5$ and for large values of β (i.e. 1 and 1.5), Fig. 2b, c show that all methods ranked the signals nearly perfectly.

The comparative performances in terms of FDR control are shown in Table 1. Using PS-CC, we were able to control for the FDR in all scenarios for both $\alpha = 5\%$ and 15% thresholds, although the procedure appeared to be rather conservative. With U-CC, the FDR was not controlled in most of the scenarios, especially for large values of β for which FDR estimates were much larger than the targeted values. The LASSO-CC was also associated with very large values of FDR estimates. Finally, Table 1 shows that the differences in terms of FDR estimates resulted in large differences in terms of the average number of detections and sensitivities with U-CC, and LASSO-CC generated a much greater number of signals than PS-CC.

3.1 Real-Case Study on AMI

3.1.1 Data and Comparison Set-Up

Our dataset consisted of patients from the EGB who experienced AMI between 1 January, 2010 and 31 December, 2016, and for whom we had records for the 13 months prior to the occurrence of AMI. For patients experiencing multiple AMI during the study period, we considered only the first occurrence. We identified a total of 4099 cases with

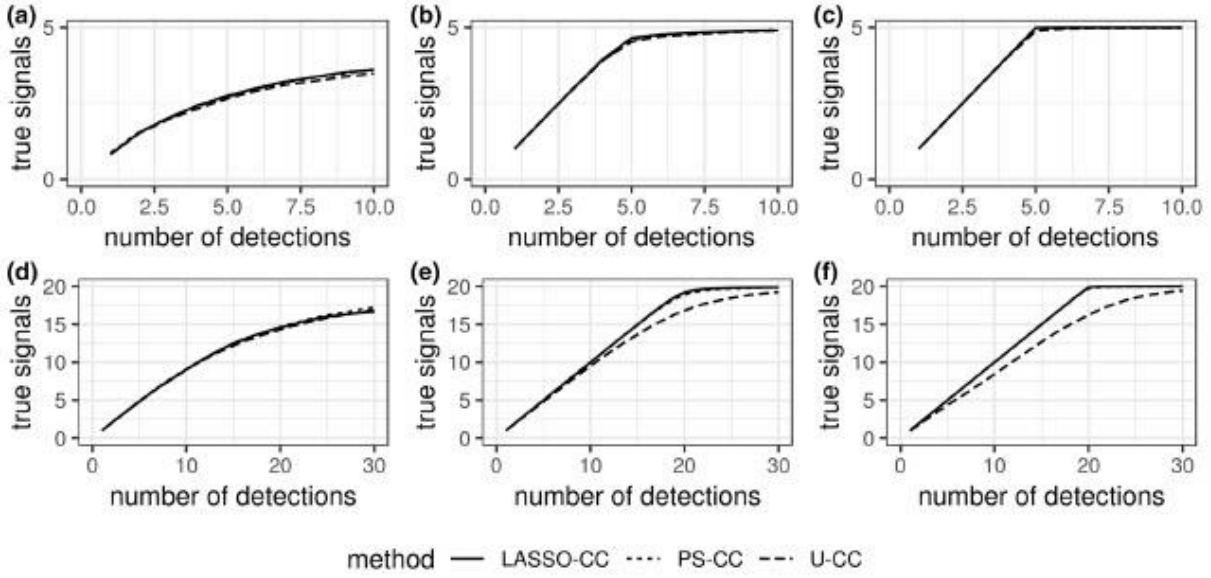

Fig. 2 Simulation results: average number of true detections according to the average number of detections by univariate case-crossover (U-CC), LASSO-based case-crossover (LASSO-CC), and propensity score-adjusted case-crossover (PS-CC) methods; each column corresponds to different β (0.5, 1, 1.5) and each row corresponds to different values of m_y (5 and 20)

