UNIVERSITE DU DROIT ET DE LA SANTE – LILLE 2
Ecole Doctorale Biologie-Santé

Thèse de doctorat

Biostatistiques, informatique médicale et technologies de communication

Réutilisation de données hospitalières pour la recherche d'effets indésirables liés à la prise d'un médicament ou à la pose d'un dispositif médical implantable

Présentée et soutenue publiquement le jeudi 11 juin 2015

Par le Docteur Grégoire Ficheur

Directeur de Thèse :	Monsieur le Professeur Régis Beuscart
Rapporteurs :	Monsieur le Professeur Alain Venot Madame le Professeur Christine Verdier
Examinateurs :	Monsieur le Professeur Régis Beuscart Monsieur le Professeur Alain Duhamel Monsieur le Professeur Marc Cuggia Monsieur le Docteur Emmanuel Chazard

Sommaire

Tableaux

Figures

Equations

Glossaire

ANSM : Agence Nationale de Sécurité du Médicament et des Produits de Santé

AP-HP : Assistance publique - Hôpitaux de Paris

ARS : Agence Régionale de Santé

ASIP : Agence des Systèmes d'Information Partagés de santé

ATC : Anatomique, Thérapeutique et Chimique

ATIH : Agence Technique de l'Information sur l'Hospitalisation

AVK : anti-vitamine K

AVP : Accident de la voie publique

BCPNN : Bayesian Confidence Propagation Neural Network

BPCO : Broncho-Pneumopathie Chronique Obstructive

CART : Classification And Regression Tree

CCAM : Classification Commune des Actes Médicaux

CDSS : Clinical Decision Support System

CH : Centre Hospitalier

CHAID : CHi-squared Automatic Interaction Detector

CHRU : Centre Hospitalier Régional Universitaire

CIM-10 : Classification Internationale des Maladies version 10

CNEDiMTS : Commission Nationale d'Evaluation des Dispositifs Médicaux et des Technologies de Santé

CNIL : Commission Nationale de l'Informatique et des Libertés

DAS : Diagnostic Associé Significatif

DM : Dispositif Médical

DMI : Dispositif Médical Implantable

DMIA : Dispositifs Médicaux Implantables Actifs

DP : Diagnostic Principal

DRG : Diagnosis Related Group

EIM : évènement iatrogène médicamenteux

ENEIS : Enquête Nationale sur les Evénements Indésirables liés aux Soins

ETL : Extract, Transform, Load

FDA : Food and Drug Administration

GHM : Groupe Homogène de Malades

GHS : Groupe Homogène de Séjours

GPS : Gamma Poisson Shrinker

HAD : Hospitalisation à Domicile

HAS : Haute Autorité en Santé

HL7 : Health Level Seven

IDS : l'Institut des Données de Santé

INSERM : Institut National de la Santé et de la Recherche Médicale

InVS : Institut de Veille Sanitaire

IRR : Incidence Ratio Rate

IUPAC : International Union of Pure and Applied Chemistry

LASSO : Least Absolute Shrinkage and Selection Operator

LEOPARD : Longitudinal Evaluation of Observational Profiles of Adverse events Related to Drugs

LGPS : Longitudinal Gamma Poisson Shrinker

LOINC : Logical Observation Identifiers Names and Codes

LPP : Liste des Produits et Prestations

MCO : Médecine Chirurgie Obstétrique

MICI : Maladies Inflammatoires Chroniques de l'Intestin

NPU : Nomenclature, Properties and Units

OMOP : Observational Medical Outcomes Partnership

OMS : Organisation Mondiale de la Santé

OSI : Open Systems Interconnection

PLS : Partial Least Squares

PMSI : Programme de Médicalisation des Systèmes d'Information

PRR : Proportional Reporting Ratios

PSIP : Patient Safety through Intelligent Procedures in medication

PSY : Psychiatrie

PTH : Prothèse Totale de Hanche

ROR : Reporting Odds Ratio

RSA : Résumé de Sortie Anonymisé

RSS : Résumé de Sortie Standardisé

SEP : Sclérose en Plaques

SNIIRAM : Système National d'Information Inter-Régimes de l'Assurance Maladie

SSR : Soins de Suite et de Réadaptation

SVM : Support Vector Machine

T2A : tarification à l'activité

Remerciements

Au Directeur de cette thèse,

Monsieur le Professeur Régis Beuscart

Professeur des Universités – Praticien Hospitalier
Professeur de Biostatistiques, Informatique Médicale et Technologies de Communication
Chef du Service d'Information et Archives Médicales du CHRU de Lille
Directeur du Centre d'Etudes et de Recherche en Informatique Médicale
Officier dans l'Ordre des Palmes Académiques

Vous m'avez fait l'honneur d'accepter de diriger cette thèse. Je vous remercie pour les conseils précieux et rigoureux que vous m'avez prodigués. Recevez ici le témoignage de ma gratitude et de mon profond respect.

Aux Rapporteurs de cette thèse,

Monsieur le Professeur Alain Venot

Professeur des Universités – Praticien Hospitalier
Professeur de Biostatistiques, Informatique Médicale et Technologies de Communication
Chef du Département interhospitalier de Santé Publique des hôpitaux Avicenne, Jean Verdier et René Muret
Directeur adjoint du « Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé » (LIMICS)

Vous m'avez fait l'honneur d'être rapporteur de cette thèse. Recevez ici le témoignage de ma gratitude et de mon profond respect.

Madame le Professeur Christine Verdier

Professeur des Universités en Informatique
Responsable de l'équipe de recherche SIGMA
Directrice de l'Unité de Formation et de Recherche « Informatique, Mathématiques et Mathématiques Appliquées » (IM²AG)

Vous m'avez fait l'honneur d'accepter de juger cette thèse. Soyez assurée de mon profond respect et de ma reconnaissance.

Aux examinateurs de cette thèse,

Monsieur le Professeur Alain Duhamel

Professeur des Universités – Praticien Hospitalier
Professeur de Biostatistiques, Informatique Médicale et Technologies de Communication
Responsable de la Plateforme d'aide méthodologique du CHRU de Lille

Vous me faites l'honneur d'accepter d'être examinateur de ce travail. Veuillez accepter mes sincères remerciements et soyez assuré de mon profond respect.

Monsieur le Professeur Marc Cuggia

Professeur des Universités – Praticien Hospitalier
Professeur de Biostatistiques, Informatique Médicale et Technologies de Communication
Responsable de l'équipe projet données massives en santé à l'Université Rennes 1

Vous me faites l'honneur d'accepter de juger cette thèse. Soyez assuré de mon profond respect et de ma reconnaissance.

Monsieur le Docteur Emmanuel Chazard

Maître de Conférences des Universités – Praticien Hospitalier
Maître de Conférences en Biostatistiques, Informatique Médicale et Technologies de Communication

Vous me faites l'honneur d'accepter d'être examinateur de cette thèse. Travailler avec vous a été un plaisir. Je vous adresse mes plus vifs remerciements.

1 Introduction

1.1 Réutilisation de données en informatique

1.1.1 Définition de la réutilisation de données

La réutilisation de données, qui se traduit par « data reuse » en langue anglaise, se définit comme l'exploitation secondaire de bases de données selon une finalité différente de celle pour laquelle elles sont initialement élaborées.

Le fait de réutiliser des bases de données pour une finalité différente se retrouve classiquement dans les secteurs bancaire, assurantiel ou de vente de biens de grande consommation. Ainsi par exemple, à l'occasion du fonctionnement quotidien d'une compagnie d'assurance, des données sont recueillies en routine lors du recrutement des clients (données démographiques), lors de l'encaissement des primes d'assurance (recettes) et lors du règlement des sinistres (dépenses). Ces données peuvent être réutilisées pour bâtir des modèles de risques individuels : ces modèles sont ensuite utilisés pour déterminer des primes d'assurance personnalisées, tenant compte des caractéristiques de l'assuré. De la même manière, les supermarchés recueillent des informations de gros volume décrivant les achats de leurs clients, chaque fois que ces clients passent en caisse. Ces informations peuvent être analysées et recoupées avec les informations démographiques et économiques disponibles à travers les programmes de fidélité. Un exemple bien connu est l'analyse du « panier de la ménagère » [1] : cette technique a rapidement permis de mettre en évidence que les personnes qui achetaient des jouets électriques achetaient également des piles électriques. Rapidement, des piles plus chères ont été installées à proximité immédiate des jouets électriques afin d'augmenter les ventes, en volume et en prix unitaire. Dans chacun de ces exemples, il existe donc une information recueillie en routine (les primes et sinistres, le ticket de caisse) à des fins transactionnelles (le remboursement, la facturation). Ces données, une fois disponibles, ont pu être analysées de manière rétrospective à une fin tout autre (l'ajustement des primes, la réorganisation des rayons) : cette nouvelle finalité peut être qualifiée de décisionnelle. Cette dissociation entre la finalité et les modalités de recueil d'une part, et la finalité et les méthodes d'analyse d'autre part, permet de définir le data reuse. Le data reuse peut également être caractérisé par les propriétés qui en découlent. Le data reuse présente de nombreux avantages : il permet des études à bas coût, s'échelonnant sur des durées très courtes, exploitant rapidement des données abondantes et accédant de fait à une forte puissance statistique. Les principaux écueils du data reuse sont la difficulté de le mener à bien du fait d'une présentation inadéquate des données, et une qualité de réponse à la question posée parfois imparfaite, du fait même de la discordance entre la finalité de recueil et la finalité de l'analyse.

Le Tableau 1 présente les propriétés des bases de données selon leur finalité initiale en « Utilisation » ou selon une nouvelle finalité en « Réutilisation ». Ce tableau permet d'insister sur la nature prospective et transactionnelle du recueil d'information dans le cadre initial de l'utilisation et sur la nature rétrospective et décisionnelle des analyses réalisées sur ces mêmes bases de données lorsqu'elles sont réutilisées. Lorsque ces données sont utilisées selon leur

finalité première, elles permettent la gestion de transactions à l'échelle d'un client unique alors qu'elles impliquent l'analyse d'un échantillon comprenant un nombre élevé de clients lorsque ces mêmes données sont réutilisées.

Tableau 1 - Propriétés des bases de données transactionnelles et décisionnelles

	Nombre de client(s) concerné(s)	Temps	Cadre	Finalité
Utilisation	1	Prospectif	Transactionnel	Cohérente avec le recueil
Ré-utilisation	Nombre élevé	Rétrospectif	Décisionnel	Différente de la finalité du recueil

De façon analogue au cas des secteurs évoqués ci-avant, la réutilisation de données médicales est progressivement devenue une réalité. Dans le champ médical, la première réutilisation de données médico-administratives dans une finalité épidémiologique remonte au début des années 1980. Celle-ci s'est accélérée avec la numérisation en routine d'un nombre toujours plus grand d'informations.

La réutilisation de données médicales peut être d'emblée illustrée par deux exemples :

Un premier exemple concerne la réutilisation d'une base de données hospitalière et plus particulièrement une table contenant les administrations médicamenteuses et une table contenant les résultats de biologie de ces patients hospitalisés. Dans cet exemple, la finalité première de ces tables de données est l'administration au patient du médicament prescrit par le médecin et l'exploitation par le médecin des résultats du bilan biologique, à des fins de soin individuel. La réutilisation de ces données peut consister à mettre en relation des administrations médicamenteuses et des résultats de biologie pour rechercher des associations entre certains médicaments et certains résultats de biologie à type d'effets indésirables médicamenteux, sur un grand nombre de séjours.

Un second exemple concerne la réutilisation d'une base de données de facturation hospitalière. Nous détaillerons la nature de ces données de facturation dans le chapitre « 1.2.1.4 Données hospitalières médico-administratives ». Une de leurs propriétés est d'être disponibles dans une base unique contenant toutes les informations des hôpitaux français. Ces données comprennent des informations sur les diagnostics et les actes du séjour hospitalier du patient. Ces informations sont utilisées pour résumer l'information en un groupe homogène de séjours (GHS) auquel correspond un prix qui sera versé à l’établissement. La finalité première de ces données est donc la facturation du séjour à l'assurance maladie et/ou au patient. Ces données peuvent néanmoins être réutilisées afin d’estimer la fréquence d'hospitalisation pour un diagnostic donné.

La réutilisation de données change le cadre classique des études épidémiologiques comprenant un recueil actif de données, qu'il soit prospectif ou rétrospectif. Construire des cohortes prospectives en épidémiologie coûte cher alors que la réutilisation de données a un coût marginal proche du coût de l'analyse des bases de données puisque les données concernées ont déjà été recueillies par ailleurs.

1.1.2 Définition des Big Data ou données massives

Les Big Data [2] ou données massives peuvent être définies comme des bases de données de grande dimension. Cette notion de grande taille peut s'entendre à travers 5 dimensions :

- Un nombre élevé d'individus statistiques (nombre de lignes)
- Un nombre élevé de variables (nombre de colonnes)
- Un nombre élevé de tables et de relations (cardinalité du modèle relationnel)
- Un nombre élevé de valeurs possibles pour les variables qualitatives mono-valuées (par exemple, le diagnostic principal d'un séjour peut prendre près de 32 000 valeurs différentes)
- Un nombre élevé de mesures du même paramètre dans le temps (par exemple, plusieurs mesures de kaliémie durant un séjour donné)

Les données recueillies en routine au décours des applications industrielles ou de santé entrent aujourd'hui de fait dans les big data. On pourra citer par exemple les bases nationales du PMSI : pour ce qui est du seul secteur MCO pour l'année 2013, cette base de données contient plus de 27 millions de séjours, et plus de 244 observations correspondant à des éléments codés (diagnostics, actes, dispositifs, etc.). Cette base est généralement considérée comme fiable depuis l'année 2008, année à laquelle la « suppression du taux de conversion » était universellement appliquée, permettant ainsi un contrôle de l'existence réelle de chaque séjour. Elle contient donc à ce jour 6 années complètes réutilisables.

Néanmoins, le terme de « big data » ne se résume pas aux données recueillies en routine. Il peut également concerner les données massives recueillies expérimentalement, notamment dans le champ des « -omics », qui regroupe notamment la génomique, la protéomique ou l'étude du métabolisme. Alors que les données recueillies en routine sont massives du fait du nombre élevé d'individus statistiques principalement, les données utilisées en « -omics » sont massives du fait principalement du nombre élevé de variables.

Il existe aujourd'hui un amalgame entre le terme de big data et la notion de data reuse, dans la mesure où une partie des big data (toute celle qui n'est pas liée aux « -omics ») peut faire l'objet d'un data reuse, et réciproquement une grande partie des études de data reuse s'appuie sur des données massives. Néanmoins, dans le présent travail, nous nous focaliserons sur le concept de data reuse, tout en gardant à l'esprit que le sujet que nous traitons est parfois appelé big data à tort.

1.1.3 Thématiques des études réutilisant des bases de données médicales

Au cours d'une analyse bibliographique, nous avons classé à dire d'expert les études réutilisant les bases de données hospitalières en dix grands types de thématiques de publication. Précisons que ce classement repose sur une estimation approximative de volume réalisée à partir des études exploitant le « nationwide inpatient sample » aux Etats-Unis, qui est un échantillon contenant des données provenant de plus de 7 millions de séjours hospitaliers chaque année. De plus, les catégories présentées peuvent être partiellement redondantes. Dix catégories sont mises en évidence. Pour chacune des catégories thématiques, nous recensons :

- les critères d'inclusion permettant l'inauguration du suivi des patients hospitalisés au sein d'une cohorte rétrospective
- les évènements spécifiquement suivis au sein de cette thématique
- les expositions d'intérêt (spécifiques de cette thématique) pour lesquelles une association avec l'évènement a été recherchée.

Nous détaillerons ensuite ci-après les évènements et expositions communes à l'ensemble des thématiques. Le recensement systématique d'une part des critères d'inclusion et d'autre part des expositions et des évènements d'intérêt pourrait être utilisé comme un outil permettant de mettre en évidence de façon systématique les associations « exposition-évènement » non encore explorées.

Voici les 10 grandes thématiques de publications réutilisant des données médicales mises en évidence dans le cadre de ce travail bibliographique.

1) Les interventions chirurgicales [3–12]
   a) Critères d'inclusion ou de stratification : appendicectomie, cystectomie, lobectomie, hystérectomie, œsophagectomie, néphrectomie, traitement de l'anévrysme abdominal, traitement d'anévrysmes cérébraux (rompus ou non),cholécystectomie, chirurgie bariatrique, arthrodèse lombaire, chirurgie des valves tricuspides, mitrales, aortiques, chirurgie coronarienne, prostatectomie, chirurgie colorectale
   b) Evènements d'intérêt spécifiques : épilepsie, thrombose, hémorragie, infection, syndrome obstructif
   c) Expositions d'intérêt spécifiques : laparoscopie versus cœlioscopie versus robotique assistée, clips versus coils

2) Les affections rares [13–16]
   a) Critères d'inclusion ou de stratification : maladies rares, donneurs vivants de rein, suivi des enfants conçus par fécondation in vitro
   b) Evènements d'intérêt spécifiques : insuffisance rénale, naissance

3) La grossesse [17–20], le post-partum et le nouveau-né
   a) Critères d'inclusion ou de stratification : grossesse et dépression, grossesse et infection, grossesse et Sclérose en Plaque (SEP), grossesse et mois de juillet,

grossesse et maladie de Willebrand, grossesse et lupus, grossesse et polyarthrite rhumatoïde, grossesse et Maladie Inflammatoire Chronique de l'Intestin (MICI), grossesse et migraine

b) Evènements d'intérêt spécifiques : nombre de semaines d'aménorrhée, poids de naissance du nouveau-né, hystérectomie, pré-éclampsie, hémorragie intracérébrale, hémorragie du post-partum, thrombose du post-partum
c) Expositions d'intérêt spécifiques : obésité de la mère

4) Le suivi des dispositifs médicaux [21–32]
   a) Critères d'inclusion ou de stratification : arthroplastie totale du genou, arthroplastie totale de hanche, arthroplastie de l'épaule, stent, pacemaker, défibrillateur implantable
   b) Evènements d'intérêt spécifiques : thrombose, descellement, infection
   c) Expositions d'intérêt spécifiques : unilatéral versus bilatéral

5) Les affections cardio-vasculaires [33–41]
   a) Critères d'inclusion ou de stratification : accidents vasculaires cérébraux, infarctus du myocarde, insuffisance cardiaque, arythmie cardiaque par fibrillation auriculaire
   b) Evènements d'intérêt spécifiques : épilepsie, insuffisance rénale, saignement versus thrombose

6) Le suivi de pathologies chroniques [42–44]
   a) Critères d'inclusion ou de stratification : MICI, diverticulose, migraine, syndrome d'apnée du sommeil, maladie de Parkinson, cirrhose, Broncho-Pneumopathie Chronique Obstructive (BPCO), certains cancers
   b) Evènements d'intérêt spécifiques : exacerbation, sortie contre avis médical, thrombose, cancer, chute
   c) Expositions d'intérêt spécifiques : traitement versus observation, comparaison de traitements

7) Les hémorragies digestives
   a) Critères d'inclusion ou de stratification : ulcères gastroduodénaux, hémorragie digestive basse

8) Les complications de certains traitements [45,46]
   a) Critères d'inclusion ou de stratification : cancers, déficit de l'acuité visuelle, thromboses
   b) Evènements d'intérêt spécifiques : traitement thrombolytique, irradiation pelvienne, chirurgie de la cataracte, chimiothérapie
   c) Expositions d'intérêt spécifiques : fracture, chute, hémorragie cérébrale

9) Les traumatismes
   a) Critères d'inclusion ou de stratification : traumatisme cérébral, chute
   b) Evènements d'intérêt spécifiques : chute, ostéoporose

10) Les études écologiques
    a) Evènements d'intérêt spécifiques : hospitalisation
    b) Expositions d'intérêt spécifiques : exposition géographique quelque soit le type (par exemple pression atmosphérique, trafic routier, luminosité)

Nous avons également identifié les expositions et évènements non spécifiques d'une thématique. Trois évènements sont fréquemment étudiés indépendamment de la thématique [47–52] :

- la mortalité
- la mortalité intra-hospitalière
- la ré-hospitalisation.

De la même manière, les huit expositions générales suivantes peuvent être combinées avec l'ensemble des thématiques :

- « centre hospitalier universitaire » versus « centre hospitalier » (ou de façon analogue « centre hospitalier urbain » versus « rural »)
- volume d'activité du centre hospitalier ou du praticien
- spécialité du professionnel de santé
- âge
- sexe
- jour de la semaine versus week-end
- type de couverture dont bénéficie l'assuré
- une comorbidité (quelle que soit la nature du diagnostic).

Les principales thématiques de réutilisation des données médico-administratives hospitalières ayant été présentées, nous nous intéressons désormais à l'importance de la réutilisation de données en termes de publications internationales au sein des principales revues scientifiques.