Table 1 Simulation results: FDR estimates ($\widehat{\text{FDR}}$), average sensitivity (sensitivity), and average number of detections (# signals) by U-CC, LASSO-CC, and PS-CC. With U-CC and PS-CC, results were obtained at two detection thresholds ($\alpha = 0.05$ and 0.15)

	β	0	0.5	0.5	1	1	1.5	1.5
	m_y	0	5	20	5	20	5	20
U-CC								
$\alpha = 0.05$	$\widehat{\text{FDR}}$	0.22	0.05	0.09	**0.04**	0.58	0.15	0.77
	Sensitivity	–	0.14	0.40	0.80	0.99	0.99	1.00
	# signals	0.26	0.80	13.3	4.25	53.23	6.54	88.75
$\alpha = 0.15$	$\widehat{\text{FDR}}$	0.82	0.21	0.17	**0.12**	0.68	0.23	0.80
	Sensitivity	–	0.33	0.53	0.89	1.00	0.99	1.00
	# signals	2.14	2.46	14.26	5.40	70.41	8.11	99.52
LASSO-CC								
	$\widehat{\text{FDR}}$	1	0.72	0.45	0.69	0.45	0.69	0.44
	Sensitivity	–	0.80	0.85	0.99	1.00	1.00	1.00
	# signals	9.29	15.98	32.17	18.68	37.2	18.55	37.06
PS-CC								
$\alpha = 0.05$	$\widehat{\text{FDR}}$	**0.01**	**0.04**	**0.03**	**0.01**	**0.01**	**0.01**	**0.02**
	Sensitivity	–	0.17	0.24	0.61	0.83	0.96	0.99
	# signals	0.02	0.93	5.04	3.06	16.66	4.83	20.11
$\alpha = 0.15$	$\widehat{\text{FDR}}$	**0.05**	**0.08**	**0.07**	**0.01**	**0.04**	**0.01**	**0.07**
	Sensitivity	–	0.24	0.39	0.75	0.91	0.99	0.99
	# signals	0.05	1.43	8.48	3.82	19.07	4.99	21.35

With LASSO-CC, results were obtained with the AIC. Settings where FDR control is achieved are indicated in bold

AIC Akaike Information Criterion, *FDR* false discovery rate, *LASSO-CC* LASSO-based case-crossover, *PS-CC* propensity score-adjusted case-crossover, *U-CC* univariate case-crossover

information on 741 single drugs reimbursed to patients during the study period.

As for the simulation study, only concomitant drugs were considered as covariates. An additional issue in the real data was the presence of potential indication biases. For example, some drugs protective against AMI may be prescribed just before it occurs, thus they may be positively associated with it. To limit this issue, we used the SIDER database of drug side effects and indications [27] to exclude cardiovascular drugs (Anatomical, Therapeutic, and Chemical World Health Organization class: C) [28] that are used for AMI, to prevent or treat angina pectoris, or to manage heart failure. Of the 741 drugs reimbursed to the source population, 132 single drugs were thus excluded from evaluation. However, they were included in the PS for PS-CC and as covariates in the LASSO because they are relevant at adjusting for potential pre-existing heart diseases.

We applied all three case-crossover methods (U-CC, LASSO-CC, and PS-CC) tested in the simulation study and used six control periods separated by 30-day washout periods. For the PS-CC algorithm, we estimated the PS on control periods only, considering AMI as a rare occurrence [25]. For U-CC and PS-CC, the threshold for the FDR procedure was set at 0.15.

Two drug safety experts evaluated all signals generated regarding their pharmacological relevance (i.e., knowledge about potential biological plausibility), and whether the association could be affected by an indication bias. They knew that the association had been generated by at least one of the three methods, but they were blinded to the actual one(s). First, the experts worked independently, then together to discuss cases of disagreement in order to reach a consensus. Each generated signal was classified as pharmacologically relevant or not and potentially subject to indication bias or not. All signals generated were then used to build a reference set in order to evaluate the ability of the methods to prioritize highly relevant signals, i.e., signals that were pharmacologically relevant and free of indication bias.

3.1.2 Results

Sixty-eight signals were detected by at least one method. Figure 3 shows the number of signals generated by all three methods: 66 by U-CC, 43 by PS-CC, and 33 by LASSO-CC. Thirty signals were generated by all methods. All signals generated by PS-CC were also generated by U-CC. Two signals were detected only by LASSO-CC, and 22 only by U-CC. As for the simulation study, the number of signals generated by U-CC was significantly greater than that by PS-CC, which also detected more signals than LASSO-CC.

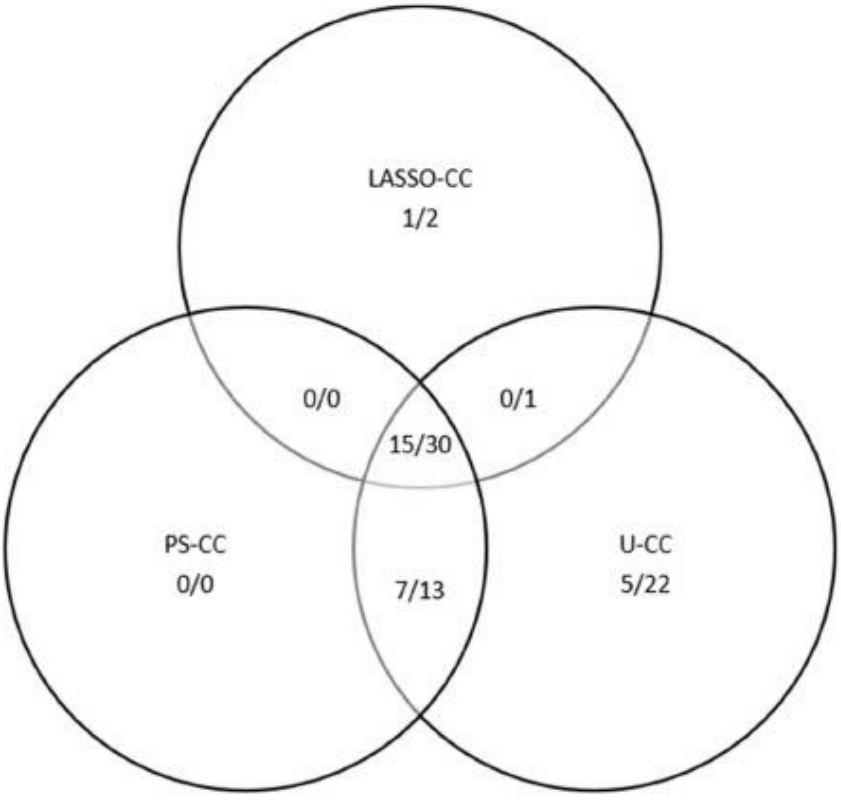

Fig. 3 Venn diagram representing the number of pharmacologically relevant signals (left number) and number of signals (right number) generated by all three methods. *LASSO-CC* LASSO-based case-crossover, *PS-CC* propensity score-adjusted case-crossover, *U-CC* univariate case-crossover

Figure 4 shows that a large proportion of signals were potentially subject to indication bias, whether considered as pharmacologically relevant or not (51/68). This proportion was comparable for the different methods tested (76%, 73%, and 77% for U-CC, LASSO-CC, and PS-CC, respectively).

Twenty-eight signals out of the 68 (41%) were considered as potentially pharmacologically relevant: 27 were detected by U-CC, 22 by PS-CC, and 16 by LASSO-CC. The proportion of pharmacologically relevant signals was greater using PS-CC: 51% (22/43) vs 41% (27/66) for U-CC and 48% (16/33) for LASSO-CC.