1.1.4 Importance de la réutilisation de données dans les publications internationales

Nous illustrons ici la place prise par les études épidémiologiques, et par la réutilisation des données au sein de cette discipline, dans les grands journaux scientifiques. Nous avons étudié de manière exhaustive les publications parues dans quatre revues internationales sur une période allant du 1er juillet 2013 au 31 décembre 2013. Les revues concernées sont :

- le *New England Journal of Medicine* (*NEJM*) [20,53–56],
- le *Journal of the American Medical Association* (*JAMA*) [57–71],
- le *Lancet* [72–75],
- *Nature Medicine*.

Nous avons identifié, pour chacun de ces journaux, les études épidémiologiques, par opposition aux essais randomisés et aux études à forte composante biologique incluant les

études cas-témoins mettant en évidence des facteurs génétiques telles que les « Genome Wide Associations Studies ». Les méta-analyses ont également été exclues de ce dénombrement.

La revue *Nature Medicine* n'a publié aucune étude observationnelle sur cette période. Les trois autres revues, qui sont à parution hebdomadaire, ont publié approximativement une centaine d'études chacune, à raison de 4-5 articles par semaine. Comme présenté dans le Tableau 2, un ordre de grandeur de la place accordée aux études observationnelles est d'1/3 pour le *JAMA* et d'1/6 pour le *NEJM* et le *Lancet*. Si l'on détaille l'origine de ces études observationnelles, on retrouve que 24 études sont issues de la réutilisation de données et 42 sont issues de cohortes plus classiques dans leur construction. Ainsi, 36% des études observationnelles publiées étaient des études réutilisant des données recueillies pour une autre finalité.

Cette proportion d'études reposant sur la réutilisation de données peut sembler élevée si l'on considère que l'analyse est réalisée à partir de données dont le recueil n'avait pas été initialement pensé dans une perspective épidémiologique. Néanmoins, il faut garder à l'esprit que le coût de ces études réutilisant des données est minime comparativement aux études de cohorte classiques. Ensuite, ces deux types de cohorte ne répondent pas forcément aux mêmes problématiques : certaines informations ne faisant pas l'objet d'une numérisation en routine ne pourraient pas être étudiées par la réutilisation de données ; inversement, la puissance statistique apportée par certaines bases de données réutilisées, comprenant parfois plusieurs millions d'individus, ne pourrait pas être obtenue par la construction de cohortes classiques. Ce gain de puissance est particulièrement précieux dans le cas du suivi d'évènements rares [16,76]. Enfin, précisons que plusieurs études utilisent conjointement des données issues de cohortes classiques et des données issues de bases de données administratives [15,77–79]. Ainsi, il ne serait pas toujours pertinent de les opposer et la présentation dichotomique réalisée dans le Tableau 2 est une simplification.

Parmi les études de cohortes rétrospectives issues de la réutilisation de données, la ventilation par pays (dont les données sont issues) est présentée dans le Tableau 2. Les principaux pays exploitant leurs bases de données administratives dans une perspective épidémiologique sont les Etats-Unis d'Amérique puis les pays du Nord de l'Europe (Danemark, Suède), le Royaume-Uni et le Canada. Il est vraisemblable que ces pays correspondent au petit nombre de pays ayant fait l'effort depuis plus de dix ans de structurer et d'analyser leurs données administratives.

Tableau 2 - Dénombrement des études observationnelles parues dans *NEJM, JAMA, Lancet* et *Nature Medicine* du 01/07/2013 au 31/12/2013

		NEJM	***JAMA***	***Lancet***	***Nature Medicine***	**Total**
Cohortes prospectives et registres classiques	Framingham, Nurses' Health Study, MIDA, REGARDS, NASS, etc.	**9**	**18**	**15**	**0**	**42**
Cohortes rétrospectives issues de la réutilisation de données		**5** [20,53–56]	**15** (6–20)	**4** (21–24)	**0**	**24**
	USA (Medicare, Veterans Affairs, California, Nationwide Inpatient Sample)	2	10	0	0	12
	Danemark	2	2	0	0	4
	UK (General Practice Research Database)	1	1	1	0	3
	Canada (Ontario)	0	2	0	0	2
	Suède	0	0	1	0	1
	Brésil	0	0	1	0	1
	Afrique du Sud	0	0	1	0	1
Etudes observationnelles (total)		**14**	**33**	**19**	**0**	**66**

Plus généralement, la réutilisation de données médico-administratives dans une perspective épidémiologique s'est développée au rythme de la numérisation en routine d'un nombre croissant de données médico-administratives. Nous les présentons dans le chapitre suivant « 1.2 Bases de données réutilisables en informatique médicale ».

1.2 Bases de données réutilisables en informatique médicale

La numérisation d'informations en routine ne cesse de croître. Cette évolution générale se produit également dans le champ de la santé. En effet, la dématérialisation du dossier patient entraîne le stockage en routine d'un volume croissant d'informations médicales. De plus, les modes ambulatoire et hospitalier de facturation en France vont de pair avec un recueil centralisé par l'assurance maladie et l'Agence Technique de l'Information sur l'Hospitalisation (ATIH) d'informations médicales structurées.

Nous présentons ici dans une première partie « 1.2.1 Données médicales recueillies en routine ou disponibles en open data » les principales bases de données pouvant être réutilisées dans le champ médical.

Tout d'abord, nous assistons depuis quelques années à un mouvement *open data* mondial d'ouverture des données dans un but de plus grande transparence des administrations. Les données disponibles en *open data* seront présentées dans la partie « 1.2.1.1 Données disponibles en *open data* ».

Il existe ensuite une base de données constituée et tenue par l'Assurance Maladie de la Sécurité Sociale : il s'agit du Système National d'Information Inter-Régimes de l'Assurance Maladie (SNIIRAM). Cette base de données inclut les données de facturation ambulatoires (telles que les consultations et actes externes, et les achats de médicaments en pharmacie). Cette base est croisée par l'Assurance Maladie avec les données de mortalité du Centre d'épidémiologie sur les causes médicales de décès (CépiDC). Ces données seront présentées en détail dans la partie « 1.2.1.2 Données ambulatoires médicales : SNIIRAM et CépiDC ».

Puis, l'informatisation croissante du dossier médical hospitalier va de pair avec le stockage en routine d'informations variées. Nous présenterons dans la partie « 1.2.1.3 Données hospitalières médicales » les données que l'on peut extraire du dossier patient hospitalier informatisé.

Il existe ensuite une base de données relative à la facturation hospitalière : il s'agit de la base nationale du Programme de Médicalisation des Systèmes d'Information (PMSI). Cette base de données permet notamment le financement des établissements de santé par l'Assurance Maladie de la Sécurité Sociale. Elle couvre par définition uniquement les activités d'hospitalisation (à l'exclusion par exemple des consultations faites dans les hôpitaux). Elle est constituée de données fournies directement par les hôpitaux et collectées par l'ATIH. Les données du PMSI (pour le champ « Médecine Chirurgie Obstétrique » (MCO)) incluent principalement les diagnostics du patient codés selon la Classification Internationale des Maladies version 10 (CIM-10) [80], les actes médicaux thérapeutiques et diagnostiques codés selon la Classification Commune des Actes Médicaux (CCAM) et certaines informations démographiques. Un des intérêts de ces données de facturation réside dans l'existence d'un numéro anonyme unique par patient permettant de suivre le parcours de soins des patients. Ces données seront présentées en détail dans la partie « 1.2.1.4 Données hospitalières médico-administratives : base nationale du PMSI ».

L'ensemble des données évoquées peuvent être réutilisées pour construire des études observationnelles. Plusieurs contraintes techniques ou réglementaires doivent néanmoins être surmontées avant de pouvoir envisager leur réutilisation : nous les présentons dans une seconde section « 1.2.2.2 Contraintes réglementaires pour la réutilisation de données médicales ».

1.2.1 Données médicales recueillies en routine ou disponibles en open data

Nous présentons ici les principales bases de données médicales existant en France dont nous distinguons 4 types principaux :

- les données ouvertes dites en *open data*,
- des données ambulatoires médicales du SNIIRAM incluant les données de mortalité (CépiDC),
- les données hospitalières médicales (données médicales hospitalières locales faisant partie du dossier patient informatisé),
- les données hospitalières médico-administratives nationales, à travers la Base Nationale du PMSI.

Enfin, après avoir présenté ces 4 types de données, nous confronterons les caractéristiques des types de données que nous exploiterons dans cette thèse, à savoir les deux types de données hospitalières.

1.2.1.1 Données disponibles en open data

Une donnée ouverte ou donnée en *open data* est une « donnée numérique d'origine publique ou privée diffusée de manière structurée selon une méthodologie et une licence ouverte garantissant son libre accès et sa réutilisation par tous » [81]. On retrouve parfois l'acronyme ODOSOS2 (de l'anglais « *open data, open source and open standards* »).

L'ouverture des données a pour but de rendre les données accessibles à tous en éliminant les restrictions sur le droit d'accès et de réutilisation. Ces restrictions concernent le plus souvent le droit d'exploitation et de reproduction, quel que soit le type de donnée numérisée. Il serait difficile de proposer une liste exhaustive des domaines intéressés par cette ouverture des données mais elle touche des secteurs aussi variés que la cartographie, la sociologie, l'environnement ou le domaine juridique.

Les données ouvertes s'inscrivent dans un mouvement de mise à disposition de données détenues le plus souvent pas des administrations publiques et va dans le sens d'une plus grande visibilité des politiques publiques. Ainsi, par exemple en France, près de 14 000 jeux de données ont progressivement été rendus accessibles sur le site data.gouv.fr dont la page d'accueil est présentée en illustration sur la Figure 1.

Figure 1 - Page d'accueil du site web http://data.gouv.fr

Dès lors que ces données sont rendues accessibles, des projets de recherche utilisant ces données peuvent être envisagés, ils reposent le plus souvent sur une des trois méthodes suivantes : tout d'abord, il est possible de mettre en relation deux bases de données en *open data* que l'on parvient à fusionner selon une variable commune aux deux bases de données. Il pourra s'agir par exemple d'individus statistiques qui ne sont pas des personnes, mais des institutions publiques ou des établissements de santé. Ensuite, il est également possible de mettre en relation des données individuelles recueillies dans le cas d'une étude épidémiologique avec des données agrégées issues d'une base disponible en open data. Ainsi par exemple, on peut imaginer mettre en relation des caractéristiques respiratoires recensées dans le cadre d'une cohorte de patients et des taux de pollution par code géographique accessibles en open data. Enfin, dans le cas où il n'est pas possible de fusionner les données individuelles (quelle que soit la granularité) issues de deux bases de données, des études écologiques peuvent toujours être envisagées.

Cette ouverture des données dépasse très largement le cadre français et peut s'observer notamment aux Etats-Unis. Dans ce pays, cette ouverture des données administratives publiques ou privées a pris une forme particulière dans le champ médical et en particulier dans le cas des essais thérapeutiques. En effet, en mai 2013, le laboratoire GlaxoSmithKline® a commencé à mettre à disposition les données de 200 essais thérapeutiques réalisés depuis 2007 [82,83]. La mise à disposition de ces données individuelles déidentifiées est réalisée en plus de l'enregistrement systématique (de tous les essais cliniques) sur le site de la FDA (Food and Drug Administration) qui permet de prévenir le biais de publication. Cette évolution apparaît comme une avancée majeure dans la reproductibilité des analyses et dans

la réplication des résultats. Notre travail porte sur la réutilisation des données issues de grandes bases de données observationnelles, ce type de données n'est pas réutilisé dans le cadre de notre travail.

1.2.1.2 Données ambulatoires médicales : SNIIRAM et CépiDC

Nous présentons maintenant le Système national d'information inter-régimes de l'Assurance Maladie de la Sécurité Sociale (SNIIRAM), que l'on peut familièrement présenter comme la base de données de l'Assurance Maladie. Créé en 1999 par la loi de financement de la sécurité sociale, le SNIIRAM est une base de données nationale dont les objectifs sont de « contribuer à une meilleure gestion de l'Assurance maladie et des politiques de santé, d'améliorer la qualité des soins et de transmettre aux professionnels de santé les informations pertinentes sur leur activité. » [84]

Le SNIIRAM poursuit quatre grandes finalités définies par l'article L.161-28-1 du code de la sécurité sociale (citées ici telles quelles) :

- Améliorer la qualité des soins, notamment par la comparaison des pratiques aux référentiels, accords de bons usages ou contrats de bonne pratique ; l'évaluation des comportements de consommation de soins ; l'analyse des caractéristiques et des déterminants de la qualité des soins ;
- Contribuer à une meilleure gestion de l'Assurance maladie, notamment par : la connaissance des dépenses de l'ensemble des régimes d'assurance maladie ; l'évaluation des transferts entre enveloppes correspondant aux objectifs sectoriels de dépenses fixés en fonction de l'objectif national de dépenses d'assurance maladie, dans le cadre de la loi de financement de la sécurité sociale ; l'analyse quantitative des déterminants de l'offre de soins et la mesure de leurs impacts sur l'évolution des dépenses d'assurance maladie ;
- Contribuer à une meilleure gestion des politiques de santé, notamment par : l'identification des parcours de soins des patients ; le suivi et l'évaluation de l'état de santé des patients et leurs conséquences sur la consommation de soins ; l'analyse de la couverture sociale des patients ; la surveillance de la consommation de soins en fonction de différents indicateurs de santé publique ou de risque ;
- Transmettre aux prestataires de soins les informations pertinentes relatives à leur activité, à leurs recettes et, s'il y a lieu, à leurs administrations.

Le SNIIRAM constitue donc une base de données complète et détaillée sur le parcours des patients à travers le système de soins qui fait l'objet d'un remboursement par l'Assurance Maladie.

Les données du SNIIRAM ont été collectées et organisées progressivement depuis 2002. Aujourd'hui, trois ensembles de restitution sont mis en service :

- Il existe tout d'abord 15 bases de données thématiques de données agrégées appelées datamarts orientées vers un sujet spécifique tel que l'analyse de l'offre de soins ou les dispositifs médicaux.
- Il existe également un échantillon général des bénéficiaires (EGB) au 97ème de la population assurée permettant la réalisation d'analyses longitudinales chez 660 000 patients.
- Il existe enfin une base de données individuelles des bénéficiaires qui permet de réaliser des études sur la consommation des soins.

Concernant les données ambulatoires médicales disponibles dans le cadre du SNIIRAM, on retrouve principalement les prestations remboursées dans le cadre de soins réalisés en médecine de ville (cette notion incluant les actes et consultations externes réalisés à l'hôpital, mais en-dehors de toute hospitalisation stricto sensu). Ces données comprennent :

- des informations sur le prestataire de soins ou biens médicaux, voire le prescripteur
- le codage détaillé des :
    - médicaments achetés en pharmacie
    - actes techniques des médecins
    - dispositifs médicaux
    - prélèvements biologiques (en tant qu'acte de prélèvement et d'analyse, mais sans le résultat de ces prélèvements)

De plus, cette base a été connectée (on parle de *linkage*) à la base nationale du PMSI et à la base des décès du CépiDC (Centre d'épidémiologie sur les causes médicales de décès). Il est courant d'observer la dénomination « SNIIRAM-PMSI » depuis que ces bases ont été fusionnées. Ces bases identifient le patient par un identifiant unique anonyme appelé le numéro ANO. Cet identifiant est permanent pour une personne dans le temps et dans l'espace, c'est-à-dire que le même patient a le même identifiant quelle que soit la structure de soins à laquelle il a affaire.

La base nationale du PMSI est présentée de façon détaillée dans le paragraphe spécifique « 1.2.1.4 Données hospitalières médico-administratives : base nationale du PMSI », nous ne la détaillons donc pas ici.

Concernant le CépiDC, il s'agit d'un laboratoire INSERM ayant pour mission le recueil puis l'analyse statistique descriptive des causes médicales de décès en France. Ces causes proviennent de la documentation en routine des certificats de décès. Le travail de *linkage* ayant été réalisé avec les données individuelles du SNIIRAM, la base de données SNIIRAM dispose désormais des données de mortalité pour l'ensemble de ses assurés. Précisons que la base nationale du PMSI ne contient que les décès survenus à l'hôpital contrairement à la base du CepiDC qui contient l'ensemble des décès, qu'ils soient survenus en ambulatoire ou à l'hôpital.

Ces données font l'objet d'une analyse récente en France et plusieurs simplifications réglementaires facilitant leur accès sont encore à l'étude, notamment par la Commission Open

Data en Santé [85]. Le cadre réglementaire d'accès à ces données pourrait évoluer rapidement. Nous verrons de façon détaillée dans la partie « 1.2.2.2 Contraintes réglementaires pour la réutilisation de données médicales » les contraintes réglementaires pour accéder à ces données.

1.2.1.3 Données hospitalières médicales

Nous présentons maintenant les données issues du dossier patient hospitalier informatisé. Précisons tout de suite qu'il s'agit à chaque fois de données constituées et disponibles dans un hôpital donné, et non de données nationales contrairement aux sections contiguës. Plus généralement, le dossier médical informatisé est une partie essentielle d'un système d'information « idéal » qui permettrait la mise en réseau de l'ensemble des informations médicales rattachées à un patient. Le dossier médical dans sa version informatisée conserve les propriétés du dossier papier ; il concerne ainsi « l'élaboration des suivis de diagnostic, les traitements, mais aussi plus généralement tous les échanges écrits entre les professionnels de santé » tels que décrits dans la loi du 4 mars 2002 [86] relative aux droits des patients. Une spécificité du dossier médical dans sa version numérisée est la nécessité de le déclarer auprès de la Commission nationale de l'informatique et des libertés (CNIL).

Informatiser un dossier de santé permet, au moins en théorie :

- A l'échelle du patient (transactionnel) : d'améliorer la coordination des soins entre professionnels de santé
- A l'échelle du professionnel de santé : de simplifier l'exercice
- A l'échelle de la connaissance collective (décisionnel) : de permettre la construction d'études épidémiologiques au sens large.

Les données présentées dans cette partie correspondent à l'ensemble des données que l'on peut extraire du dossier patient hospitalier informatisé, l'ensemble des types de données disponibles peut ainsi varier géographiquement d'un établissement à l'autre et au cours du temps. Encore une fois, ces données ne sont disponibles que dans une structure de soins donnée et ne font pas l'objet d'une base nationale accessible à tous les chercheurs.

Ces données informatisées peuvent contenir plusieurs types de données :

- les données du PMSI détaillées dans le paragraphe « 1.2.1.4 Données hospitalières médico-administratives : base nationale du PMSI »
- Les résultats de biologie médicale (ce terme regroupant l'hématologie biologique, la biochimie clinique, la microbiologie médicale et l'immunopathologie)
- Les prescriptions ou administrations médicamenteuses
- Les courriers (le plus souvent en texte libre), notamment les comptes-rendus d'actes et la lettre de sortie
- Les sorties de dispositifs médicaux électroniques, comme les électrocardiogrammes

- Les images issues des activités d'imagerie médicale (radiographies, IRM, TDM, etc.)
- Etc.

La Figure 2 illustre un extrait de dossier médical informatisé : cet affichage permet de visualiser la prescription d'un médicament sur un logiciel de prescription connectée (« Computerized Physician Order Entry » [87] en anglais).

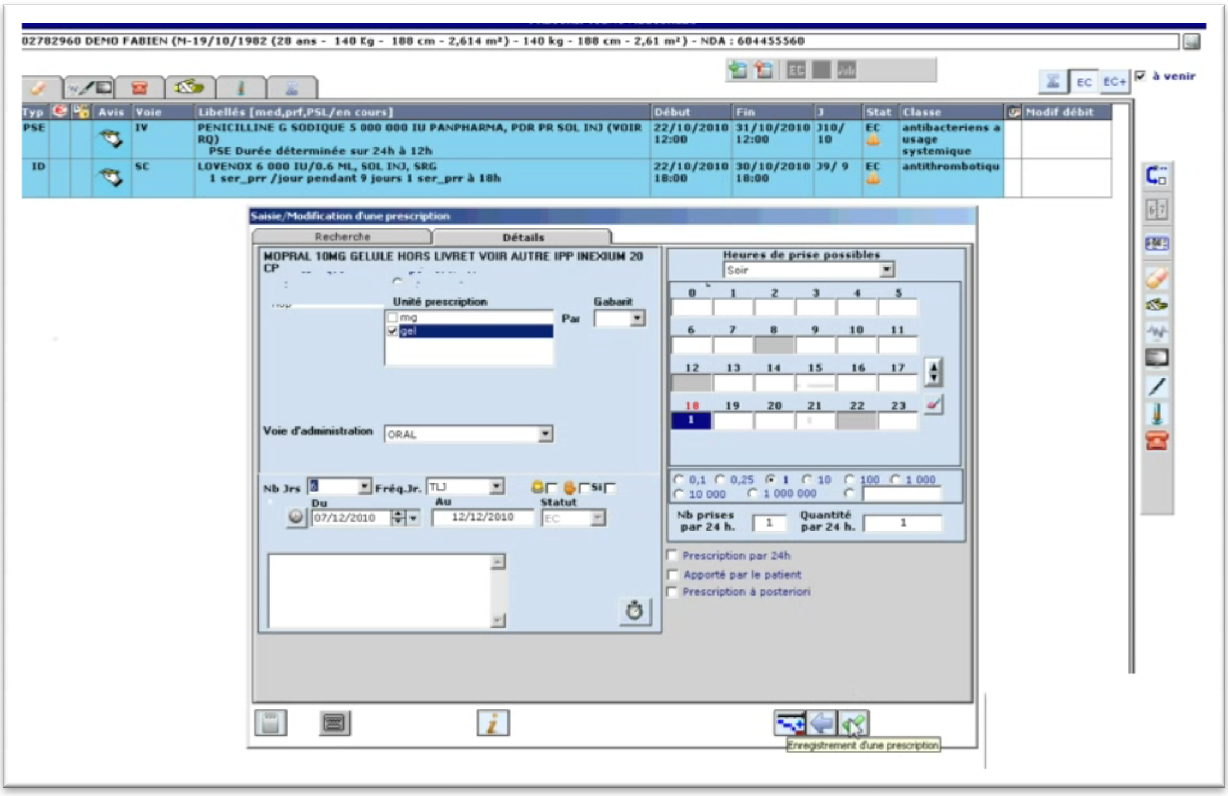


Figure 2 - Illustration d'un extrait de dossier médical informatisé

Ces données hospitalières sont plus riches que les données SNIIRAM car elles comprennent notamment des informations très détaillées sur l'administration de médicaments et sur les résultats de biologie médicale. En revanche, il est en général compliqué de pouvoir fusionner des données issues d'hôpitaux différents, nous abordons de façon plus détaillée cette problématique dans le paragraphe « 1.2.2.1.1 Interopérabilité des données en santé ». Plus prosaïquement, chaque étude demande également une extraction locale, chronophage et source éventuelle de délais ou blocages d'origine technique, organisationnelle ou politique.