Eight signals out of the 68 signals (12%) were considered both pharmacologically relevant and free from indication biases, and thus regarded as highly relevant signals (Table 2). Focusing on these eight signals also gave a slight advantage to the proposed PS-CC in terms of the proportion of true-positive detections: 16% (7/43) for PS-CC vs 12% (8/66) for U-CC and 12% (4/33) for LASSO-CC.

4 Discussion

In this work, we propose an approach that combines high-dimensional variable selection PSs and a case-crossover design. As shown in the simulation study, the approach is a solid alternative to LASSO penalized case-crossover regression and provides similar performances in terms of signal ranking while making it possible to control for the FDR. The univariate approach (U-CC) fails to control for the FDR. By construction it does not account for correlations between drugs. The larger the associated coefficients, the higher the probability for U-CC to select "innocent" drugs correlated with true predictors. In most extreme simulated scenarios ($\beta = 1.5$) ranking and FDR are affected. In less extreme scenarios, the U-CC ranking remains similar to that of other approaches, but the FDR is still poorly controlled.

In the simulation study, PS-CC seemed to reduce the number of generated signals as compared with the univariate approach. The real-case study confirmed this tendency, with a greater number of signals generated by U-CC, but at the cost of many non-relevant signals. In routine practice, this could mean a low level of efficiency and time wasted in their assessment. LASSO-CC selected more signals in the simulation study than PS-CC and less signals in the application study. This inconsistency in LASSO-CC results may be induced by the fact that in the simulation study, we considered, for computational reasons, a subset of the most prevalent drugs. In the real-life study, PS-CC and LASSO approaches provided the best performance in detecting the highly relevant signals of drug-related AMI in terms of

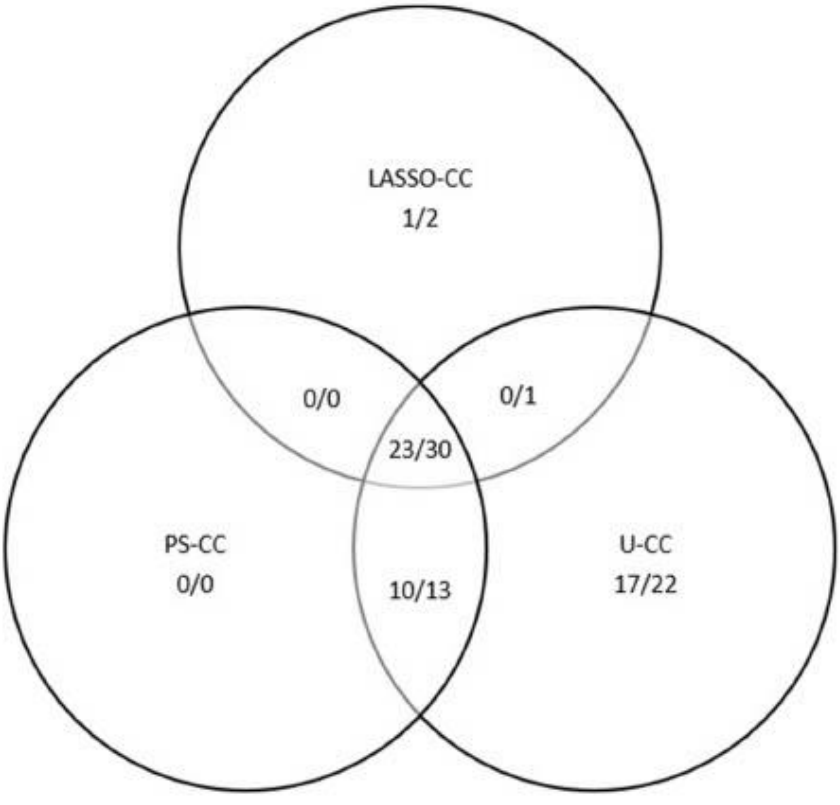

Fig. 4 Number of generated signals with a suspicion of indication bias (left number) and the number of signals (right number) generated by all three methods. *LASSO-PC* LASSO-based case-crossover, *PS-CC* propensity score-adjusted case-crossover, *U-CC* univariate case-crossover