1.2.1.4 Données hospitalières médico-administratives : base nationale du PMSI

Le Programme de médicalisation des systèmes d'information (PMSI) commence son histoire en 1991 : la loi n° 91-748 du 31 juillet 1991 portant réforme hospitalière [88] instaure alors que « les établissements de santé, publics et privés, doivent procéder à l'évaluation et à l'analyse de leur activité ». Tandis que le recueil d'activité en termes de journées

d’hospitalisation réalisées existait déjà, afin de mesurer l'activité médicale des établissements, il apparaît alors nécessaire de disposer d'informations standardisées : le PMSI offre un standard de recueil de l’information médicale.

Le PMSI a été inspiré par le modèle américain des Diagnosis Related Groups (DRG). Selon ce modèle, une phase de recueil de l’information permet dans un second temps d’identifier des « groupes homogènes de malades » (GHM). En effet, chaque séjour est unique lorsqu’on regarde en détail ses attributs mais, intuitivement, on remarque qu’il existe pourtant des séjours d’appendicectomie non compliquée, des séjours d’appendicectomie compliquée, d’accouchements par voie basse, etc. L’objectif du groupage des séjours est de les classer automatiquement dans de tels groupes d’activités, homogènes d’un point de vue médical, et également d’un point de vue économique. Tandis que cette homogénéité médicale est le point d’entrée du modèle (en bref, les critères médicaux constituent les critères algorithmiques de classement), l’homogénéité économique est la sortie du modèle (en bref, les groupes sont constitués de manière à réduire l’hétérogénéité des coûts du séjour, ces coûts ayant été évalués sur un échantillon d’étude).

Le PMSI était initialement un outil de description de l'activité hospitalière, ce n'est qu'en 2004 qu'il est devenu un outil d'allocation des ressources dans le cadre de la tarification à l'activité (T2A) dans le secteur du court séjour (MCO). Jusqu’en 2004, les GHM correspondaient à des coûts relativement homogènes mais les recettes de l’établissement étaient fixées de manière déconnectée, à l’aide d’un budget global ou d’un prix de journée selon les établissements, ces systèmes étant hérités des décennies précédentes. En 2004 et 2005, les GHM (ou plus précisément les GHS, groupes homogènes de séjours) se sont vu affecter un tarif qui a progressivement constitué la majorité des ressources liées aux soins des établissements de santé de MCO.

Les informations recueillies dans le secteur MCO comprennent principalement des codes diagnostiques encodés selon la CIM-10, des codes d'actes diagnostiques ou thérapeutiques encodés selon la classification CCAM et des informations démographiques et administratives. Ces informations sont produites pour chaque séjour hospitalier en France et agrégées dans un Résumé de Sortie Standardisé (RSS) puis anonymisées en un Résumé de Sortie Anonymisé (RSA) qui est alors transmis à l'Agence Régionale de Santé (ARS), à l’ATIH et à la caisse pivot de l'assurance maladie qui rémunère en retour l'hôpital sur la base de cette déclaration. Le tarif d'un séjour hospitalier est calculé à partir du RSA : le RSA est analysé par une fonction de groupage permettant d'affecter chaque séjour dans un GHS et *in fine* de lui attribuer un tarif.

Ainsi, pour ces informations, les hôpitaux français publics ou privés respectent le format unique défini par l'ATIH. La conséquence directe de ce format unique respecté partout en France est l'interopérabilité physique, syntaxique et sémantique de ces données, nous y reviendrons dans la partie « 1.2.2.1.1.1 Interopérabilités physique, syntaxique et sémantique ».

Enfin, d'autres données sont disponibles dans le secteur public (ex-DGF) pour le MCO : ces données sont les Dispositifs Médicaux Implantables (DMI) et les Molécules Onéreuses (MON). La Figure 3 présente les types de données disponibles pour le secteur MCO ainsi que pour les autres secteurs hospitaliers que sont le secteur SSR, le secteur HAD et le secteur PSY. Le secteur EXT représenté sur la Figure 3 présente lui les soins externes réalisés à l'hôpital, il ne s'agit donc pas d'un secteur d'hospitalisation.

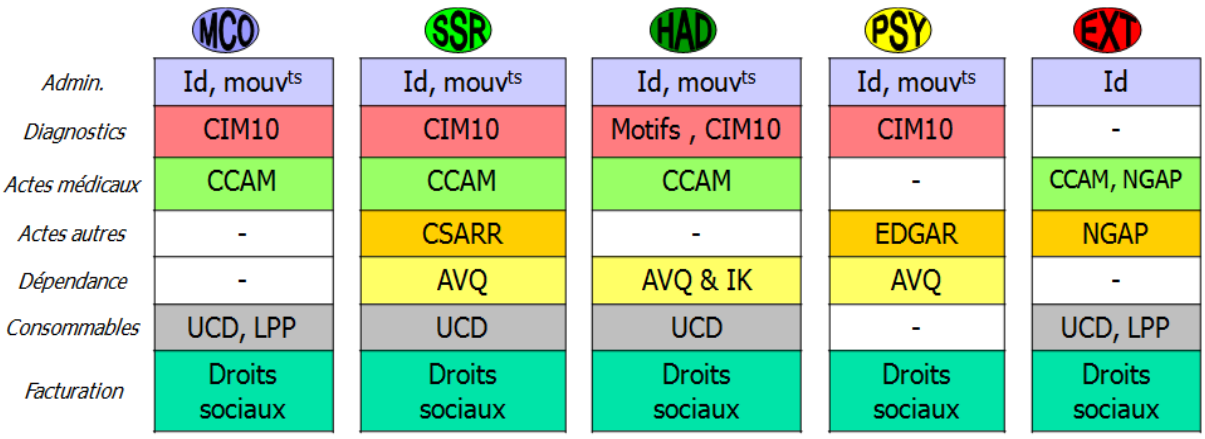


Figure 3 - Informations recueillies dans le PMSI en fonction des secteurs d'activité

1.2.1.5 Présentation et comparaison des données réutilisées dans cette thèse

Nous détaillons et comparons maintenant plus en avant les données que nous utiliserons dans le cadre de cette thèse, à savoir les données hospitalières médico-administratives du PMSI et les données médicales du dossier patient informatisé d'un hôpital partenaire. Ces données sont disponibles pour les années 2007 à 2013.

Le Tableau 3 présente la disponibilité des types de données au sein de nos deux bases de données principales, à savoir la base nationale du PMSI pour les années 2007 à 2013 et la base hospitalière d'un centre hospitalier partenaire pour les années 2007 à 2013. La notion d'ordre éventuel pouvant concerner ces données ainsi que leur disponibilité sont également présentées.

Les services hospitaliers présents au sein du centre hospitalier partenaire sont la cardiologie, la pneumologie, la médecine interne, l'hépato-gastro-entérologie, la chirurgie, la gynécologie-obstétrique, un service de chirurgie et un service d'urgences. Les données du centre hospitalier partenaire sont reçues selon un modèle de données proposé dans le cadre du projet PSIP [89]. Nous obtenons ainsi de cet hôpital partenaire :

- les administrations médicamenteuses codées selon la classification Anatomique, Thérapeutique et Chimique (ATC)

- les résultats de biologie médicale codés selon la terminologie International Union of Pure and Applied Chemistry (IUPAC)
- les courriers hospitaliers en texte libre
- les données PMSI dans le champ MCO incluant les actes selon la CCAM, les codes diagnostiques selon la CIM-10 et des informations démographiques et administratives.

Dans le cas de la base nationale, les données que nous exploitons sont :

- les données du PMSI dans le champ MCO décrites ci-avant
- les données relatives aux DMI (disponibles dans le secteur ex-DGF) codées selon la Liste des Produits et Prestations (LPP) de l'assurance maladie
- le fichier ANO permettant le chaînage des données à l'échelle du patient.

D'autres types de données sont disponibles en MCO ou dans d'autres champs mais nous ne les présentons pas ici.

On peut observer dans le Tableau 3 que le nombre de séjours hospitaliers disponibles dans l'hôpital partenaire est très inférieur au nombre de séjours hospitaliers disponibles dans la base nationale du PMSI : cette différence importante illustre la nature interopérable des données du PMSI en France contrairement aux données médicales hospitalières pour lesquelles il est beaucoup plus compliqué de pouvoir construire des bases inter-hospitalières. L'interopérabilité des données est présentée de façon plus détaillée dans la partie « 1.2.2.1.1 Interopérabilité des données en santé », les modèles de données supportant ces données hospitalières sont eux présentés dans la partie « 1.2.2.1.1.4 Modèles de données supportant les données utilisées dans cette thèse ». En revanche, les types de données disponibles sont plus importants au sein de notre hôpital partenaire qu'elles ne le sont dans le cadre du PMSI. En synthèse, les données de notre hôpital partenaire sont riches de par la nature des champs qu'elles contiennent mais concernent un nombre de patients plus réduit comparativement aux données médico-administratives du PMSI.

L'analyse de ces données hospitalières revêt plusieurs contraintes réglementaires que nous détaillons dans la partie « 1.2.2.2 Contraintes réglementaires pour la réutilisation de données médicales ». Les données réutilisées dans le cadre de cette thèse ont fait l'objet de deux autorisations CNIL : (i) une première (enregistrement numéro VIa0335797v) concernant les données du centre hospitalier partenaire de 2007 à 2013 et une seconde concernant les données de la base nationale du PMSI (autorisation AE141087). Les courriers hospitaliers issus du centre hospitalier partenaire ont été anonymisés par la méthode FASDIM [90].

Pour chacune des données du Tableau 3, leur nature ordonnée ou non est précisée ainsi que la nature de l'ordre. Deux types d'ordre peuvent être distingués dans les données :

(i) Le premier type d'ordre concerne le jour de survenue de l'information qui est ensuite codée au cours du séjour hospitalier : cette notion d'ordre concerne le jour d'administration d'un médicament ou le jour de prélèvement de l'échantillon pour un résultat de biologie

médicale. Les données ordonnées, pour lesquelles la chronologie est disponible en cours de séjour, sont l'administration médicamenteuse et les résultats de biologie médicale. Cette notion d'ordre pourrait également concerner le jour de production d'un courrier hospitalier ou de survenue d'une pathologie (ou symptôme) mais ces deux dernières informations ne sont disponibles qu'*a posteriori* du séjour et ne sont pas datées. Du fait des règles de codage, qui impliquent de coder toute maladie active durant le séjour, il n'est pas certain que le diagnostic codé ait été mis en évidence pour la première fois au cours du séjour concerné et il n'est pas non plus possible de connaître d'emblée la chronologie de survenue de ces diagnostics et/ou symptômes. Concernant les actes CCAM, la chronologie n'est pas toujours disponible car la date de réalisation au cours du séjour n'est pas rendue obligatoire par l'ATIH. Néanmoins, lorsqu'elle est disponible, c'est bien la date de réalisation de l'acte qui est indiquée.

(ii) Le second type d'ordre concerne la position des mots dans la phrase.

Tableau 3 - Disponibilité des données utilisées et types d'ordres contenus dans ces données

Type de données [terminologie]	Donnée ordonnée	Type d'ordre intéressant	Disponibilité de l'ordre dans la base de données	Exemple de donnée brute pour un séjour hospitalier	Centre hospitalier partenaire Années 2007 à 2013 (80 000 séjours et séances)	Base nationale du PMSI Années 2007 à 2013 (170 000 000 séjours et séances)
Administrations médicamenteuses [ATC]	Oui	Jour d'administration du médicament	Oui	Chlorure de Potassium [A12BA01] (jour 2, jour 3, jour 5); Ampicilline [J01CA01] (jour 1, jour 2, jour 3, jour 4, jour 5, jour 6, jour 7);	Oui	Non
Résultats de biologie médicale [IUPAC]	Oui	Jour de prélèvement de l'échantillon dans le milieu biologique (ex : sang, urine, etc.)	Oui	Kaliémie=3.8mmol/l (jour 3); Kaliémie=5.2mmol/l (jour 6); Créatininémie=10mg/l (jour 4); Créatininémie=14mg/l (jour 5); Créatininémie=17mg/l (jour 6);	Oui	Non
Lettre de sortie [texte libre]	Oui	Position du mot au sein du courrier hospitalier (pos)	Oui	Hyperkaliémie (pos 42) était (pos 43) associée (pos 44) à (pos 45) une (pos 46) détérioration (pos 47) de (pos 48) la (pos 49) fonction (pos 50) rénale (pos 51);	Oui	Non
Diagnostics [CIM-10]	Oui	Jour de survenue du diagnostic ou symptôme	Non (encodage *a posteriori* du séjour)	Insuffisance rénale aiguë [N17.9]; Hypertension artérielle essentielle [I10]; Hyperkaliémie [E87.5];	Oui	Oui
Actes diagnostiques et thérapeutiques [CCAM]	Oui	Jour de réalisation de l'acte	Oui (partiellement)	Remplacement de l'articulation coxofémorale par prothèse totale [NEKA020];	Oui	Oui
Dispositifs médicaux implantables [LPP]	Oui	Jour de pose du DMI	Non	Hanche, tête ou tête à jupe, céramique [3111390];	Oui	Oui (ex-DGF)
Démographie	Non			Sexe="féminin"; Age=54;	Oui	Oui

1.2.2 Contraintes réglementaires et techniques pour la réutilisation de données médicales

L'accumulation d'un nombre croissant d'informations en routine dans le cadre de bases de données permet d'envisager la réutilisation de grands volumes de données. Néanmoins, pour pouvoir être exploitées, ces données informatisées doivent être accessibles d'un point de vue réglementaire, et interopérables. Nous aborderons ces deux questions ci-dessous.

1.2.2.1 Contraintes techniques pour la réutilisation de données médicales

1.2.2.1.1 Interopérabilité des données en santé

1.2.2.1.1.1 Interopérabilités physique, syntaxique et sémantique

L'interopérabilité revêt classiquement trois dimensions que sont l'interopérabilité physique, l'interopérabilité syntaxique et l'interopérabilité sémantique. Ces trois aspects font partie de la norme « Open Systems Interconnection » (OSI) qui comprend 7 niveaux. L'interopérabilité n'est pas indispensable en soi à la mise en relation de deux bases de données : un humain serait tout à fait capable de corriger l'absence d'interopérabilité à la main. En revanche, l'interopérabilité est indispensable à l'automatisation de ces mises en relation ou traitements : elle garantit que les bases peuvent être combinées ou traitées sans avoir à mettre en place de prétraitement ad hoc. Elle s'avère d'autant plus indispensable qu'il s'agit de traiter des bases de données d'origines différentes.

1.2.2.1.1.1.1 Interopérabilité physique

L'interopérabilité physique rend compte de la transmission effective des signaux entre les interlocuteurs. Son service correspond typiquement à l'émission et la réception d'un bit entre deux ordinateurs. Elle définit par exemple le flux (HTTP, FTP, etc.), l'encodage (ASCII, UTF-8, etc.) et la structuration des données (XML, texte tabulé, etc.).

Un contrexemple de données interopérables physiquement serait constitué par un fichier en texte tabulé compressé au format ZIP d'une part, et un flux de données transmis par protocole SOAP (XML en HTTP) d'autre part.

1.2.2.1.1.1.2 Interopérabilité syntaxique

L'interopérabilité syntaxique rend compte de la mise en correspondance de champs entre deux bases de données, c'est à dire de la possibilité de pouvoir accéder à la table puis au champ correspondant dans la table. Elle tient compte du fait qu'une même information peut être décrite par des données structurées différemment. Des outils de type « Extract, Transform, Load » (ETL) jouent un rôle central dans la mise en correspondance de bases de données.

Un contrexemple de données interopérables syntaxiquement serait constitué par une base de données comprenant l'âge en années, et l'autre base de données comprenant notamment la date de naissance et la date du séjour. Dans les deux cas l'âge peut être obtenu, mais il faut définir un mécanisme de transformation ad hoc.

1.2.2.1.1.1.3 Interopérabilité sémantique

L'interopérabilité sémantique rend compte de la mise en correspondance du contenu des champs en termes de représentation de l'information, en particulier l'information qualitative (non numérique). Ce type d'interopérabilité est rendu possible par l'utilisation de terminologies comprenant des codes et des libellés, ces derniers pouvant être déclinés dans différentes langues. Les terminologies internationales sont pour la plupart proposées par l'Organisation Mondiale de la Santé (OMS).

Un contrexemple de données interopérables sémantiquement serait constitué par une base de données codant « érysipèle », l'autre « érésipèle », l'autre « dermo-hypodermite infectieuse », et la dernière « A46 ».

L'adhésion des systèmes d'information hospitaliers à ces terminologies varie sensiblement d'un type d'information à l'autre. Ainsi, la CIM pour les diagnostics et la classification ATC pour les médicaments sont fortement utilisées à travers le monde. Il en va différemment des classifications « Logical Observation Identifiers Names and Codes » (LOINC) et IUPAC pour les résultats de biologie médicale. En effet, le plus souvent, les hôpitaux utilisent une terminologie locale qui se rapproche du libellé en langage naturel du paramètre. En France, l'Assistance publique - Hôpitaux de Paris (AP-HP) a réalisé la migration vers la terminologie LOINC pour les résultats de biologie médicale et rapporte un nombre élevé de codes manquants dans LOINC, codes néanmoins progressivement ajoutés à la classification. Une recommandation de l'Agence des Systèmes d'Information Partagés de santé (ASIP) incite à l'utilisation de LOINC en France mais cette utilisation n'est pas obligatoire.

1.2.2.1.1.1.4 Données interopérables hospitalières en France

Les principales données hospitalières interopérables en France sont les données relatives à la facturation hospitalière (base nationale du PMSI). Ces données concernent l'ensemble des établissements de santé pour les secteurs MCO, Soins de Suite et de Réadaptation (SSR), Psychiatrie (PSY) et Hospitalisation à Domicile (HAD). L'interopérabilité inter-hospitalière syntaxique et sémantique est acquise pour ces données, la contrainte légale et financière (en MCO et HAD, deux secteurs soumis au financement par la T2A) l'ayant rendue possible.

Ces données ne comprennent ni les administrations médicamenteuses, ni les actes de biologie médicale intra-hospitaliers. L'utilisation de bases de données de grande dimension comprenant ces types de données est pourtant d'un intérêt particulier, la création de bases de données inter-hospitalières accueillant des données issues de systèmes d'information différents est donc nécessaire. A ce stade, la puissance statistique attendue de telles bases inter-hospitalières reste néanmoins très inférieure à celle des bases de données administratives correspondant aux données PMSI.

En synthèse, l'exploitation des bases de données hospitalières peut s'envisager selon deux axes : d'une part, des bases hospitalières administratives de très grande dimension mais ne comprenant que les données du PMSI et d'autre part des bases hospitalières de moins grande dimension mais plus riches en termes de types de données.

1.2.2.1.1.2 Développement de standards en santé

Dans le but de faire progresser l'interopérabilité syntaxique et sémantique des données en santé, plusieurs standards ont été proposés. De façon schématique, un standard américain « Health Level Seven » (HL7) [91] et un standard européen « CEN EN13606 » [92] sont en concurrence. Il est important de noter que le standard européen est un sous-ensemble des spécifications du standard openEHR [93]. Ces standards permettent principalement de construire des messages afin d'échanger des données entre des systèmes d'information différents reposant sur des modèles de données variés.

1.2.2.1.1.3 Définition de modèles de données communs en pharmaco-épidémiologie

Plusieurs projets de recherche en pharmaco-épidémiologie ont proposé des modèles de données compatibles avec leur problématique scientifique. Ainsi, plusieurs projets ont proposé des modèles de données communs [94] dont le projet PSIP [89] (Patient Safety through Intelligent Procedures in medication).