Table 2 Highly relevant detected signals

Drug	Pharmacological relevance	U-CC	PS-CC	LASSO-CC
Alimemazine	Known to be cardiotoxic and arrhythmogenic	✓		
Domperidone	Known to be cardiotoxic and arrhythmogenic	✓	✓	✓
Fexofenadine	Can cause tachycardia and palpitations. In patients at risk, it could trigger AMI	✓	✓	
Hydroxyzine	Known to be cardiotoxic and arrhythmogenic	✓	✓	✓
Ibuprofen	Risk of AMI confirmed at high doses	✓	✓	✓
Metoclopramide	Antidopaminergic and serotoninergic on 5-HT_4 (like cisapride)	✓	✓	
Metopimazine	Known to be arrhythmogenic	✓	✓	✓
Oxymetazoline	Alpha-1 receptor agonist, vasoconstrictor	✓	✓	

AMI acute myocardial infarction, *LASSO-CC* LASSO-based case-crossover, *PS-CC* propensity score-adjusted case-crossover, *U-CC* univariate case-crossover

positive predictive value, i.e., the proportion of true-positive detections. Better results from methods handling multiple variables over the univariate approach were expected. A more remarkable result was the similar performances of LASSO-CC and PS-CC as these two methods rely on two different mechanisms to deal with multi-dimensionality.

In this work, we made a number of choices regarding our PS methodology. In our real data study, we used only concomitant drugs as potential confounders but in future work other variables may be used. We used the HDPS algorithm to filter out potential instrumental variables. We arbitrarily set the maximum number of variables to be included in the PS model at $U = 50$. In practice, the number of variables selected by the boosted tree model never reached this maximum (data not shown). We also built our PS model with gradient-boosted trees, although other machine learning algorithms could be considered. A potential limitation of these flexible machine learning models is that they may require larger sample size for estimation of nuisance parameters (e.g., the PS) [29]. Finally, PS was accounted for using adjustment in the conditional regression model (Eq. 6). Although being a classical approach to ensure balance regarding confounders, there are other alternatives such as matching or weighting. Nevertheless, the adjustment had the advantage of beings straightforward to implement in the case-only design.

We identified a theoretical limitation of the PS-CC approach: despite excluding the case period when estimating the PS, the use of the PS with a rare event like AMI could still lead to a biased estimation as in a case-control design [25, 30]. However, despite using some causal inference tools, the primary objective of signal detection is to provide a list of prioritized associations for further analysis by drug safety experts. At the stage of signal detection, unbiasedness of parameter estimation is not a primary concern and other methods such as the LASSO case-crossover also yield biased parameter estimations. The proposed method builds on the theoretical properties of PSs to correct for indication bias, and those of self-controlled designs to implicitly account for fixed confounding, either measured or not. The simulation study provides empirical evidence of the statistical properties of combining both.

Concerning the relevance of the detected signals, it is important to underline that any method for detecting drug safety signals is affected by a high rate of false positives. Signals detected by using health record databases are mostly affected by indication bias, as was the case here. Examples of these signals include treatments for chronic obstructive pulmonary disease. Even if a systemic sympathomimetic action could be hypothesized, they are used in patients with chronic obstructive pulmonary disease, which increases the risk of AMI itself and represents the most important chronic disease in tobacco smokers. Other signals were detected, mainly by U-CC, which were considered as being due exclusively to indication bias. Such is the case of proton-pump inhibitors, which are indicated for dyspeptic disorders. The latter are frequently prodromal symptoms of AMI, yet proton-pump inhibitors are commonly co-prescribed with at-risk drugs such as non-steroidal anti-inflammatory drugs or antiplatelet medications for managing gastrointestinal toxicity. A few other drugs such as oxazepam or alprazolam are possibly related neither to AMI nor to an apparent indication bias.