1.2.2.1.1.4 Modèles de données supportant les données utilisées dans cette thèse

Nous présentons plus spécifiquement dans cette partie les modèles de données utilisés dans cette thèse. Les données de notre hôpital partenaire sont structurées selon le modèle proposé dans le cadre du projet PSIP [89]. Une version simplifiée de ce modèle de données est présentée sur la Figure 4.

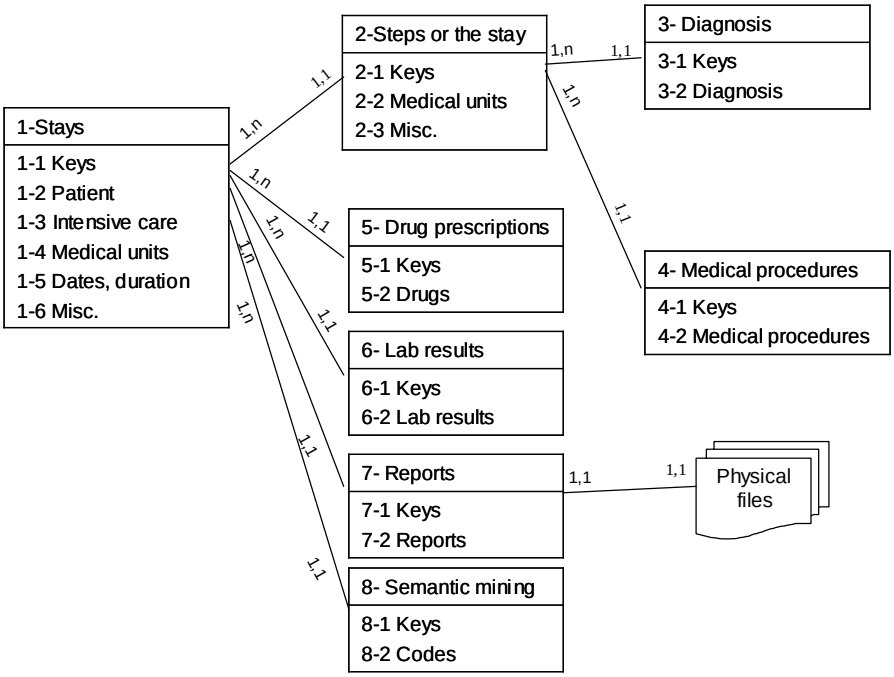


Figure 4 - Représentation simplifiée du modèle de données du centre hospitalier partenaire

Dans le cas de la base nationale des données du PMSI, le modèle de données retenu est d'une conception différente. Il repose sur un schéma simplifié comprenant une première table « épisodes » contenant une ligne par séjour avec un identifiant unique de séjour et une seconde table « observations » construite sur le modèle entité-attribut-valeur. Cette table « observations » contient ainsi des informations relatives au séjour hospitalier telles que les codes d'actes CCAM ou les codes diagnostiques CIM-10 qui lui sont rattachés. Ces observations peuvent volontiers correspondre à des variables multivaluées ou des mesures répétées. L'intérêt du modèle entité-attribut-valeur utilisé réside dans le fait que la liste des types d'informations rattachées au séjour n'est pas fixée *a priori*, elle peut facilement être étendue à un nouveau type d'information puisque le type de l'information décrite est lui-même renseigné dans une colonne d'attribut. Ce modèle est amplement utilisé en e-commerce : tandis qu'un T-shirt sera caractérisé par sa couleur et sa taille, une paire de chaussures le sera également par d'autres attributs (lacets/scratches/boutons/rien, derby/richelieu/autre, pointure, matière, etc.), et ces attributs ne doivent pas être fixés par avance lors de la conception du modèle de données : le site de e-commerce pourrait très bien dans un second temps référencer des lecteurs MP3.

Ensuite, les valeurs renseignées pour chacun des attributs sont volontairement enregistrées selon un format proche des données brutes, la nature de ce format étant renseignée dans une colonne dédiée. Par exemple, la racine du RSA contenant l'âge, le sexe ou encore le GHM est stockée de façon brute sans découpage préalable : la conséquence est que la valeur de l'âge n'est pas accessible d'emblée dans la base, sa valeur est obtenue lors d'une extraction par une interprétation à la volée de son format. Enfin, l'identifiant unique du patient (ANO) est une observation rattachée au séjour dans la table « observations ».

1.2.2.1.2 Qualité des données

La qualité de ces données administratives réutilisées doit être évaluée [95]. Deux études sur la base nationale des résumés de sortie anonymisés (RSA) illustrent la difficulté d'avoir une information de qualité. Ainsi, jusqu'à 10% d'erreurs de chaînage sont mises en évidence dans la base nationale [96]. De plus, les individus non assurés sociaux ne peuvent pas être suivis dans la base nationale. De même, la qualité des données de mortalité hospitalière est critiquée. Dans une étude récente, D. Blum [97] a ainsi observé que 4,3% des séjours sans décès comportaient néanmoins des codes CIM-10 traceurs de décès, comme par exemple le code « R98 : Décès sans témoin ».

Il a également observé que 3% des patients décédés d'après leur mode de sortie étaient néanmoins ré-hospitalisés ultérieurement. Pour ce qui est des données de décès, la liaison des bases SNIIRAM-PMSI et celle du CépiDc pourrait néanmoins corriger cette approximation.

Plus généralement, la qualité des bases de données administratives est évaluée par la confrontation de ces dernières avec les registres plus classiques [98–105]. La perspective d'utiliser ces bases pour certaines surveillances infectieuses n'est pas pleinement satisfaisante [106] à l'exception des études s'intéressant aux tendances de certaines infections [107]. Ce dernier point rejoint une hypothèse implicite essentielle des études

réutilisant des données : ces données pourraient être davantage pertinentes pour les études analytiques comparativement aux études descriptives. En effet, il est possible que le biais issu des erreurs de codage [108] soit un biais non différentiel : des définitions différentes d'un même évènement indésirable peuvent certes fausser l'incidence ou la prévalence d'un évènement (épidémiologie descriptive), mais ne pas modifier profondément la nature des couples médicament-évènement identifiés par la méthode et donc ne pas influencer la performance de cette dernière dans une étude d'épidémiologie analytique [109]. Enfin, précisons que des méthodes ont été proposées pour prendre en compte ce biais [110].

Afin de remédier aux erreurs de codage, de nombreux algorithmes permettant d'identifier les cas d'intérêt ont été proposés et/ou comparés [111–115]. Certains de ces algorithmes peuvent reposer sur les codes utilisés pour la facturation mais ils peuvent également s'appuyer sur les courriers hospitaliers ou les médicaments administrés au patient. De façon analogue, plusieurs études cherchent à transcrire certains scores selon le système de codage utilisé pour les bases de données administratives afin de pouvoir les utiliser à grande échelle [116–119]. Certains scores ont même été d'emblée construits à partir de ces bases de données administratives [120].

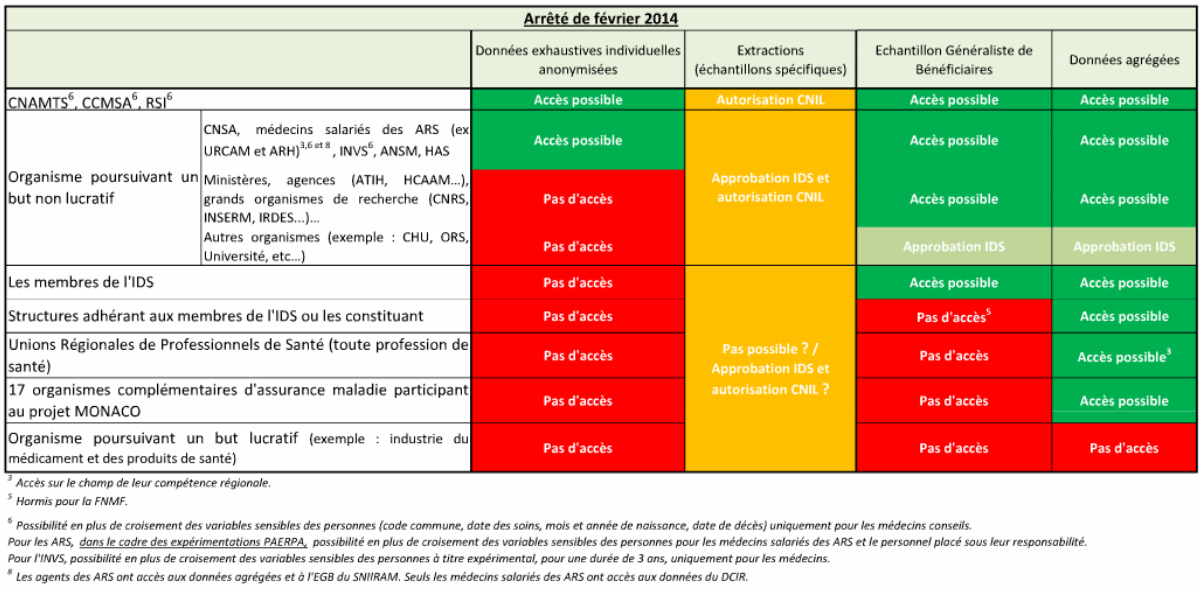
Enfin, plusieurs cadres méthodologiques ont été proposés pour assurer un processus de contrôle qualité sur les données administratives, notamment dans une perspective de réutilisation en pharmaco-épidémiologie [121].

1.2.2.2 Contraintes réglementaires pour la réutilisation de données médicales

Les autorisations obtenues pour l'exploitation des données réutilisées dans cette thèse ont été présentées dans la partie « 1.2.1.5 Présentation et comparaison des données réutilisées dans cette thèse ». Nous présentons maintenant plus généralement le cadre réglementaire d'accessibilité aux données de la base SNIIRAM-PMSI incluant le CepiDC. L'Institut des Données de Santé (IDS) jour un rôle d'accompagnement pour les demandes d'utilisation des grandes bases de données en France.

La réglementation de l'accès aux données du SNIIRAM est définie dans un arrêté [122] du ministère des affaires sociales et de la santé transmis pour avis à la CNIL.

Ces données font l'objet d'une analyse récente en France et plusieurs simplifications réglementaires facilitant leur accès sont encore à l'étude, notamment par la Commission Open Data en Santé [85]. Leur accès est coordonné par l'Institut des Données de Santé (IDS). L'analyse de ces données fait partie des missions de l'Agence Nationale de Sécurité du Médicament et des Produits de Santé (ANSM) et de l'Institut de Veille Sanitaire (InVS), la mise en place d'une analyse effective de ces bases de données relève donc de leur responsabilité. Ces deux agences de santé ainsi que la Haute Autorité en Santé (HAS) peuvent exploiter les données des dix années antérieures. Les possibilités d'accès aux données du SNIIRAM-PMSI sont détaillées sur la Figure 5 [123].



Arrêté de février 2014					
		Données exhaustives individuelles anonymisées	Extractions (échantillons spécifiques)	Echantillon Généraliste de Bénéficiaires	Données agrégées
CNAMTS[6], CCMSA[6], RSI[6]		Accès possible	Autorisation CNIL	Accès possible	Accès possible
Organisme poursuivant un but non lucratif	CNSA, médecins salariés des ARS (ex URCAM et ARH)[3,6 et 8], INVS[6], ANSM, HAS	Accès possible	Approbation IDS et autorisation CNIL	Accès possible	Accès possible
	Ministères, agences (ATIH, HCAAM…), grands organismes de recherche (CNRS, INSERM, IRDES…)…	Pas d'accès		Accès possible	Accès possible
	Autres organismes (exemple : CHU, ORS, Université, etc…)	Pas d'accès		Approbation IDS	Approbation IDS
Les membres de l'IDS		Pas d'accès	Pas possible ? / Approbation IDS et autorisation CNIL ?	Accès possible	Accès possible
Structures adhérant aux membres de l'IDS ou les constituant		Pas d'accès		Pas d'accès[5]	Accès possible
Unions Régionales de Professionnels de Santé (toute profession de santé)		Pas d'accès		Pas d'accès	Accès possible[3]
17 organismes complémentaires d'assurance maladie participant au projet MONACO		Pas d'accès		Pas d'accès	Accès possible
Organisme poursuivant un but lucratif (exemple : industrie du médicament et des produits de santé)		Pas d'accès		Pas d'accès	Pas d'accès

[3] Accès sur le champ de leur compétence régionale.
[5] Hormis pour la FNMF.
[6] Possibilité en plus de croisement des variables sensibles des personnes (code commune, date des soins, mois et année de naissance, date de décès) uniquement pour les médecins conseils.
Pour les ARS, dans le cadre des expérimentations PAERPA, possibilité en plus de croisement des variables sensibles des personnes pour les médecins salariés des ARS et le personnel placé sous leur responsabilité.
Pour l'INVS, possibilité en plus de croisement des variables sensibles des personnes à titre expérimental, pour une durée de 3 ans, uniquement pour les médecins.
[8] Les agents des ARS ont accès aux données agrégées et à l'EGB du SNIIRAM. Seuls les médecins salariés des ARS ont accès aux données du DCIR.

Figure 5 - Accessibilité des données SNIIRAM-PMSI [123]

La présente section achève la présentation des principales bases de données réutilisables en France ainsi que les contraintes techniques et réglementaires inhérentes à cette réutilisation. Dans la partie suivante « 1.3 Effets indésirables liés à la prise d'un médicament ou à la pose d'un dispositif médical implantable », nous présenterons la finalité médicale de cette thèse pour la réutilisation des données hospitalières présentées dans la partie « 1.2.1.5 Présentation et comparaison des données réutilisées dans cette thèse », à savoir la mise en évidence d'effets indésirables liés à la prise d'un médicament ou à la pose d'un dispositif médical.

1.3 Effets indésirables liés à la prise d'un médicament ou à la pose d'un dispositif médical implantable

Les effets indésirables médicamenteux (« Adverse Drug Reaction » en anglais) et les événements iatrogènes médicamenteux (« Adverse Drug Event » en anglais) sont impliqués dans près de 100 000 décès aux Etats-Unis chaque année [124,125] et pourraient être impliqués dans 5% des décès de patients hospitalisés [126] en Suède. En France, une enquête nationale a révélé que l'incidence au cours de l'hospitalisation des événements iatrogènes médicamenteux graves est de 7,6 pour 1000 jours d'hospitalisation [127] ; parmi ceux-là, 3 événements iatrogènes médicamenteux graves pour 1000 jours d'hospitalisation pourraient être évités.

1.3.1 Définitions de l'effet indésirable

1.3.1.1 Cas du médicament

1.3.1.1.1 Définition de l'effet indésirable médicamenteux (ADR en Anglais)

L'OMS définit les effets indésirables médicamenteux comme « une réaction nocive et non voulue à un médicament, se produisant aux doses utilisées chez l'homme pour la prophylaxie,

le diagnostic ou le traitement » [128]. Cette définition se réfère à des effets qui se produisent dans un cadre thérapeutique « normal » en termes de dosage et exclut ainsi les erreurs d'administration (erreur de dose, de mode d'administration ou de site d'injection par exemple).

1.3.1.1.2 Définition de l'événement iatrogène médicamenteux (ADE en Anglais)

Un événement iatrogène médicamenteux peut être défini comme « un dommage résultant de l'intervention médicale liée à un médicament » (par opposition à un dommage résultant d'une « affection sous-jacente du patient ») [129]. Ainsi, les événements iatrogènes médicamenteux comprennent à la fois les effets indésirables médicamenteux et les erreurs de prescription ou d'administration. Connaître la proportion d'évènements iatrogènes médicamenteux liés à des erreurs de prescription ou d'administration est essentiel d'un point de vue épidémiologique, car une partie de ces évènements pourraient en théorie être évités. Pour cette raison, nous avons choisi d'étudier les événements indésirables en général. Pour lever toute ambiguïté concernant l'acronyme utilisé, nous utiliserons l'acronyme EIM pour désigner les évènements iatrogènes médicamenteux (ADE en Anglais) au cours de ce travail.

Selon la FDA, un EIM est considéré comme grave lorsqu'il aboutit à l'une des conséquences suivantes [130] :

- un décès
- une situation de danger pour la vie
- l'hospitalisation (initiale ou prolongée)
- un handicap ou des lésions permanentes
- une anomalie ou malformation congénitale
- une intervention pour prévenir une déficience ou dommage permanents.

Notons toutefois que la notion d'intervention, dernier point de la liste ci-dessus, est assez floue. A notre connaissance, cette définition pourrait inclure par exemple l'arrêt initialement non prévu d'un médicament ou l'introduction d'un antidote.

1.3.1.2 Cas du dispositif médical

Dans le cas du dispositif médical, nous commençons par le définir et présenter son cadre réglementaire dans une première partie avant de définir l'effet indésirable qui peut lui être associé dans une seconde partie.

1.3.1.2.1 Définition et cadre réglementaire du dispositif médical

1.3.1.2.1.1 Définition du dispositif médical

Selon le Code de la Santé Publique, on entend par dispositif médical (DM) « tout instrument, appareil, équipement, matière, produit destiné à être utilisé chez l'homme à des fins médicales. Les DM qui sont conçus pour être implantés et qui dépendent d'une source d'énergie sont dénommés dispositifs médicaux implantables actifs (DMIA). » La définition des DM englobe des produits très divers dont les caractéristiques intrinsèques, les indications

et les conditions d'utilisation peuvent être très variables. Les DM sont regroupés en 4 classes en France selon leur niveau de risque. Ces classes sont :

- Classe 1 : faible degré de risque
- Classe 2a : degré moyen de risque
- Classe 2b : potentiel élevé de risque
- Classe 3 : potentiel très sérieux de risque (comprend les DM implantables actifs)

1.3.1.2.1.2 Cadre réglementaire des DM

1.3.1.2.1.2.1 Cadre actuel de commercialisation du DM : le marquage CE

La première étape en vue d'une mise à disposition d'un DM est l'obtention du marquage CE. Le cadre français est hérité du cadre réglementaire européen régi par trois directives (une sur les DMIA, une sur les DM de diagnostic in vitro et une sur les « autres DM ») complétées depuis mars 2010 (directive 2007/47/CE [131]) par une directive complémentaire qui rend nécessaire l'apport de données cliniques dans le cadre de la procédure de certification.

1.3.1.2.1.2.2 Cas des DM non stériles ou n'ayant pas de fonction de mesure

Ces DM sont auto-certifiés par le fabriquant.

1.3.1.2.1.2.3 Cas de tous les autres DM

Dans tous les autres cas, l'intervention d'un organisme notifié, choisi parmi ceux figurant sur la liste de la commission européenne, est nécessaire. Le certificat de conformité délivré par l'organisme notifié est valable 5 ans au maximum et renouvelable et, depuis l'avènement dans le droit français de la directive complémentaire en mars 2010 [132], il est nécessaire d'apporter dans le cadre de cette procédure de certification des données cliniques :

(i) « La démonstration de la conformité aux exigences essentielles doit inclure une évaluation clinique. » (Annexe I - 6 bis)

(ii) « En règle générale, la confirmation du respect des exigences concernant les caractéristiques et performances […] dans des conditions normales d'utilisation d'un dispositif ainsi que l'évaluation des effets indésirables et du caractère acceptable du rapport bénéfice/risque […] doivent être fondées sur des données cliniques. […] L'évaluation de ces données, dénommée l'évaluation clinique, doit, en tenant compte, le cas échéant, des normes harmonisées pertinentes, suivre une procédure définie et fondée au plan méthodologique […] » (Annexe X - 1 – 1.1)

(iii) En théorie, depuis 2010 : l'essai clinique est la règle pour les DMIA ainsi que pour les DM de classe III, « sauf à justifier de pouvoir y déroger ». La charge de la preuve est renversée, le chapitre clinique devant désormais être systématiquement documenté dans tout dossier de marquage CE d'un DM. Pour les autres DM certifiés, il est requis au moins une « démonstration d'équivalence » entre le dispositif à évaluer et celui objet des données

cliniques disponibles (en théorie essai mais peu clair en pratique), ces démonstrations reposent alors parfois sur des comparaisons indirectes.

On note que l'ANSM n'intervient pas dans le marquage CE ; l'ANSM n'intervient que pour autoriser et assurer le suivi des essais thérapeutiques ainsi que pour contrôler les organismes de certification. Le rôle de contrôle qu'elle a pour le médicament est d'une certaine façon délégué à des structures privées.

1.3.1.2.2 Définition de l'effet indésirable concernant un dispositif médical

Le cadre général de suivi des évènements indésirables liés aux dispositifs médicaux mis sur le marché est celui de la matériovigilance.

Deux types d'effets indésirables peuvent être distingués dans le code de la santé publique [133] :

(i) les effets indésirables graves : « est considéré comme incident ou risque d'incident grave tout incident ou risque d'incident mettant en cause un dispositif médical ayant entraîné ou susceptible d'entraîner la mort ou la dégradation grave de l'état de santé d'un patient, d'un utilisateur ou d'un tiers :

- décès du patient ou menace du pronostic vital,
- invalidité ou incapacité permanente ou importante,
- nécessité d'hospitalisation ou prolongation d'hospitalisation,
- toute circonstance nécessitant une intervention médicale ou chirurgicale,
- survenue d'une anomalie ou malformation congénitale. »

(ii) Les autres événements indésirables sont les suivants :

- « Réaction nocive et non voulue se produisant lors de l'utilisation d'un dispositif médical conformément à sa destination,
- Réaction nocive et non voulue résultant d'une utilisation d'un dispositif médical ne respectant pas les instructions du fabricant,
- Tout dysfonctionnement ou toute altération des caractéristiques ou des performances d'un dispositif médical,
- Toute indication erronée, omission et insuffisance dans la notice d'instruction, le mode d'emploi ou le manuel de maintenance. »

1.3.2 Méthodes courantes d'étude des effets indésirables

Nous présentons maintenant les méthodes courantes d'identification des effets indésirables. Ces méthodes concernent principalement les médicaments et les EIM, sans être toutefois spécifiques des médicaments.

Deux aspects méthodologiques doivent être différenciés : d'une part les méthodes qui permettent de connaître la nature des EIM survenant pour un médicament donné, et d'autre

part les méthodes permettant d'évaluer l'incidence (ambulatoire ou hospitalière) d'EIM déjà connus.

1.3.2.1 Méthodes permettant de connaître la nature des EIM pour un médicament donné

Quatre méthodes principales peuvent être recensées par ordre chronologique de survenue au cours de la vie du médicament :

(i) Tout d'abord, certains EIM peuvent être mis en évidence lors d'essais thérapeutiques. En effet, en plus de son efficacité, la tolérance du médicament est systématiquement évaluée. Néanmoins, la rareté des EIM est telle qu'une part importante des types d'EIM du médicament concerné par l'essai n'est pas connue au moment de la mise sur le marché du médicament.

(ii) Ensuite, au cours de la vie du médicament (en post-marketing), tout praticien confronté à un évènement qu'il considère être un EIM rare ou grave est censé rapporter cet évènement au centre régional de pharmacovigilance. Ainsi les centres de pharmacovigilance et l'ANSM sont informés des EIM potentiels. A ce stade, ce sont des cas qui sont déclarés et ces cas font l'objet d'une validation individuelle selon des critères dits « d'imputabilité ». Plusieurs algorithmes permettent d'imputer l'anomalie observée à la prise d'un médicament. Ces algorithmes sont basés sur des questionnaires et estiment la vraisemblance de survenue d'un EIM impliquant le médicament étudié. Les algorithmes de Naranjo [134], Kramer [135] et Bégaud [136] peuvent être évoqués dans ce contexte. Le niveau de preuve associé à ces validations individuelles de cas est néanmoins réduit.

(iii) Puis, la base de pharmacovigilance peut être analysée par des méthodes dites de « disproportionnalité » afin de mettre en évidence des associations médicament-évènement surreprésentées dans la base ; nous détaillons ces méthodes ci-après dans la partie « 1.4.3.1 Transposition dans les bases observationnelles des méthodes dites de disproportionnalité ». Il est important de comprendre que ces méthodes exploitent le fait que l'ensemble des médicaments prescrits au moment de la survenue de l'évènement, et non le(s) seul(s) médicament(s) imputé(s), sont recensés dans les bases de pharmacovigilance.

(iv) Enfin, de véritables études pharmacoépidémiologiques peuvent être conduites afin de confirmer l'association entre un médicament et un effet observé, l'intérêt de bases observationnelles dans ce contexte est présenté dans les parties « 1.4.1 Place de la réutilisation de données observationnelles dans la mise en évidence d'effets indésirables médicamenteux » et « 1.4.2 Place de la réutilisation de données observationnelles pour le suivi des dispositifs médicaux. » Ces études peuvent être conduites sur des bases médico-administratives et plusieurs travaux de cette thèse entrent directement dans cette quatrième voie d'identification. Les méthodes pouvant être utilisées sur des bases de données observationnelles sont détaillées dans la partie « 1.4.3 Types d'études permettant la mise en évidence d'évènements indésirables à partir des bases de données observationnelles ».

1.3.2.2 Méthodes permettant d'évaluer l'incidence d'effets indésirables connus à l'hôpital

Dès lors qu'un effet indésirable a été mis en évidence pour un médicament, il est utile de pouvoir estimer l'incidence ou le taux d'incidence de cet effet indésirable. La base nationale de pharmacovigilance devrait théoriquement permettre l'estimation de l'incidence des EIM survenant en ambulatoire et/ou à l'hôpital. Néanmoins, cela n'est que partiellement le cas. En effet, il est décrit que les déclarations (pourtant obligatoires) souffrent d'une participation insuffisante des praticiens [137,138]. Ainsi, l'enquête nationale sur les événements indésirables liés aux soins (ENEIS) réalisée en France se base sur une revue systématique experte d'un échantillon représentatif (selon un sondage en grappes) de dossiers de séjours hospitaliers. Ce travail fastidieux est néanmoins la seule façon jusqu'ici de produire cette estimation d'incidence.

Toujours dans cette perspective d'estimation de l'incidence des EIM, plusieurs projets dont le projet PSIP (Patient Safety through Intelligent Procedures in medication) ont développé des méthodologies permettant d'automatiser au moins pour partie la détection des EIM.

Le projet PSIP [139] est un projet de recherche européen financé par l'*European Research Council* dans le cadre du septième programme cadre (FP7), un appel à projet en informatique médicale et technologies de la santé. Il a débuté en janvier 2008 et s'est étalé sur une période de 40 mois. Celui-ci était coordonné par le Centre Hospitalier Régional Universitaire (CHRU) de Lille et associait 13 partenaires européens se répartissant 13 workpackages. Ces partenaires étaient académiques (Université d'Aalborg en Suède, AUTH en Autriche, Université de Thessalonique en Grèce), hospitaliers (CHRU de Lille, CHRU de Rouen, CH de Denain, Hôpital d'Ushate en Bulgarie, Hôpitaux de Frederiksberg et Nordsjaelland au Danemark) et industriels (Oracle™, IBM™, Medasys®, Vidal®). Ce projet s'est poursuivi par le projet PSIP-EVAL à partir de janvier 2012.

L'idée initiale était d'exploiter rétrospectivement des bases de données hospitalières en les détournant de leur fonction initiale, en vue de détecter puis de prévenir la survenue d'EIM. Les bases hospitalières utilisées contenaient les administrations médicamenteuses, les résultats de biologie médicale, le codage des actes et diagnostics selon le PMSI ou ses équivalents danois et bulgares, les comptes-rendus hospitaliers ainsi que des données démographiques et administratives. PSIP s'était fixé pour objectif de proposer des méthodes innovantes pour détecter de façon automatisée les EIM en développant des règles de détection, puis produire des connaissances épidémiologiques et des méthodes de prévention de ces EIM.

La génération de règles de détection d'évènements iatrogènes médicamenteux s'est déroulée selon deux cadres méthodologiques : tout d'abord, nous nous sommes attachés à développer des règles fabriquées à dire d'expert, c'est à dire que nous avons traduit dans un langage informatique des règles issues de bases pharmacologiques telles que VIDAL® ou Thériaque® ; ensuite, des règles ont également été construites à partir des bases de données médicales elles-mêmes en utilisant des méthodes de fouille de données [140,141]. L'objectif dans ce second cadre méthodologique était de détecter des EIM complexes et parfois des EIM

combinés qu'un expert n'aurait pas nécessairement identifiés : des modèles prédisant la survenue d'anomalies biologiques ont été construits par fouille de données. Nous avons notamment combiné des notions de médicament administré, médicament suspendu (ce qui était inhabituel jusqu'alors), pathologie chronique du patient, données démographiques, anomalie biologique du patient et notion de parcours (par exemple admission par les urgences, naissance, etc.).

Plusieurs méthodes ont été proposées et une liste finale de 236 règles intégrant des règles construites pour partie par fouille de données a été proposée. Un des résultats du projet PSIP a été le développement d'un outil de revue des séjours sélectionnés automatiquement par les règles de détection. Cet outil dénommé « ADE Scorecards » [142] s'attache à présenter de façon ergonomique les séjours automatiquement détectés par les règles comme ayant potentiellement présenté un évènement iatrogène médicamenteux. L'un des travaux de cette thèse consiste à évaluer la pertinence d'une liste de règles de détection pour estimer de façon automatisée les EIM à type d'hyperkaliémie.

1.4 Mise en évidence d'effets indésirables par la réutilisation de bases de données observationnelles

Les validations individuelles de cas décrites dans la partie « 1.3.2.1 Méthodes permettant de connaître la nature des EIM pour un médicament donné) » devraient idéalement, sauf cas particulier [143], être confirmées par de véritables études épidémiologiques telles que décrites dans cette même partie. Dans ce contexte, les bases hospitalières et administratives semblent avoir toute leur place pour ce type d'étude [16,144].

1.4.1 Place de la réutilisation de données observationnelles dans la mise en évidence d'effets indésirables médicamenteux

En 2007, le congrès américain a adopté le « Food and Drug Administration Amendment Act » qui a demandé la mise en place d'une surveillance active post-commercialisation s'appuyant sur l'exploitation de données observationnelles de 100 millions de patients à partir de 2012. Il est envisagé que ce système de surveillance puisse utiliser des méthodes statistiques évoluées. Les données observationnelles privilégiées ont été les données administratives de facturation et les dossiers patients informatisés [145,146].

Ce type de recherche active d'un effet indésirable post-commercialisation en réutilisant des bases de données médico-administratives a également pris forme en France ponctuellement. Un exemple récent a concerné la pioglitazone : cet antidiabétique a été retiré du marché en 2011 suite à une étude épidémiologique réalisée sur la base de données SNIIRAM-PMSI [147,148]. Nous présentons ici brièvement le contenu de l'étude menée sur cette base de données. Parmi les patients diabétiques, le sous-groupe de patients recevant de la pioglitazone a été comparé aux patients diabétiques n'en recevant pas : ces deux sous-groupes de patients avaient été identifiés au sein de la base de données SNIIRAM des remboursements de médicaments achetés en ambulatoire en pharmacie. Ensuite, l'évènement d'intérêt était le fait d'avoir une hospitalisation dont le diagnostic principal (DP du PMSI) – correspondant au motif d'hospitalisation depuis mars 2009 – était un cancer. Un excès de cas de cancers de la

vessie a alors été mis en évidence dans le groupe de patients traités par pioglitazone, ce résultat venant confirmer une hypothèse préalable à la réalisation de l'étude. Quelques études analogues sont actuellement conduites par l'ANSM et concourent à la surveillance des hémorragies survenant suite à la prise des nouveaux anticoagulants ou encore la survenue de Sclérose En Plaques (SEP) secondairement à certaines vaccinations.

De façon analogue, l'utilisation d'informations hospitalières objectives telles que celles contenues dans les dossiers patients informatisés, semble utile pour assurer cette surveillance. Les informations contenues dans ces systèmes (consommations médicamenteuses, résultats de biologie médicale et données du PMSI) ont été collectées en routine et de façon exhaustive. L'exploitation de ces données apparaît pertinente en particulier pour le suivi d'EIM de survenue aiguë durant l'hospitalisation. D'une manière générale, la recherche systématique dans les bases administratives et hospitalières de potentiels EIM pourrait bénéficier, au moins dans une phase exploratoire, des méthodes de fouille de données, ces méthodes sont détaillées dans la partie « 1.4.3.4 Construction de modèles prédictifs par fouille de données ».

1.4.2 Place de la réutilisation de données observationnelles pour le suivi des dispositifs médicaux

Dans le cas des dispositifs médicaux (DM), nous allons présenter en deux temps l'intérêt potentiel de l'analyse de données observationnelles pour leur suivi. Dans une première partie, nous allons présenter les spécificités et les difficultés méthodologiques de l'évaluation des dispositifs médicaux avant leur commercialisation ce qui nous permettra, dans une seconde partie, de présenter l'intérêt particulier de l'analyse de données observationnelles (recueillies après le début de la commercialisation) dans ce contexte.

1.4.2.1 Spécificités méthodologiques de l'évaluation des DM

L'analyse de l'efficacité des DM de classe III et des DMIA devrait en théorie (et idéalement) se calquer sur les essais cliniques des médicaments. Ainsi, ils devraient être prospectifs, randomisés, en aveugle (patients, cliniciens et analystes ce qui peut être vu comme une forme de triple aveugle), avec groupe contrôle simultané.

En pratique, la méthode des essais cliniques randomisés est difficile à appliquer aux DM. Cette difficulté, ainsi qu'un tissu économique très différent (les entreprises fabriquant des DM sont très nombreuses et souvent de très petite taille) est probablement à l'origine du déficit actuel de leur évaluation clinique.

1.4.2.1.1 Problématique et éléments de réponse

Il existe d'abord une très grande variabilité parmi l'ensemble des types de DM donc il est impossible de définir un cadre méthodologique détaillé compatible avec l'hétérogénéité du monde des DM.

Il est possible de recenser plusieurs caractéristiques fréquentes des DM : nous présentons à chaque fois les méthodes envisagées pour surmonter la difficulté évoquée (ces méthodes

peuvent concerner la phase d'évaluation antérieure à la commercialisation ou la phase de suivi postérieure à celle-ci).

1.4.2.1.1.1 Le DM a des propriétés techniques (et peut incorporer un principe actif)

Il faut dissocier la performance technique du bénéfice clinique lors de l'évaluation du DM. La performance technique est un préalable indispensable, mais le bénéfice clinique n'en est pas pour autant acquis : la performance technique ne valide pas en soi la pertinence thérapeutique d'une pose de DM.

1.4.2.1.1.2 Le DM peut être posé secondairement à un acte chirurgical

La réalisation d'un acte chirurgical induit plusieurs biais méthodologiques pouvant influer le résultat brut de l'étude :

- les compétences de l'opérateur et la compliance à l'intervention décrite dans le protocole (effet opérateur)
- le niveau d'équipement, le nombre d'actes et la post-intervention réalisés dans le centre (effet centre)
- la proportion de rejets par le patient de l'intervention évaluée est supérieure dans ce contexte

Pour ce qui des effets opérateur et centre, voici une liste de points méthodologiques à respecter :

- l'intervention doit être standardisée (pour un acte chirurgical : standardiser l'anesthésie, l'acte chirurgical, les soins postopératoires, la rééducation, etc.) en décrivant précisément la procédure [149] puisqu'elle est une partie du traitement ; cela peut impliquer l'utilisation d'un mannequin d'apprentissage avant le début de l'étude puis la réalisation de films et d'audits en cours d'étude
- les courbes d'apprentissage des intervenants doivent être évaluées.

Sur le plan méthodologique, les analyses en cluster ont toute leur place dans ce contexte.

1.4.2.1.1.3 L'aveugle peut difficilement être maintenu (l'intervention chirurgicale est une cause fréquente de ce biais)

Pour des raisons assez évidentes, l'aveugle est plus difficile à réaliser et à maintenir [150,151] au cours des essais cliniques évaluant l'efficacité des DM, lorsque le procédé de référence n'est pas un DM similaire. Les pistes décrites pour essayer de le contenir sont réunies ci-dessous.

(i) L'aveugle du patient :

- le patient doit être maintenu en aveugle des hypothèses de l'étude
- le patient ne doit pas avoir de contact avec les participants du (des) bras différent(s)
- on peut simuler l'utilisation du DM en rééducation

- on peut simuler certaines interventions ou l'utilisation de certains DM (mais cela ne va pas sans poser une question éthique évidente) :
    - intervention sous anesthésie locale pour certains implants en ophtalmologie
    - évaluation de la stimulation du nerf vague (traitement de l'épilepsie pharmacorésistante) avec un stimulateur inactivé dans le groupe contrôle
    - utilisation de prothèses d'aspect identique

(ii) L'aveugle du médecin évaluateur :

- un médecin évaluateur distinct du médecin pratiquant la pose du DM pourrait intervenir en aveugle (on peut alors évoquer un « triple aveugle ») des hypothèses de l'étude pour évaluer le critère de jugement principal, ce point est néanmoins difficile à respecter lorsque l'évaluateur est responsable de la rééducation

La conséquence des difficultés décrites ci-dessus est le contrôle insuffisant de l'effet placebo, et par conséquent le risque d'augmentation artificielle de la taille de l'effet observé. Enfin, la qualité de l'aveugle peut également varier en fonction de la nature similaire ou non du traitement de référence.

1.4.2.1.1.4 Le traitement de référence n'est pas toujours facilement identifiable

Le traitement de référence est le meilleur traitement qui serait donné dans la situation clinique où le DM est indiqué. Il peut avoir une nature différente de celle du DM. Le DM sera ainsi comparé après une analyse bibliographique rigoureuse à :

- un autre DM
- un médicament
- une prise en charge chirurgicale
- une autre prise en charge (exemple : kinésithérapie)
- un placebo (cf. section ci-dessus sur l'aveugle)

On peut citer enfin plusieurs autres caractéristiques des DM n'ayant pas de solution méthodologique simple parmi lesquelles :

- l'évolution rapide des produits mis sur le marché
- le faible effectif de la population cible
- la durée de vie de certains DM supérieure à la durée de l'étude (notamment les DMIA)
- le coût unitaire élevé de certains DM

Citons ici les « Tracker Studies » [152] qui sont des essais cliniques randomisés conçus pour recueillir les données d'efficacité clinique d'une technologie innovante, susceptible d'évoluer rapidement et pour laquelle le rôle de l'opérateur est important. Le plan d'investigation clinique de ces études doit prévoir les ajustements méthodologiques et statistiques qui permettront de tenir compte des effets opérateur/apprentissage et des modifications apportées aux DM.

1.4.2.1.2 Orientations données par CONSORT et la CNEDiMTS face à l'insuffisance méthodologique actuelle des essais

Le groupe CONSORT [151] a proposé une mise à jour (par rapport au traitement médicamenteux) impliquant une modification de onze items (et l'ajout d'un item supplémentaire) regroupés en 5 sections et 22 items.

Les recommandations décrites par la CNEDiMTS (quant à la qualité des essais) évoquent plusieurs des biais évoqués ci-dessus, ils recommandent néanmoins le cadre général assez classique présenté sur la Figure 6 pour les essais cliniques préalables à la mise à disposition du DM.

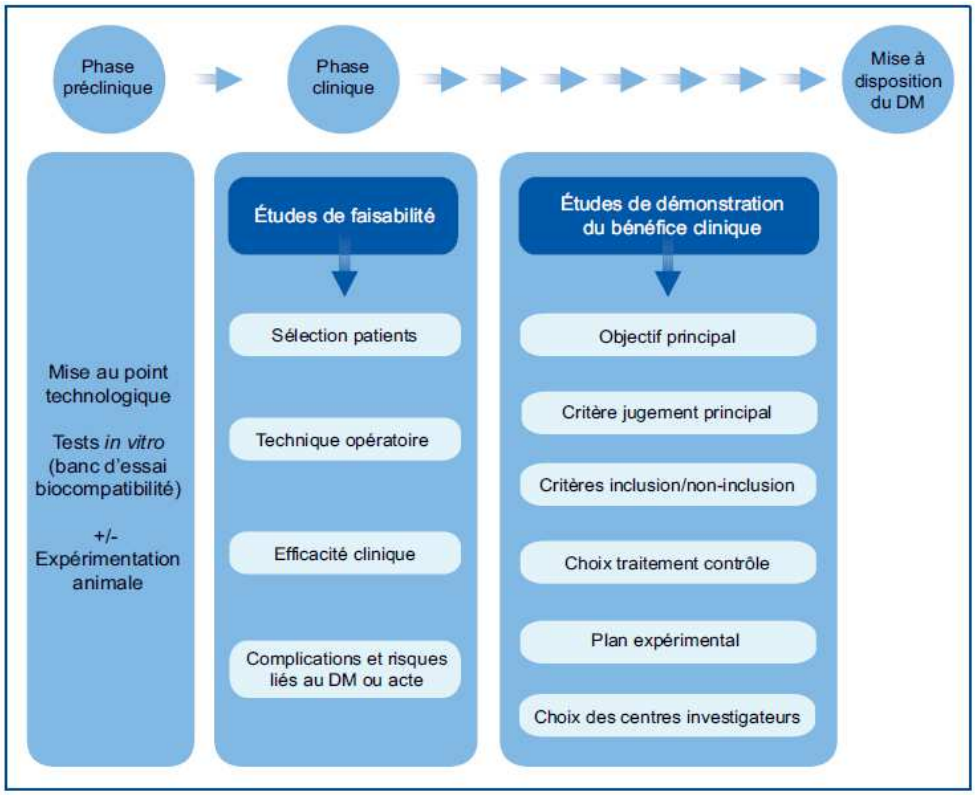


Figure 6 - Essais cliniques préalables à la mise à disposition du dispositif médical

La CNEDiMTS formule également plusieurs recommandations pour les produits pour lesquels elle a rendu un avis favorable, elle peut en effet demander des études complémentaires au cours de la vie du produit et/ou lors d'un renouvellement. Enfin, l'ANSM peut également « intervenir à tout moment de la vie du dispositif » en particulier en cas de déclaration d'incident(s) grave(s).