More than 10% of the detected signals were considered as highly relevant. These included three antihistamines (alimemazine, hydroxyzine, and fexofenadine) known for their arrhythmogenic potential [31, 32], and which could trigger AMI in patients at risk of this event. These also included three drugs used for controlling nausea and vomiting, once again for their arrhythmogenic potential: metopimazine, domperidone, and metoclopramide [32, 33]. Evidence of a risk of AMI was already offered for domperidone and metoclopramide [34], while a class effect could be hypothesized for metopimazine, even if the evidence is less clear. Oxymetazoline was also considered as

pharmacologically relevant because it is an alpha-1 receptor agonist used for its vasoconstrictor properties, which could thus increase the risk of AMI [35–37]. Finally, ibuprofen is a non-steroidal anti-inflammatory drug for which a large body of evidence suggests an increased risk of AMI, especially when used at high doses [38–40].

Our study demonstrates that confounding by indication is tricky to overcome and seems to have a large influence on results. In our opinion, it should be considered as a primary concern for signal detection from computerized administrative healthcare databases. The method we used in our real-life study was filtering by using the SIDER database of drug indications and adverse effects. However, while it is possible to constitute a comprehensive list of drugs used to treat AMI, it is difficult to obtain a fully exhaustive list of drugs prescribed for all AMI-related conditions and symptoms. Other methods introducing control individuals, such as the case-time control [41], have also been developed to address this issue but were not studied in the scope of this paper, as it was not feasible to select relevant controls.

It is highly likely that computerized administrative healthcare databases will be increasingly used to detect drug safety signals. This calls for the development of sophisticated methods able to account for the wealth, the complexity, but also the lack of information about several confounders and precise drug indications. These methods should also provide good insights into the rate of false-positive signals in order to limit the amount of evaluation work done by drug safety experts, and hence facilitate research and assessment. The present work shows that combining high-dimensional causal inference tools with appropriate epidemiological designs to perform signal detection from computerized administrative databases may lead to more efficient pharmacovigilance.

5 Conclusions

The present work shows that combining high-dimensional causal inference tools with appropriate epidemiological designs to perform signal detection from computerized administrative databases may lead to more efficient pharmacovigilance. Specifically, we investigated the interest of the PS combined with case-crossover approaches for mining high-dimensional healthcare databases. Overall, our study showed similar performances compared to the LASSO case-crossover. The simulation study showed that the FDR was correctly controlled with the proposed approach. In the real-case study, the number of false positives was limited and seven highly relevant signals were highlighted. Our study also confirms that indication bias has an important impact on the results, which remains a challenge when mining medico-administrative databases.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40264-022-01148-5.

Declarations

Funding This research was partially supported by the ANSM (Agence Nationale de Sécurité du Médicament et des Produits de Santé) as part of the 2014 'young researchers' call for projects (AAP-2014-033).

Conflict of interest The authors have declared no conflict of interest.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Availability of data and material The data on which the findings are based cannot be made freely available because of legal restrictions. Data used for the present study come from the French National Health Insurance databases and include many variables that, when combined, can lead to reidentifying subjects and then collecting health information on these individuals. Therefore, the French Data Protection Authority (CNIL) forbids making such data freely available. Access to the raw data of the French National Health Insurance must be requested from the National Health Data System (https://www.snds.gouv.fr/).

Code availability An R script is provided as Electronic Supplementary Material to illustrate implementation.

Author contributions EV, EC, SE, PTB, and IA planned and designed the study. EV drafted the manuscript and performed the research. FS and AP conducted the pharmacological assessment. All authors reviewed the manuscript and approved the final version.