1.4.2.2 Rôle des études observationnelles pour les DMI

Devant les difficultés rencontrées pour mener des essais cliniques de qualité suffisante, il est parfois préférable d'appliquer une autre méthodologie et notamment celle des essais fondés sur l'observation avec des suivis adaptatifs. Ce type d'essai clinique est souvent proposé en

substitution ou en complément des essais cliniques randomisés. Il existe plusieurs types d'essais cliniques observationnels que l'on peut classer par niveau de preuve scientifique décroissant :

- Méthodes avec des cas et des témoins :
  - Essai clinique non randomisé avec témoin simultané ou antérieur,
  - Etude de cohorte prospective,
  - Etude cas-témoin rétrospectif,
- Méthodes avec des cas sans témoin :
  - Essai clinique de suivi de cas (registre de survie par exemple),
  - Série de cas consécutifs,
  - Simples rapports de cas.

Cette liste d'études épidémiologiques montre l'intérêt tout particulier des études observationnelles dans le cas des DM. Pour cette raison, plusieurs bases observationnelles sont administrées telles que la « base de données en épidémiologie et chirurgie thoracique (épithor) » ou « l'OBSERVAtoire des Pratiques en Urologie (OBSERVAPUR) » mais ces bases sont également décrites comme manquant de données de suivi en particulier.

Enfin, dans le cadre d'une mission commune d'information portant sur les DMI faisant suite au scandale des prothèses PIP, le Professeur Eric Vicaut a évoqué lors de son audition au Sénat [153] en mars 2012 l'idée d'une « plateforme d'évaluation des DM en lien avec les CIC et CIC-IT » et a de plus indiqué qu'il considérait la base nationale du PMSI comme un matériel intéressant pour les études observationnelles sur les DM.

1.4.3 Types d'études permettant la mise en évidence d'évènements indésirables à partir des bases de données observationnelles

L'utilisation de bases observationnelles implique l'utilisation d'une méthodologie statistique rigoureuse. En effet, en l'absence de tirage au sort, les caractéristiques moyennes du groupe des patients recevant un traitement peuvent être différentes des caractéristiques moyennes du groupe contrôle des patients ne recevant pas le traitement étudié. L'absence d'égalisation des contextes *a priori* rend indispensable le contrôle des nombreux facteurs de confusion potentiels.

Ce contrôle actif des facteurs de confusion dans ce contexte observationnel peut être envisagé selon plusieurs axes méthodologiques. Certains résultats présentés et retranscrits dans cette partie sont issus du projet OMOP [154], ce dernier s'étant attaché à comparer rigoureusement plusieurs méthodologies à la recherche de quatre types d'EIM (insuffisance hépatique aiguë, infarctus du myocarde, insuffisance rénale aiguë et saignement digestif haut) sur 10 jeux de données (4 bases de données administratives, 1 base de dossiers patients informatisés et 4 jeux de données simulés [155]) : chaque méthode s'attache à mettre en évidence une liste de 399 paires de « médicament-évènement gold standard » comprenant des vrais négatifs et des vrais positifs [156]. Ce projet de grande envergure a apporté des réponses rigoureuses empiriques à la nature des méthodes pouvant être envisagées pour l'exploitation des bases de données

observationnelles dans une perspective pharmaco-épidémiologique. Les résultats présentés dans cette partie ont ensuite été répliqués sur la base de données du projet EU-ADR [157].

Les méthodes présentées ici n'ont pas pour but le dénombrement d'EIM connus mais bien la mise en évidence de certains EIM survenant secondairement à une prise en charge thérapeutique. Ainsi, ces méthodes peuvent permettre d'une part la mise en évidence d'un signal dans une démarche exploratoire et d'autre part de conforter la pertinence d'un tel signal dans une démarche confirmatoire.

Les résultats finaux [158] de ce projet montrent l'intérêt tout particulier des méthodes utilisant le patient comme son propre témoin, que ce soient les cohortes (rétrospectives) en cross-over [159] ou les cas-témoins en cross-over [160]. D'autres méthodes d'intérêt présentées ici sont les cohortes [161] ou les cas-témoins [162] sans cross-over : ces méthodes font appel à des modèles de régression avec ajustement plus classiques, éventuellement par l'intermédiaire d'un score de propension. Ensuite, la transposition au sein des bases observationnelles des méthodes issues de l'analyse des bases de pharmacovigilance telles que la disproportionnalité [163] sera explorée. Enfin, des méthodes de fouille de données seront envisagées parmi lesquelles les arbres de classification, les réseaux de neurones ainsi que les algorithmes LGPS et LEOPARD [164].

Nous précisons ici que les EIM recherchés sont des évènements binaires : ils sont présents ou absents, sans notion de grade. Cette précision est donnée car la méthodologie d'un de nos travaux traitera d'effets quantitatifs. Ensuite, l'ensemble de ces méthodes s'envisage préférentiellement au sein du cadre de recommandations en épidémiologie proposé par STROBE [165].

Une difficulté des études pharmaco-épidémiologiques est le biais de survie [166–168]. De façon schématique, l'idée est que les patients doivent survivre un certain temps sans réaliser d'évènement avant d'être considérés comme appartenant au groupe traité. Ainsi, un biais est souvent introduit dans ces études par l'erreur d'affectation d'un patient réalisant par exemple l'évènement avant la fin de la période minimale d'exposition au traitement. Il semble nécessaire dans ce cas d'utiliser des modèles temps-dépendants afin de limiter ce biais [169]. L'étude d'expositions ponctuelles ayant un effet limité dans le temps semble moins exposée à ce biais.

Nous précisons de façon préalable que les méthodes exploitant les données textuelles non structurées ne sont pas abordées ici. Un choix *a priori* d'exploiter les données structurées est réalisé. Des travaux recherchant des signaux à travers les courriers hospitaliers, à travers les publications scientifiques ainsi qu'à travers les réseaux sociaux ou les forums médicaux sont une piste complémentaire de travail en pharmacovigilance que nous n'explorons pas ici. De même, des modèles multi-agents [170] ont été proposés pour assurer la gestion complexe des signaux générés par les méthodes décrites dans cette partie mais ils ne sont pas présentés dans le cadre de ce travail.

1.4.3.1 Transposition dans les bases observationnelles des méthodes dites de disproportionnalité

Les analyses de disproportionnalité représentent le principal cadre méthodologique pour l'analyse des bases de pharmacovigilance contenant les cas d'EIM spontanément déclarés. Ces analyses recherchent les couples médicament-évènement déclarés plus fréquemment que l'on ne s'attendrait à les observer par hasard. Une déclaration spontanée d'effet indésirable contient typiquement la date, l'âge, le sexe, la liste complète de médicaments auxquels le patient a été récemment exposé (et non uniquement le médicament auquel serait imputé l'effet) et une liste d'effets indésirables typés le plus souvent selon la classification MedDRA. Les principales méthodes d'analyse de disproportionnalité sont le Bayesian Confidence Propagation Neural Network (BCPNN) [171] utilisé par l'OMS (Uppsala Monitoring Centre) et le Gamma Poisson Shrinker (GPS) [172] utilisé par la Food and Drug Administration (FDA). Le Proportional Reporting Ratios (PRR) (utilisé au Royaume-Uni par la Medicines Control Agency) et le Reporting Odds Ratio (ROR) sont deux autres méthodes pouvant être citées. Le tableau de contingence 2*2 croisant le médicament et l'évènement indésirable retourne souvent un petit effectif pour la cellule correspondant à la présence du médicament ET de l'effet indésirable. Ce petit effectif a pour conséquence d'augmenter l'incertitude de la mesure d'association réalisée à partir de ce même tableau. L'inférence bayésienne des méthodes BCPNN et GPS apporte une réponse à cette difficulté.

Une évaluation originale de cette méthode est réalisée dans le cadre du projet OMOP [163] à partir des bases de données observationnelles. Dans cette étude, les auteurs miment dans un premier temps ce que seraient les déclarations spontanées générées par ces données longitudinales observationnelles, sous l'hypothèse que ces déclarations soient systématiquement réalisées. Les méthodes de disproportionnalité sont ensuite appliquées à la base théorique de déclarations constituée. Cette analyse permet d'une part de comparer ces méthodes de disproportionnalité (transposées aux données observationnelles) aux autres méthodes pharmaco-épidémiologiques et d'autre part d'apprécier plus généralement l'intérêt des bases de pharmacovigilance dans la mise en évidence d'associations médicament-évènement.

La majorité des aires sous la courbe calculées retrouvent des valeurs comprises entre 0,35 et 0,6 (il est rappelé qu'une aire sous la courbe de 0,5 correspond à une affectation aléatoire), à l'exception de certaines analyses retrouvant une aire sous la courbe à 0,7 pour l'effet insuffisance rénale aiguë et sur la base de données de General Electric™.

Les auteurs concluent dans ce cas que les méthodes de disproportionnalité ne permettent pas de discriminer correctement les vrais positifs et les vrais négatifs sur les données observationnelles analysées contrairement à ce qu'elles semblent pouvoir faire au sein des bases de pharmacovigilance comprenant les déclarations spontanées.

Nous avons présenté dans ce paragraphe la façon dont les méthodes de disproportionnalité telles que GPS pouvaient être utilisées à partir de bases de pharmacovigilance construites en mimant les déclarations systématiques depuis une base observationnelle. Nous verrons ci-

après dans la partie consacrée à « LGPS et LEOPARD » que ces méthodes peuvent également être appliquées directement au sein des bases observationnelles.

1.4.3.2 Designs d'études contrôlées

Les designs d'études présentés dans cette partie n'utilisent pas le patient comme son propre témoin (qui seront présentés dans la partie « 1.4.3.3 Designs d'études utilisant le patient comme son propre témoin »). Les deux designs d'études courants de ce type sont l'étude de cohorte et l'étude cas-témoins. Leur développement dans une perspective pharmaco-épidémiologique revêt néanmoins plusieurs spécificités que nous présentons dans ce paragraphe.

1.4.3.2.1 Etude de cohorte de type « nouvel utilisateur »

L'étude de cohorte de type « nouvel utilisateur » a pendant longtemps été considérée comme le principal design pour les études relatives à la sécurité du médicament. La définition de la période à risque [173] et la méthode d'ajustement (reposant préférentiellement sur le score de propension) sont les principaux paramètres de ces études. De façon schématique, deux groupes sont construits : le premier comprend les patients ayant reçu au moins une fois le traitement étudié et la date d'index correspond à la première prise du traitement ; le second comprend possiblement l'ensemble des patients de la base ne recevant jamais le traitement étudié.

Ce second groupe (contrôle) peut être défini selon l'une des méthodologies suivantes :

- ce groupe peut être composé des patients recevant le traitement le plus fréquemment utilisé pour la même indication que celle du traitement étudié,
- ce groupe peut être composé des patients recevant l'ensemble des autres traitements existant pour la même indication que celle du traitement étudié,
- ce groupe peut être composé des patients ayant un code diagnostique correspondant à l'indication du traitement étudié, en conservant uniquement les patients traités par des médicaments connus pour ne pas provoquer l'effet d'intérêt,
- ce groupe peut être composé des patients ayant un code diagnostique correspondant à l'indication du traitement étudié, en conservant tous les patients à compter de la date du code diagnostique,
- ce groupe peut être composé de tous les autres patients de la base ou d'un échantillon tiré au sort de ces patients.

De façon analogue, on peut s'intéresser aux complications survenant pour une pathologie donnée : dans ce cas, les patients sont inclus à la première date où le diagnostic d'intérêt est retrouvé. Le groupe contrôle construit à partir d'une base de données hospitalière (ne comportant pas de patients totalement indemnes) pourra être composé de patients ayant été hospitalisés pour des pathologies ponctuelles telles que par exemple la pose d'une prothèse de hanche ou une amygdalectomie. Une liste de ces pathologies ponctuelles neutres est retrouvée dans plusieurs articles exploitant exclusivement des bases de données hospitalières. Ce groupe

peut également être composé de tous les autres patients de la base ou d'un échantillon tiré au sort de ces patients.

Un délai suffisant avant la date « d'inclusion » est systématiquement fixé dans le groupe des patients ayant le traitement étudié afin de s'assurer qu'il s'agit véritablement d'un nouvel entrant pour ce traitement. Un délai identique minimal de suivi est fixé dans le groupe « contrôle ».

Comme évoqué ci-avant, les évènements d'intérêt sont possiblement des évènements temps-dépendants mais ces derniers sont souvent traités comme des variables binaires non censurées après avoir défini *a priori* un délai maximal de survenue. Ainsi, un modèle logistique et non un modèle de survie peut être utilisé pour ce type d'analyse. Toujours dans la construction de ce modèle, l'utilisation de variables d'ajustement permet de contrôler les caractéristiques des cohortes à « l'inclusion ». Il n'est pas rare que les effets d'intérêts soient pour partie redondants avec l'indication du traitement, auquel cas il peut exister un biais d'indication. Plusieurs des designs décrits ci-avant permettent la prise en compte pour partie de ce biais d'indication. La construction d'une variable d'ajustement traduisant la probabilité du patient de recevoir le traitement est un autre élément de réponse pouvant être apporté. Cette variable porte le nom de score de propension (ou variable instrumentale), elle constitue une variable agrégée d'ajustement, son caractère agrégé la rend particulièrement utile dans les études n'ayant pas la puissance statistique suffisante pour intégrer la liste complète des variables d'ajustement. Nous précisons en ce sens que cette variable agrégée doit autant que possible être « surajustée » dans le modèle final par les variables ayant participé à sa construction.

L'utilisation du score de propension est discutée par plusieurs auteurs et certains articles évoquent une sous-estimation de l'effet rendu [174] par la variable traitement dans un contexte d'ajustement par le score de propension, si l'effet obtenu est comparé à celui retrouvé dans le cadre d'un essai randomisé. Un travail plus récent contredit néanmoins ce résultat [175].

Plusieurs modèles sont décrits pour la construction du score de propension : (i) la méthode « high-dimensional propensity score algorithm » [176] sélectionnant un nombre défini de covariables associées à l'exposition et à l'évènement, (ii) l'implémentation proposée par Brookhart [177] sélectionnant un jeu de covariables associées à l'exposition exclusivement et (iii) l'approche « large scale Bayesian regression » [178] qui utilise toutes les covariables représentant les conditions à *baseline*, l'exposition à des traitements et les actes pour classifier le statut de l'exposition.

Dans certains cas, la méthode utilisée permet d'obtenir un résultat modeste avec des aires sous la courbe supérieures à 0,65 pour au moins une base de données (à l'exception de l'effet infarctus du myocarde). Cependant, dans de nombreux cas, la qualité prédictive est faible avec des aires sous la courbe égales à 0,50. Notons également que ce résultat ne réplique pas un résultat initial du projet OMOP [179].

Les auteurs concluent dans ce cas que les cohortes de type « nouvel utilisateur » donnent une qualité prédictive modeste. Les auteurs insistent sur l'impact majeur sur le résultat de la nature du groupe contrôle.

1.4.3.2.2 Etude cas-témoins

Les études cas-témoins sont largement utilisées pour la mise en évidence d'effets indésirables médicamenteux. Dans le cas de la réutilisation de données administratives, les études cas-témoins ont moins de place car ces bases de données administratives sont, du fait de leur exhaustivité, de véritables cohortes rétrospectives. Des études cas-témoins peuvent néanmoins s'envisager en complément de ces dernières dans le cadre d'études cas-témoins nichées dans la cohorte, afin de s'intéresser aux expositions possiblement associées à certains évènements rares.

La construction du groupe « témoins » se fait classiquement par l'appariement selon certaines caractéristiques telles que l'âge ou le sexe, ainsi que certains facteurs de confusion voulant être pleinement contenus. L'appariement est consommateur d'un nombre exponentiel de patients à chaque ajout de variable, la réutilisation de ces bases administratives de grande dimension apporte un confort dans cette étape. Il est courant de choisir 5 à 10 témoins pour un cas, le gain de puissance étant néanmoins réduit au-delà de 5 témoins par cas.

L'analyse statistique réalisée s'appuie ensuite sur une régression logistique conditionnelle ou sur une régression logistique bayésienne. Lorsque la régression logistique (conditionnelle) est utilisée, il est utile de pouvoir prendre en compte d'autres facteurs de confusion que sont par exemple le nombre de médicaments reçus par le patient, le nombre de visites ou encore certains scores de comorbidité. L'ajustement dans le cadre de la régression logistique conditionnelle ne peut concerner que des variables non constantes au sein d'une classe d'appariement, les variables constantes pour une classe d'appariement ne pouvant être étudiées qu’en interaction : la régression logistique conditionnelle diffère sur ce point du modèle linéaire mixte généralisé.

Les scores de comorbidité évoqués ci-avant tels que le score de Charlson [180] sont des scores prédisant la mortalité. Ils ont été construits historiquement sur des cohortes de petite dimension et comportant des informations cliniques plus fines que celles accessibles dans le cadre des données administratives. Plusieurs adaptations du score de Charlson [116,181,182] (dont certaines il y a plus de 20 ans [182]) et le score de Elixhauser [183] ont ensuite été proposés pour les bases de données administratives [184].

Les auteurs ayant expérimenté cette méthode [162] dans une finalité pharmaco-épidémiologique concluent que les études cas-témoins ont une contribution faible dans l'identification des paires médicament-effet indésirable.

1.4.3.2.3 Méthode LGPS et LEOPARD

Les méthodes Longitudinal Gamma Poisson Shrinker (LGPS) et Longitudinal Evaluation of Observational Profiles of Adverse events Related to Drugs (LEOPARD) ont été développées

spécifiquement pour cette tâche pharmaco-épidémiologique [185]. LGPS étend la méthode GPS en utilisant les temps d'exposition des personnes plutôt qu'un décompte ne prenant pas en compte le temps pour estimer le nombre d'évènements attendus.

LGPS compare le taux d'incidence d'un évènement pendant la période d'exposition à un taux d'incidence calculé à partir de l'ensemble des sujets afin de calculer un « *Incidence Ratio Rate* » (IRR). L'estimateur naïf IRR peut être combiné avec d'autres informations afin de prendre en compte les expositions concomitantes à d'autres médicaments. Cependant, la réalisation d'un tel ajustement est rendue difficile par le nombre très grand de médicaments potentiels et expose ainsi au sur-ajustement. Afin de remédier à cette difficulté, des méthodes de sélection de variables sont utilisées et en particulier la méthode du LASSO (*Least Absolute Shrinkage and Selection Operator*) [186] bayésien permettant d'injecter la connaissance *a priori* de réalisation de l'exposition médicamenteuse telle qu'elle est observée sur la base analysée. Un modèle de Poisson (fonctionnant avec une variable à expliquer binaire) est utilisé dans ce contexte pour permettre la réalisation de ce Shrinkage et les IRR élevés sont identifiés pour la génération de signaux.

Une fois la méthode LGPS exécutée, la méthode LEOPARD est utilisée pour gérer le biais protopathique (relatif au respect de la temporalité entre l'exposition et l'évènement) en vérifiant que la proportion d'exposition est supérieure après l'évènement comparativement à une période antérieure à l'effet. Cette méthode permet de filtrer les signaux renvoyés par LGPS.

La combinaison de LGPS et LEOPARD a obtenu les résultats les plus performants au cours d'un challenge organisé dans le cadre du projet OMOP. Ces résultats n'ont cependant pas été répliqués lors de la synthèse générale réalisée dans le cadre de ce même projet, amenant les auteurs à critiquer cette méthode et à encourager des recherches complémentaires permettant d'expliciter ces différences. En effet, la capacité discriminante de LGPS et LEOPARD est retrouvée faible dans cette seconde étude [164]. Les auteurs confirment toutefois l'intérêt de LEOPARD.

1.4.3.2.4 Statistiques de scan

Les statistiques de scan [187] sont des méthodes de détection locale temporelles et/ou spatiales ne formulant pas d'hypothèse *a priori* sur la localisation d'éventuels clusters. Elles ont un intérêt tout particulier dans le champ de la surveillance épidémiologique. En effet, un biais classique est, sachant qu'on a observé une concentration anormale de cas dans l'espace ou dans le temps, de constituer le groupe d'exposés autour des cas observés (en choisissant soi-même une zone géographique ou une fenêtre de temps), et de le comparer au groupe de non-exposés (en-dehors de la zone, ou hors de la fenêtre temporelle). Cette approche, autrefois assez fréquente, a pour effet de maximiser la taille de l'effet artificiellement, car le groupe des exposés n'est pas défini d'après une exposition « neutre », mais au contraire pour inclure le maximum de cas et donc maximiser le risque relatif. Les statistiques de scan visent au contraire à reproduire le fait que, par le simple fait du hasard, les cas peuvent paraître regroupés dans l'espace ou dans le temps. Ainsi par exemple, si on jette des grains de riz au

sol, certains d'entre eux formeront inévitablement des zones plus denses alors que leur chute a été aléatoire. Ces statistiques de scan reposent sur la construction d'une statistique de test dont la loi de probabilité sous l'hypothèse nulle est construite à partir de simulations répétées d'échantillons ; ces échantillons simulés comprennent la réalisation d'évènements selon une distribution (le plus souvent paramétrique) compatible avec les probabilités observées. Ces simulations peuvent donc, par le seul fait du hasard, aboutir à la formation de concentrations de cas sous l'hypothèse nulle. Le calcul de la vraisemblance de l'observation tient alors compte de cet effet.

Les statistiques de scan sont débutantes dans le champ de la pharmaco-épidémiologie et ont été proposées selon deux méthodologies. Tout d'abord par la construction d'une statistique de test permettant l'analyse de la distribution du délai entre le moment de l'administration du médicament et la survenue d'un évènement puis la mise en évidence d'un pic postérieur au début de l'exposition. Ensuite par la construction d'une statistique de test permettant d'évaluer l'association entre une administration de médicament et un effet d'intérêt. Cette seconde voie [188] a été couplée avec l'utilisation d'une classification thérapeutique afin de tester tous les médicaments individuellement mais également tous les niveaux parents de cette classification : le nom de la méthode concernée est « Tree-based scan statistics ». L'utilisation des niveaux d'agrégation d'une telle classification permet de réaliser un choix *a priori* sur les combinaisons de médicaments pouvant être testées, ce choix *a priori* est nécessaire tant le nombre de combinaisons de médicaments est élevé.

Un article a comparé les performances de GPS à celui de la méthode « Tree-based scan statistics » [189] et a retrouvé un résultat équivalent pour ces deux méthodes.

Enfin, on peut citer le nom d'une autre méthode décrite comme pouvant également jouer un rôle dans la surveillance épidémiologique. Cette méthode est nommée « The maximized sequential probability ratio test » [190], elle permet l'analyse répétée de données en tenant compte du nombre de tests réalisés dans l'estimation du risque, ce nombre étant naturellement déterminé par le rythme de surveillance.

1.4.3.3 Designs d'études utilisant le patient comme son propre témoin

Les études quasi-expérimentales (dites études avant-après) peuvent être utilisées pour l'évaluation d'une prise en charge. Le fait de pouvoir comparer le patient avant et après l'intervention permet de contrôler efficacement certains facteurs de confusion liés aux caractéristiques du patient qui sont constantes au cours de l'étude. En revanche, deux phénomènes bien connus limitent l'intérêt méthodologique de telles études, l'idée générale étant que le motif de prise en charge du patient a tendance à être pour partie spontanément résolutif à travers d'une part la régression vers la moyenne [191] et d'autre part l'effet placebo. Les essais contrôlés randomisés apportent une solution à cette limite méthodologique. De plus, ils sont les seuls conformes au cadre théorique des tests statistiques supposant la réalisation d'un tirage au sort. Nous insistons ici sur le positionnement intermédiaire des essais en cross-over qui permettent le contrôle de certains facteurs de confusion constants pour le patient dans chaque bras mais également le contrôle du fait que la première phase est

différente de la seconde (de par la régression vers la moyenne) par la randomisation de l'ordre des phases de traitement.

Nous présentons maintenant deux designs d'études épidémiologiques qui fonctionnent de façon analogue à ces essais en cross-over. Ils sont d'un intérêt tout particulier car ils permettent de contrôler les facteurs de confusion interindividuels en prenant le patient comme son propre témoin. Nous précisons d'emblée que la limite des études quasi-expérimentales est moins présente dans le cas de la mise en évidence d'effets indésirables car l'évènement d'intérêt n'est pas directement en lien dans ce cas avec les critères d'inclusion. Ces deux types d'études sont la cohorte en cross-over et le cas-témoin en cross-over. Dans chacune de ces études, chaque patient sera pris comme son propre contrôle à un autre moment où il n'a pas présenté l'exposition (dans les études de cohorte en cross-over) ou à un autre moment où il n'a pas présenté l'évènement (dans les études cas-témoin en cross-over).

Nous détaillons maintenant un exemple illustrant la recherche d'association entre le post-partum immédiat et la thrombose [18]. Pour ce faire, une étude de cohorte en cross-over ou une étude cas-témoins en cross-over peuvent être envisagées.

L'étude de cohorte en cross-over est illustrée sur la Figure 7. Dans cette étude, toutes les patientes accouchant sur une période donnée sont incluses et les thromboses survenant au cours d'une période à risque (dont la durée est fixée *a priori*) sont dénombrées. Cette quantité d'évènements survenant au cours de la période d'exposition est comparée à celle retrouvée au cours d'une période contrôle de même durée fixée arbitrairement exactement un an après l'accouchement.

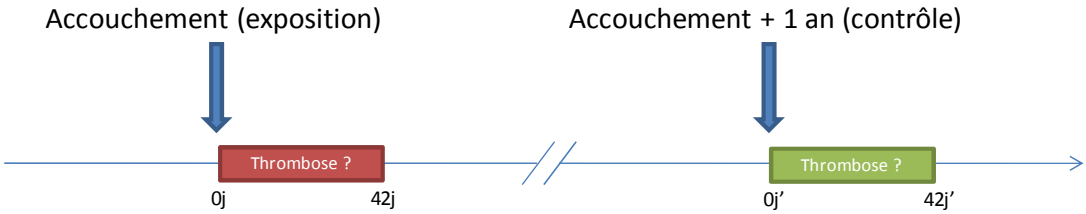


Figure 7 - Illustration d'un cas de cohorte en cross-over

De façon analogue, le risque de thrombose au cours du post-partum peut être évalué selon une étude cas-témoins en cross-over [18]. Cette méthodologie initiée par Maclure [192] et sa déclinaison en série de cas autocontrôlés [193] proposée par Farrington est illustrée sur la Figure 8. Dans cette étude, toutes les patientes ayant présenté une thrombose sur une période donnée sont incluses et les accouchements survenant au cours d'une période à risque (antérieure à la thrombose et dont la durée est fixée *a priori*) sont dénombrés. Cette quantité d'expositions survenant au cours de la période des cas est comparée à celle retrouvée au cours d'une période témoin de même durée fixée arbitrairement exactement un an avant l'accouchement.

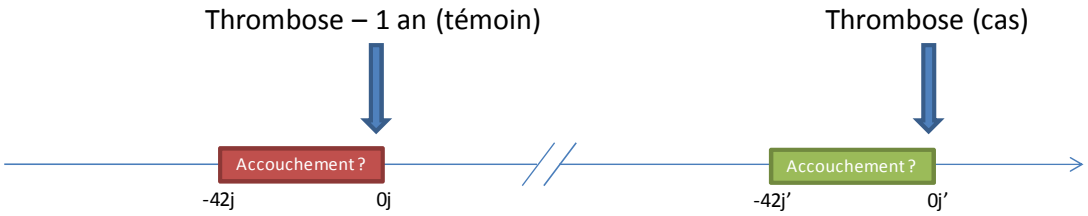


Figure 8 - Illustration de cas-témoin en cross-Over

Le design de ces études doit respecter plusieurs conditions :

- tout d'abord, les expositions d'intérêt doivent avoir une durée limitée afin de pouvoir construire sans difficulté une période contrôle
- ensuite, les évènements d'intérêt doivent être de préférence des cas aigus résolutifs afin de disposer d'un *wash-out* entre les deux périodes comparées
- enfin, il est fréquemment observé que les évènements d'intérêt diminuent la réalisation de l'exposition. Par exemple, si l'on s'intéresse à l'association post-partum et thrombose, le fait d'avoir eu une thrombose (évènement) diminue la probabilité d'être enceinte. Ainsi, il est préférable de choisir une période contrôle postérieure à la période d'exposition dans les études de cohorte en cross-over et de choisir une période témoin antérieure à la période cas dans les études cas-témoin en cross-over. Ainsi il n'y a pas, au cours de la période d'étude, d'exposition d'intérêt étudiée postérieurement à la survenue de l'évènement.

Plusieurs points doivent être discutés dans la construction de ces designs :

- l'analyse doit-elle considérer toutes les occurrences ou seulement le premier évènement ?
- l'analyse doit-elle être ajustée pour les expositions concurrentes qui varient au cours de la période d'observation ?
- quelle fenêtre à risque doit être définie, et à partir de quel délai minimal entre l'exposition et l'évènement ?
- doit-on définir une ou plusieurs fenêtres d'analyse ?
- quel délai minimal de suivi doit-on fixer parmi les critères d'inclusion ?

Ces types d'études comprennent classiquement deux mesures binaires de réalisation de l'évènement ou de l'exposition par patient. L'analyse repose alors classiquement sur une régression logistique conditionnelle (avec appariement sur le patient) ou sur une régression de Poisson conditionnelle pour l'analyse des séries de cas en cross-over. L'utilisation d'un modèle linéaire mixte généralisé est peu courante dans ce contexte car il existe un risque élevé de catégories uniformes, c'est à dire de catégories pour lesquelles les effets observés ont la même valeur. Les différences de convergence entre la régression logistique conditionnelle et le modèle linéaire généralisé à effet aléatoire dépassent le cadre de ce travail. La réalisation d'une telle régression logistique conditionnelle se réalise avec le logiciel de programmation en

statistiques *R* [194] classiquement avec la fonction *clogit* du package *survival* [195]. Il est noté ensuite qu'il est possible de combiner la réalisation d'une régression logistique conditionnelle (compatible avec l'appariement par exemple sur le patient) et d'un modèle linéaire mixte généralisé (compatible avec par exemple un effet centre tel que le lieu d'hospitalisation) en réalisant une régression logistique conditionnelle mixte implémentée sous *R* avec le package *mclogit* [196].

Qualité de ces méthodes dans une perspective pharmaco-épidémiologique :

Le design en cas-témoin cross-over étudié dans le cadre du projet OMOP aboutit à une forte performance discriminante [160]. Ainsi 12 des 20 scénarios étudiés retrouvent une aire sous la courbe supérieure à 0,75 pour tous les médicaments. Les scénarios incluent d'une part le design en cross-over mais également la construction de modèles multivariés avec ajustement sur les médicaments concomitants pouvant favoriser l'effet étudié. Les auteurs nuancent toutefois ce résultat car la qualité de ces mêmes modèles pour la prédiction de la quantité d'effet est en revanche insuffisante.

Le design en cohorte cross-over étudié dans le cadre du projet OMOP aboutit également à une forte performance discriminante [159]. En effet, une aire sous la courbe supérieure à 0,76 est retrouvée pour l'ensemble des scénarios. De même que pour l'étude en cas-témoin cross-over, ce résultat doit néanmoins être tempéré par la qualité insuffisante de la calibration de la quantité d'effet estimé par ces modèles.

Ensuite, il est intéressant de remarquer que, pour un même jeu de données (à savoir ces données réutilisées qui sont de véritables cohortes rétrospectives), le design cas-témoins en cross-over aboutit à des résultats équivalents à ceux du design cohorte en cross-over. Il est tout à fait pertinent de les utiliser en analyse de sensibilité l'une de l'autre [18].

Enfin, des méthodes bayésiennes du type « Lasso Poisson regression » et « Ridge Poisson regression » sont proposées pour l'exploitation des cas-témoins en cross-over [197]. La combinaison de ces deux méthodes semble une piste privilégiée à explorer.

1.4.3.4 Construction de modèles prédictifs par fouille de données

Les méthodes définies ci-avant peuvent être considérées comme appartenant pleinement au champ épidémiologique en ce sens qu'elles formulent une hypothèse entre une exposition et un effet à l'exception des méthodes LGPS et LEOPARD qui ont un positionnement intermédiaire. Certaines méthodes ont moins pour objectif la mise en évidence de couples médicament-effet indésirable au sein d'un modèle cohérent que de réaliser un modèle prédictif de survenue de l'effet que celui-ci comprenne ou non un médicament. On parle volontiers de méthodes de fouille de données pour ces modèles prédictifs issus de l'informatique et adaptés pour l'exploitation des bases de données de grande dimension. Il est néanmoins difficile d'établir une dichotomie entre des méthodes qui relèveraient exclusivement de l'épidémiologie et d'autres qui relèveraient exclusivement de la fouille de données. En effet, certaines méthodes avec sélection de variables dans le cadre de modèles classiquement retrouvés en épidémiologie ont des positionnements intermédiaires. Ainsi, dans le cas de

LGPS, une sélection de variables est réalisée afin d'ajuster l'exposition d'intérêt dans une démarche épidémiologique conservatrice. Si une transition doit être faite entre cette démarche épidémiologique et cette démarche prédictive, on peut considérer que la même méthode utilisée pour réaliser une prédiction de survenue d'évènement mais sans aucune hypothèse *a priori* sur les variables à inclure dans le modèle est alors plutôt dans le champ de la fouille de données. Parmi les méthodes avec sélection de variables utilisant des types de modèles retrouvés classiquement en épidémiologie, on peut citer les régressions PLS (Partial Least Squares), les régressions sur les composantes principales, les régressions selon la méthode *stepwise* combinée avec une méthode de *bootstrap* et les régressions *shrinkage* incluant la méthode du LASSO, la méthode du LASSO bayésien et la méthode *ridge*.

De nombreuses méthodes de fouille de données sont retrouvées. Malgré la grande qualité prédictive de méthodes telles que les réseaux de neurones ou les *Support Vector Machine* (SVM), leur utilisation est rendue difficile par la très grande difficulté d'interprétation des modèles obtenus. Ainsi, les méthodes dont les résultats sont interprétables sont privilégiées. Parmi elles, les arbres de décision et les règles d'association, deux méthodes utilisées dans le projet PSIP [140,198,199], sont présentées ci-dessous.

1.4.3.4.1 Arbres de décision

L'arbre de décision est une méthode non paramétrique et supervisée qui met en relation une variable à expliquer avec un ensemble de variables explicatives. Cette méthode segmente l'échantillon étudié par itérations successives en déterminant étape par étape les variables les plus liées à l'effet dans le sous-groupe étudié. Elle divise à chaque étape l'échantillon en deux sous-groupes homogènes en leur sein (variance intra-groupe) et les plus différents possibles entre eux (variance inter-groupes).

Il existe plusieurs types d'arbres en fonction de la nature de la variable à expliquer :

- Les arbres de classification lorsque la variable à expliquer est qualitative ou binaire
- Les arbres de régression lorsque la variable à expliquer est quantitative
- Les arbres de Poisson lorsque la variable à expliquer correspond à un décompte

Les arbres de survie lorsque la variable à expliquer est une variable censurée. Les techniques de construction d'arbres de décision varient également selon plusieurs points :

- la fonction utilisée pour segmenter les nœuds, c'est-à-dire pour choisir la variable explicative la plus discriminante à chaque étape et éventuellement le paramètre de séparation (un seuil pour les variables quantitatives, des éventuels regroupements pour les variables nominales)
- la méthode utilisée pour trouver la taille optimale de l'arbre (pré-pruning ou post-pruning de l'arbre), en privilégiant une certaine parcimonie
- la prise en compte du risque lié à l'échantillonnage (méthodes de cross-validation)

La méthode développée historiquement est la méthode CHAID (CHi-squared Automatic Interaction Detector) [200], permettant uniquement la classification. Elle utilise la statistique du Chi 2 et réalise un pruning *a priori* (pré-élagage), c'est à dire qu'elle s'arrête à un moment où la quantité du chi 2 devient inférieure à une valeur fixée préalablement. La méthode la plus utilisée aujourd'hui est la méthode CART (Classification And Regression Tree) [201], qui a généralisé cette approche aux autres types de variables à expliquer. Cette méthode utilise l'indice de gini (lorsque la variable à expliquer est binaire) pour sélectionner les variables explicatives utilisées dans chaque nœud, et réalise un pruning *a posteriori* (post élagage), c'est à dire qu'elle construit dans un premier temps l'arbre le plus grand possible et réduit sa dimension dans un second temps. La méthode CART peut de plus inclure une cross-validation, qui permet de tester à chaque étape la stabilité du nœud en fonction d'échantillonnages aléatoires.

Voici les principaux avantages de chaque méthode (l'une par rapport à l'autre) :

- CHAID [200] présente le principal avantage de fonctionner avec des variables multi-classes, c'est à dire qu'il peut très bien, au niveau d'un nœud, se diviser non pas en deux mais en trois branches. Cette propriété vise autant les variables explicatives que les variables à expliquer.
- CART [201] peut également travailler avec des variables qualitatives en les transformant en variables binaires, au prix d'un regroupement de certaines modalités. CART peut de plus travailler avec des variables quantitatives car il les transforme en variables binaires. Pour ce type de variable, il détermine le seuil maximisant le gain de la fonction gini. Par exemple pour l'âge, il détermine l'âge seuil séparant au mieux le groupe étudié et tous les âges sont ensuite interprétés comparativement à ce seuil : si l'âge est supérieur à ce seuil, la variable vaut 1 sinon elle vaut 0.

Plusieurs travaux [202,203] comparant leurs performances respectives ont néanmoins montré une légère supériorité pour CART. Dans la méthode CART, la variable à expliquer peut être une variable quantitative (un arbre de régression est alors réalisé) ou une variable qualitative. Pour une variable binaire, un arbre de classification ou un arbre de régression peut être réalisé. Enfin, il existe, pour les variables binaires à temps-dépendant, une version d'arbre de survie. La méthode est présentée dans le cas de l'arbre de classification pour une variable qualitative à deux modalités.

L’indice de gini utilisé dans CART traduit la probabilité que 2 individus, choisis aléatoirement dans un nœud, appartiennent à 2 classes différentes. Dans une première étape, une segmentation itérative est réalisée jusqu'à obtenir un arbre ne comportant que des feuilles terminales « pures », c'est à dire ne contenant que des individus statistiques ayant tous le même comportement vis à vis de la variable à expliquer ou présentant les mêmes valeurs de variables explicatives. Cependant, ces groupes sont très nombreux, avec des effectifs faibles et leur expression est donc peu reproductible. L’erreur de classement à ce stade est alors généralement quasiment nulle mais pourtant l’arbre construit dans cette étape n’est pas du tout optimal et il apparaît que le taux de mauvais classement dans l’échantillon d’apprentissage n’est pas un critère suffisant pour contrôler la construction de l’arbre : il faut également contrôler sa dimension.

Pour cette raison, Breiman [201] a introduit un paramètre de complexité pénalisant l’arbre proportionnellement à son nombre de feuilles. Ainsi l’erreur globale de l’arbre est une fonction croissante du nombre d’éléments mal classés mais également du nombre de nœuds terminaux de l’arbre. Ce paramètre de complexité permet de trouver un compromis entre la dimension de l’arbre et la proportion de mal classés. Plus l’arbre est grand, plus le nombre de mal classés est petit mais plus le coût lié à la dimension de l’arbre est grand. Et inversement. Des paramètres de complexité croissants sont testés et le meilleur arbre est déterminé pour chaque valeur de paramètre. On choisit ensuite l’arbre ayant l’erreur associée la plus petite. Ces deux dernières étapes sont des étapes de déconstruction de l’arbre maximal obtenu précédemment : ces deux étapes élaguent l’arbre *a posteriori* (« post pruning »).

Ensuite, il est possible de pénaliser davantage le fait de ne pas détecter l’effet alors qu’il est présent (faux négatif) par rapport au fait de détecter l’effet alors qu’il est absent (faux positif), cela a un intérêt tout particulier dans le cas des effets indésirables médicamenteux pour lesquels on souhaite détecter des sous-groupes à risque ayant une bonne sensibilité.

Nous présentons enfin les avantages et inconvénients de la méthode comparativement aux autres méthodes de fouille de données :

- Avantage : les arbres de régression sont facilement interprétables notamment car il existe une analogie avec les arbres de décision fréquemment utilisés en médecine et construits, quant à eux, à dire d’expert. Ce point est essentiel pour l’étape ultérieure de validation qualitative des règles obtenues avec l’arbre.
- Inconvénients : le résultat obtenu dépend de la variable choisie au rang 1 (et on pourrait appliquer ce raisonnement aux rangs suivants). Cette variable « l’emporte » pourtant le plus souvent de très peu au cours du classement des variables candidates par l’indice de gini. Une des conséquences de cela est la grande instabilité des arbres [204] : la modification de quelques paramètres d’agrégation ou la suppression de quelques individus peut suffire à modifier nettement l’arbre issu d’un même échantillon et donc les règles qui en découlent. Mais, autant le nouveau jeu de règles obtenu peut différer fortement, autant la prédiction finale peut rester relativement stable : c’est justement parce que des variables explicatives sont très corrélées entre elles que ce problème peut survenir. Pour compenser ce défaut de stabilité, la méthode

des arbres a été étendue par la méthode des forêts aléatoires (*random forest* [205,206]) permettant un échantillonnage des variables et des individus statistiques permettant la construction de jeux d'arbres (forêts) participant tous à la réalisation du modèle. La méthode des forêts d'arbres est d'une plus grande robustesse mais rend plus compliquée la validation d'éventuelles règles de prédiction : la méthode retourne bien un poids moyen pour chaque variable explicative, mais ne permet pas d'obtenir un arbre consensuel unique, contenant donc un jeu de règles consensuel.