References

1. Roux E, Thiessard F, Fourrier A, Bégaud B, Tubert-Bitter P. Evaluation of statistical association measures for the automatic signal generation in pharmacovigilance. IEEE Trans Inf Technol Biomed. 2005;9:518–27.
2. Tatonetti NP, Ye PP, Daneshjou R, Altman RB. Data-driven prediction of drug effects and interactions. Sci Transl Med. 2012;4:125ra31.
3. Harpaz R, DuMouchel W, Shah NH, Madigan D, Ryan P, Friedman C. Novel data mining methodologies for adverse drug event discovery and analysis. Clin Pharmacol Ther. 2012;91:1010–21.
4. Caster O, Norén GN, Madigan D, Bate A. Large-scale regression-based pattern discovery: the example of screening the WHO global drug safety database. Stat Anal Data Min ASA Data Sci J. 2010. https://doi.org/10.1002/sam.10078.
5. Ahmed I, Pariente A, Tubert-Bitter P. Class-imbalanced subsampling lasso algorithm for discovering adverse drug reactions. Stat Methods Med Res. 2018;27:785–97.
6. Pariente A, Gregoire F, Fourrier-Reglat A, Haramburu F, Moore N. Impact of safety alerts on measures of disproportionality in spontaneous reporting databases: the notoriety bias. Drug Saf. 2007;30:891–8.
7. Arnaud M, Salvo F, Ahmed I, Robinson P, Moore N, Bégaud B, et al. A method for the minimization of competition bias in signal detection from spontaneous reporting databases. Drug Saf. 2016;39:251–60.

8. Salvo F, Leborgne F, Thiessard F, Moore N, Bégaud B, Pariente A. A potential event-competition bias in safety signal detection: results from a spontaneous reporting research database in France. Drug Saf. 2013;36:565–72.
9. Arnaud M, Bégaud B, Thiessard F, Jarrion Q, Bezin J, Pariente A, et al. An automated system combining safety signal detection and prioritization from healthcare databases: a pilot study. Drug Saf. 2018;41:377–87.
10. Ryan PB, Stang PE, Overhage JM, Suchard MA, Hartzema AG, DuMouchel W, et al. A comparison of the empirical performance of methods for a risk identification system. Drug Saf. 2013;36(Suppl. 1):S143–58.
11. Farrington CP. Relative incidence estimation from case series for vaccine safety evaluation. Biometrics. 1995;51:228–35.
12. Ryan PB, Madigan D, Stang PE, Overhage JM, Racoosin JA, Hartzema AG. Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the Observational Medical Outcomes Partnership. Stat Med. 2012;31:4401–15.
13. Maclure M. The case-crossover design: a method for studying transient effects on the risk of acute events. Am J Epidemiol. 1991;133:144–53.
14. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiology. 2009;20:512–22.
15. Franklin JM, Eddings W, Glynn RJ, Schneeweiss S. Regularized regression versus the high-dimensional propensity score for confounding adjustment in secondary database analyses. Am J Epidemiol. 2015;182:651–9.
16. Courtois É, Pariente A, Salvo F, Volatier É, Tubert-Bitter P, Ahmed I. Propensity score-based approaches in high dimension for pharmacovigilance signal detection: an empirical comparison on the French Spontaneous Reporting Database. Front Pharmacol. 2018;9:1010.
17. Demailly R, Escolano S, Haramburu F, Tubert-Bitter P, Ahmed I. Identifying drugs inducing prematurity by mining claims data with high-dimensional confounder score strategies. Drug Saf. 2020;43:549–59.
18. Wang SV, Maro JC, Baro E, Izem R, Dashevsky I, Rogers JR, et al. Data mining for adverse drug events with a propensity score-matched tree-based scan statistic. Epidemiology. 2018;29:895–903.
19. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol. 1995;57:289–300.
20. Simpson SE, Madigan D, Zorych I, Schuemie MJ, Ryan PB, Suchard MA. Multiple self-controlled case series for large-scale longitudinal observational databases. Biometrics. 2013;69:893–902.
21. Mittleman MA, Mostofsky E. Exchangeability in the case-crossover design. Int J Epidemiol. 2014;43:1645–55.
22. Hardin JW. Generalized estimating equations (GEE). In: Everitt BS, Howell DC, editors. Encyclopedia of statistics in behavioral science. Wiley; 2005. pp. 39–42. https://doi.org/10.1002/0470013192.bsa250.
23. Avalos M, Grandvalet Y, Adroher ND, Orriols L, Lagarde E. Analysis of multiple exposures in the case-crossover design via sparse conditional likelihood. Stat Med. 2012;31:2290–302.
24. Ahmed I, Dalmasso C, Haramburu F, Thiessard F, Broët P, Tubert-Bitter P. False discovery rate estimation for frequentist pharmacovigilance signal detection methods. Biometrics. 2010;66:301–9.
25. Månsson R, Joffe MM, Sun W, Hennessy S. On the estimation and use of propensity scores in case-control and case-cohort studies. Am J Epidemiol. 2007;166:332–9.
26. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 13–17 August, 2016; San Francisco (CA), pp. 785–94. https://doi.org/10.1145/2939672.2939785.
27. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Res. 2016;44:D1075–9.
28. WHO Collaborating Centre for Drug Statistics Methodology. ATC classification index with DDDs. Oslo, Norway 2018; 2019.
29. Zivich PN, Breskin A. Machine learning for causal inference: on the use of cross-fit estimators. Epidemiology. 2021;32:393–401.
30. Rose S, van der Laan MJ. Simple optimal weighting of cases and controls in case-control studies. Int J Biostat. 2008;4:Article 19.
31. Poluzzi E, Raschi E, Godman B, Koci A, Moretti U, Kalaba M, et al. Pro-arrhythmic potential of oral antihistamines (H1): combining adverse event reports with drug utilization data across Europe. PLoS ONE. 2015;10: e0119551.
32. Woosley RL, Heise CW, Romero KA. QT drugs list. www.crediblemeds.org. Accessed 1 Feb 2022.
33. Stoetzer C, Voelker M, Doll T, Heineke J, Wegner F, Leffler A. Cardiotoxic antiemetics metoclopramide and domperidone block cardiac voltage-gated Na+ channels. Anesth Analg. 2017;124:52–60.
34. Coloma PM, Schuemie MJ, Trifirò G, Furlong L, van Mulligen E, Bauer-Mehren A, et al. Drug-induced acute myocardial infarction: identifying 'prime suspects' from electronic healthcare records-based surveillance system. PLoS One. 2013;8: e72148.
35. Rajpal S, Morris LA, Akkus NI. Non-ST-elevation myocardial infarction with the use of oxymetazoline nasal spray. Rev Port Cardiol. 2014;33(51):e1-4.
36. Oosterbaan R, Burns MJ. Myocardial infarction associated with phenylpropanolamine. J Emerg Med. 2000;18:55–9.
37. Akay S, Ozdemir M. Acute coronary syndrome presenting after pseudoephedrine use and regression with beta-blocker therapy. Can J Cardiol. 2008;24:e86–8.
38. Bally M, Dendukuri N, Rich B, Nadeau L, Helin-Salmivaara A, Garbe E, et al. Risk of acute myocardial infarction with NSAIDs in real world use: Bayesian meta-analysis of individual patient data. BMJ. 2017;357: j1909.
39. Bhala N, Emberson J, Merhi A, Abramson S, Arber N, et al. Vascular and upper gastrointestinal effects of non-steroidal anti-inflammatory drugs: Meta-analyses of individual participant data from randomised trials. Lancet. 2013;382:769–79.
40. Varas-Lorenzo C, Riera-Guardia N, Calingaert B, Castellsague J, Salvo F, Nicotra F, et al. Myocardial infarction and individual nonsteroidal anti-inflammatory drugs meta-analysis of observational studies. Pharmacoepidemiol Drug Saf. 2013;22:559–70.
41. Suissa S. The case-time-control design. Epidemiology. 1995;6:248–53.