1.4.3.4.2 Règles d'association et règles d'association temporelle

La méthode des règles d'association, dans sa version supervisée, permet de rechercher des règles de prédiction comprenant une ou plusieurs conditions et un évènement (Équation 1) au sein de bases de données de grande dimension. Cette méthodologie a été initiée par Piatetsky-Shapiro [207] qui a décrit des règles fortes découvertes dans une base de données puis Agrawal [1] a introduit le concept de règle d'association initialement expérimenté sur les items des transactions réalisées en supermarché.

$$Condition1 \ \& \ Condition2 => Evènement$$

Équation 1

La construction de règles d'association se décompose en deux étapes : (i) une première étape non supervisée recherchant les co-occurrences fréquentes présentes dans la base et (ii) une seconde étape supervisée construisant des règles d'association à partir des jeux d'items fréquents obtenus dans la première étape. Plus que statistique, le côté novateur de la méthode tient à l'utilisation d'heuristiques simples mais efficaces permettant de limiter le nombre de combinaisons à tester et rendre acceptable le temps de calcul.

Ces deux étapes sont réalisées conditionnellement à des critères de support (fréquence minimale pour les jeux d'items fréquents) et de confiance (confiance minimale pour les règles d'association) définis avant l'exécution de l'algorithme. Le nombre maximal d'items composant un jeu d'items ainsi que les évènements à prédire sont également définis. La première étape est complexe sur un plan combinatoire car le nombre de jeux d'items devant être évalués augmente de façon exponentielle avec le nombre maximal d'items pouvant composer un jeu d'items. Pour cette raison, Agrawal a proposé l'algorithme Apriori dont la principale caractéristique est de ne calculer le support d'un jeu d'items qu'à la condition que tous les jeux d'items qu'il contient soient d'un support supérieur à celui défini de façon préalable à l'exécution de l'algorithme. Ainsi, le support du jeu d'items A&B&C ne sera calculé qu'à la condition que les 3 jeux d'items A&B, A&C et B&C aient un support suffisant.

Un grand nombre de règles sont possiblement identifiées pour chaque évènement d'intérêt [141]. Pour cette raison, plusieurs indicateurs additionnels sont utilisés afin de diminuer le nombre de règles devant être validées par un expert et afin de ne conserver que les règles différant significativement de règles obtenues par hasard. Ainsi, certains indicateurs comme le lift, l'hyperlift et l'hyperconfiance ont été proposés pour tester l'indépendance entre

les conditions et l’évènement de la règle. L’hyperconfiance est proche d’un test exact unilatéral de Fisher, ces indicateurs sont donc robustes y compris sur des items peu fréquents. De plus, des méthodes ont été proposées afin de repérer les règles redondantes, c'est-à-dire les jeux d’items proches (« *closed frequent itemsets* ») [208].

Cette méthode a ensuite été généralisée à la recherche de règles d’association pouvant contenir une contrainte d’ordre [209], on parle de motifs séquentiels et de règles d’association temporelles si cette contrainte d’ordre est le temps.

1.5 Objectif de la thèse

1.5.1 Objectif principal

L'objectif de ce travail est d'étudier la réutilisation de bases de données hospitalières de grande dimension pour la mise en évidence d'effets indésirables liés à la prise d'un médicament ou à la pose d'un dispositif médical implantable.

Cet objectif principal se décline en 2 objectifs opérationnels réalisés sur deux bases de données hospitalières que sont d'une part la base nationale des données du PMSI de 2007 à 2013 et d'autre part une base d'un centre hospitalier (CH) partenaire pour la même période. La première base contient les codes diagnostiques, les actes et des informations démographiques pour environ 170 000 000 de séjours et séances ; la seconde base comprend les mêmes informations mais également les résultats de biologie médicale, les médicaments administrés et les courriers hospitaliers pour environ 80 000 séjours hospitaliers.

1.5.2 Objectif opérationnel 1 : réutilisation de données hospitalières pour la recherche d'effets indésirables médicamenteux

Notre objectif est tout d'abord d'évaluer une méthode automatisée (basée sur un jeu de règles de détection complexes) afin de réaliser une détection rétrospective des EIM à type d'hyperkaliémie.

Nous réutiliserons la base hospitalière du CH partenaire.

Notre objectif est ensuite de réaliser une analyse descriptive de la variation de kaliémie associée à la présence d'un motif séquentiel d'administrations médicamenteuses et de résultats de biologie.

Nous réutiliserons la base hospitalière du CH partenaire.

1.5.3 Objectif opérationnel 2 : réutilisation de données hospitalières pour le suivi des dispositifs médicaux implantables

Notre objectif est d'abord d'estimer le risque thrombotique secondaire à la pose d'une prothèse totale de hanche.

Nous réutiliserons la base nationale des données du PMSI.

Notre objectif est ensuite de construire un outil web permettant à un utilisateur de visualiser dynamiquement, pour un groupe de DMI, la distribution démographique, temporelle et géographique d'une part et la survenue de ré-hospitalisations pour un diagnostic d'intérêt d'autre part.

Nous réutiliserons la base nationale des données du PMSI.

2 Présentations des études réalisées

Quatre travaux correspondant aux 2 objectifs opérationnels sont présentés successivement dans cette partie. Ils correspondent à 4 publications (acceptées ou en cours de finalisation) réalisées dans le cadre du travail de recherche présenté dans cette thèse. Deux publications concernent les effets indésirables des médicaments et deux autres portent sur les effets indésirables survenant suite à la pose d'un dispositif médical.

Les deux premiers travaux présentés dans les parties « 2.1 Première publication - Analyse descriptive des variations de kaliémie associées à un motif séquentiel d'administrations médicamenteuses et de résultats de biologie » et « 2.2 Deuxième publication - Construction et évaluation de règles de détection des effets indésirables médicamenteux à type d'hyperkaliémie » correspondent au premier objectif opérationnel présenté en « 1.5.2 Objectif opérationnel 1 : réutilisation de données hospitalières pour la recherche d'effets indésirables médicamenteux ». Ce premier objectif opérationnel réutilise les données hospitalières de notre hôpital partenaire telles que présentées dans les parties « 1.2.1.3 Données hospitalières médicales » et « 1.2.1.5 Présentation et comparaison des données réutilisées dans cette thèse ».

Les deux études suivantes présentées dans les parties « 2.3 Troisième publication - Estimation du risque thrombotique secondaire à la pose d'une prothèse totale de hanche » et « 2.4 Quatrième publication - Proposition d'un outil web permettant le suivi des dispositifs médicaux implantables » correspondent au second objectif opérationnel présenté en « 1.5.3 Objectif opérationnel 2 : réutilisation de données hospitalières pour le suivi des dispositifs médicaux implantables ». Ce second objectif opérationnel réutilise les données hospitalières de la base nationale du PMSI telles que présentées dans les parties « 1.2.1.4 Données hospitalières médico-administratives : base nationale du PMSI » et « 1.2.1.5 Présentation et comparaison des données réutilisées dans cette thèse ».

Chacune des quatre études est présentée de façon plus détaillée au début de chaque partie.

A titre d'information, ces travaux ont été respectivement :

- soumis et accepté dans le cadre du congrès international MEDINFO pour le premier (« 2.2 Deuxième publication - Construction et évaluation de règles de détection des effets indésirables médicamenteux à type d'hyperkaliémie »). Cette publication est indexée dans la base Pubmed au même titre qu'un article [210].
- soumis et accepté dans le journal « BMC Medical Informatics and Decision Making » pour le deuxième (« 2.2 Deuxième publication - Construction et évaluation de règles de détection des effets indésirables médicamenteux à type d'hyperkaliémie »). Cette publication est indexée dans la base Pubmed [211].
- en cours de finalisation pour le troisième (« 2.3 Troisième publication - Estimation du risque thrombotique secondaire à la pose d'une prothèse totale de hanche »)
- soumis et accepté dans le congrès Medical Informatics Europe pour le quatrième (2.4 Quatrième publication - Proposition d'un outil web permettant le suivi des dispositifs

médicaux implantables). Cette publication est en cours d'indexation dans la base Pubmed.

Bien qu'ayant été rédigés en langue anglaise, ces quatre travaux seront présentés ici en langue française et parfois complétés par des éléments utiles au lecteur mais que les limites de taille imposées par les journaux ne permettaient pas de détailler dans la version originale. Inversement, certaines parties redondantes avec des éléments notamment de l'introduction ont été supprimées.

2.1 Première publication - Analyse descriptive des variations de kaliémie associées à un motif séquentiel d'administrations médicamenteuses et de résultats de biologie

L'étude présentée dans cette partie a porté sur l'analyse descriptive des variations des taux de potassium dans le sang. Comme expliqué dans l'introduction ci-dessous « 2.1.1 Introduction », l'analyse des EIM est à notre connaissance le plus souvent réalisée par l'étude d'évènements binaires tels que, par exemple, la survenue d'une hyperkaliémie (taux de potassium sanguin excédant un certain seuil) ou la survenue d'un évènement thrombotique. L'effet qui nous intéresse dans ce travail est une variation de kaliémie (taux de potassium sanguin) par unité de temps et non l'évènement binaire lié à l'hyperkaliémie. D'autre part, les effets indésirables médicamenteux surviennent volontiers dans un contexte clinico-biologique particulier. Ainsi certains EIM surviennent-ils préférentiellement, par exemple, chez un patient insuffisant rénal. Pour cette raison, nous souhaitons étudier non pas seulement l'évolution de la kaliémie associée à l'administration d'un médicament, mais l'évolution de la kaliémie associée à l'administration d'un ou plusieurs médicaments et en tenant compte du contexte biologique dans lequel ce(s) médicament(s) est (sont) administré(s). Pour ce faire, nous procédons en deux étapes. Tout d'abord, sans tenir compte de la variation de kaliémie, nous construisons des motifs fréquents composés de résultats de biologie médicale et d'administrations de médicaments. Ces motifs fréquents sont construits directement à partir des données de notre hôpital partenaire. Dans un second temps, nous étudions la variation de kaliémie associée à la présence de chacun de ces motifs.

Les données utilisées pour cette analyse sont obtenues auprès de notre hôpital partenaire. Elles correspondent aux données disponibles dans le dossier hospitalier informatisé et sont présentées en détail dans les parties « 1.2.1.3 Données hospitalières médicales » et « 1.2.1.5 Présentation et comparaison des données réutilisées dans cette thèse ». Plus précisément, nous utilisons pour cette étude :

- les résultats de biologie médicale (codés selon la terminologie C-NPU dite IUPAC)
- les administrations de médicaments (codées selon la classification ATC)
- les caractéristiques démographiques des patients telles que retrouvées parmi les données du PMSI.

Les données analysées sont celles de l'année 2009, elles sont structurées selon le modèle de données du projet PSIP présenté dans la partie « 1.2.1.5 Présentation et comparaison des données réutilisées dans cette thèse ».

2.1.1 Introduction

La numérisation des dossiers hospitaliers entraîne le stockage d'une quantité croissante d'informations médicales nécessaires au fonctionnement du dossier patient informatisé [212]. Les hôpitaux collectent ainsi de grandes bases de données qui peuvent être réutilisées dans une perspective d'amélioration de la sécurité des patients. Ces données peuvent ainsi être utilisées pour détecter des effets inattendus ; elles peuvent également être utilisées pour étudier les effets secondaires connus et fréquents de certains médicaments.

Ces bases de données peuvent être utilisées pour étudier les administrations de médicaments dans un contexte observationnel. Des études post-commercialisation peuvent ainsi être envisagées. Un travail bibliographique étendu nous a montré que ces bases de données étaient principalement utilisées pour la détection d'effets indésirables médicamenteux ou d'effets iatrogènes médicamenteux plutôt que pour l'étude d'effets attendus voire pour préciser l'efficacité d'un médicament dans un contexte observationnel.

Les méthodes de détection automatisée des EIM sont basées sur l'utilisation de règles de détection qui peuvent utiliser des types de données différents et selon des niveaux de complexité variés [213,214] : certaines règles utilisent exclusivement les résultats de biologie médicale indépendamment du contexte d'administration de médicaments alors que certaines règles utilisent des résultats de biologie médicale et des administrations de médicaments. Certaines règles peuvent utiliser exclusivement des médicaments tels que des antidotes pour identifier des EIM de façon rétrospective.

Plusieurs études ont montré que l'utilisation des résultats de biologie est particulièrement intéressante pour détecter les EIM. Les résultats de biologie peuvent être utilisés dans ce but de deux manières :

1. avec un seuil absolu, par exemple en considérant qu'un évènement survient lorsque le taux de kaliémie est supérieur à 5,3 mmol/l
2. avec un seuil relatif, par exemple en considérant qu'un évènement survient lorsque la kaliémie est augmentée de 30% par rapport à la valeur précédente. Cette seconde approche semble donner des résultats intéressants [215].

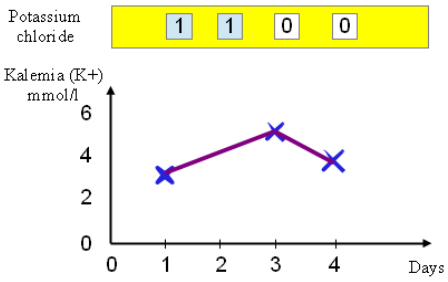


Figure 9 - Exemple de données médicales temporelles liées à un séjour hospitalier (en haut : administration de médicament. En bas : résultat de biologie médicale).

La Figure 9 montre un échantillon de données provenant d'un séjour hospitalier. La partie supérieure présente le chlorure de potassium administré chaque jour. La partie inférieure montre la variation de kaliémie concomitante à cette administration de médicament. Dans cet exemple, un médicament, du chlorure de potassium, est administré pendant 2 jours (au jour 1 et au jour 2) et l'effet associé peut être vu sur la courbe de la kaliémie qui augmente. L'administration est alors arrêtée au jour 3 et une diminution des valeurs de kaliémie est observée. Cet exemple simple illustre l'association possible entre une administration de médicament et une variation d'un résultat de biologie médicale.

L'objectif de ce travail est d'évaluer la variation d'un paramètre de biologie médicale associée à une administration médicamenteuse. Une analyse observationnelle est conduite à partir d'une base de données hospitalière dans le but de prendre en compte le contexte clinico-biologique dans lequel le médicament d'intérêt est administré.

2.1.2 Méthode

Les données ordonnées pour lesquelles l'information chronologique est disponible sont les médicaments et les résultats de biologie médicale. Ces deux types de paramètres présentent l'intérêt que leur chronologie détaillée est disponible dans les données structurées décrivant le séjour hospitalier. Nous précisons que 745 médicaments différents sont disponibles dans la base de données et que 234 types de résultats de biologie différents sont retrouvés.

Les résultats de biologie médicale sont des variables quantitatives mesurées de manière répétées : chaque paramètre de biologie médicale constitue donc une donnée longitudinale. Pour l'administration de médicaments, la dose est disponible ce qui représente également une variable quantitative avec des mesures répétées dans le temps. Ces données peuvent être traitées comme des quantités. Ainsi, nous conservons les résultats de biologie médicale comme des données quantitatives quand on mesure l'effet de la prise d'un médicament (l'effet de la règle). Toutefois, pour tenir compte du contexte de l'administration de médicaments (la cause de la règle), nous utilisons ultérieurement une méthode qui nécessite la transformation des doses de médicament et des résultats de biologie médicale en variables catégorielles. Ainsi, chaque paramètre de biologie médicale est transformé selon le schéma suivant : les quintiles sont calculés pour chaque paramètre et cinq classes sont construites à partir des quintiles. Nous déterminons aussi les valeurs maximales autorisées pour chaque résultat de biologie médicale, ces valeurs sont fixées respectivement à 2 et à 7 mmol/l dans le cas de la kaliémie.

Voici les 5 classes obtenues pour la kaliémie (mmol/l)

- $2 <$ kaliémie < 3.5
- $3.5 <$ kaliémie < 3.9
- $3.9 <$ kaliémie < 4.2
- $4.2 <$ kaliémie < 4.6
- $4.6 <$ kaliémie < 7

Comme le montre la Figure 9, nous avons des informations ordonnées incluant l'administration de médicaments ainsi que la fluctuation des résultats de biologie médicale pour la même période. L'objectif est de décrire les variations à court terme des résultats de biologie médicale en présence de certaines administrations médicamenteuses.

Nous avons ainsi des « règles observationnelles » incluant :

- Causes : un ou plusieurs médicaments éventuellement complétés par certains résultats de biologie médicale reflétant le contexte de l'administration de médicaments. La méthode permet de prendre en compte le contexte de l'administration sur un ou deux jours consécutifs. Les « causes » peuvent donc être des résultats de biologie médicale ou des administrations de médicaments.
- Effet observé : un paramètre quantitatif reflétant la variation du paramètre de biologie médicale en présence des causes.

Pour illustrer cette méthode générique, ce travail est présenté pour le cas de la supplémentation en potassium (chlorure de potassium : code ATC = A12BA01) administrée à faible dose (moins de 2 g/jour strictement). Le chlorure de potassium doit donc être l'une des causes de la règle. Dans cet exemple, l'effet observé est le résultat quantitatif de variation de la kaliémie.

2.1.2.1 Etape 1/3 : construction des causes des règles

Des motifs séquentiels sont recherchés parmi les tables de données des administrations médicamenteuses et de résultats de biologie médicale. Chacun de ces motifs est caractérisé par sa fréquence de réalisation. Les filtres suivants sont appliqués pour cette recherche (les seuils sont fixés arbitrairement) :

- le motif séquentiel doit être trouvé parmi au moins 40 séjours hospitaliers
- le motif séquentiel doit contenir au plus deux éléments par jour
- le motif séquentiel peut s'étendre sur au plus 2 jours consécutifs
- le motif séquentiel doit contenir au moins un médicament (et au moins du chlorure de potassium dans notre exemple)

La méthode utilisée pour construire ces motifs séquentiels est l'algorithme SPADE [209]. Comme indiqué précédemment, les motifs séquentiels ne contenant aucun médicament n'ont pas été conservés ce qui permet en outre de contrôler efficacement le développement exponentiel du nombre de motifs pour chaque paramètre candidat supplémentaire utilisé en entrée de l'algorithme. Cependant, pour illustrer la faisabilité d'une telle recherche, le motif séquentiel ne contenant pas de chlorure de potassium a été construit manuellement et est également présenté dans les résultats.

2.1.2.2 *Etape 2/3 : construction de l'effet de la règle*

Nous sommes intéressés par la variation d'un paramètre de biologie médicale en présence d'une cause (telle que définie à l'étape précédente). Les fluctuations des résultats de biologie médicale peuvent difficilement être analysées dans leur intégralité, c'est à dire pour l'ensemble du séjour. Le séjour à l'hôpital est donc divisé en phases courtes monotones (1 ou 2 jours) entre deux résultats consécutifs d'un même paramètre de biologie médicale.

Plusieurs arguments justifient cette division :

- certaines données sont manquantes (dans ce contexte observationnel), parfois pendant plusieurs jours lorsque le paramètre de biologie n'est pas mesuré
- les phases doivent être compatibles avec la durée pendant laquelle les causes (produites dans la première étape) sont présentes
- il est beaucoup plus difficile (bien que non impossible) d'analyser l'ensemble de la courbe d'un séjour comme un seul effet.

Cette division du séjour en courtes phases est illustrée dans la Figure 10. Dans cet extrait de données d'un séjour, il y a deux phases de variations monotones de la kaliémie; le séjour est donc divisé en deux phases (la phase 1 dure deux jours, la phase 2 dure un jour). Pour chaque phase, une pente qui reflète l'évolution journalière moyenne du paramètre de biologie médicale est calculée (en mmol/l/jour pour la variation de la kaliémie) selon la formule présentée dans l'Équation 2.

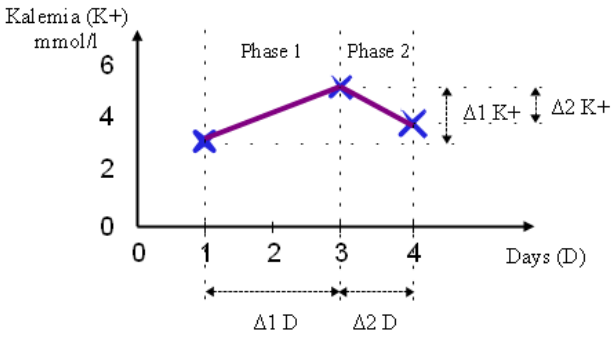


Figure 10 - Division du séjour en phases homogènes (pour la kaliémie)

$$\text{Slope } 1 = \frac{\Delta 1\ \text{K+}}{\Delta 1\ \text{D}} \quad \text{Slope } 2 = \frac{\Delta 2\ \text{K+}}{\Delta 2\ \text{D}}$$

Équation 2

2.1.2.3 Etape 3/3 : Relation entre les causes et l'effet de la règle

Les causes potentielles produites dans la première étape doivent ensuite être rapprochées des pentes produites dans la seconde étape. Deux conditions sont définies pour que la pente (effet) puisse être considérée comme étant temporellement compatible avec une cause :

- la cause doit commencer en même temps que la phase
- la cause doit se terminer un jour avant la fin de la phase (à noter que ce délai d'activité supplémentaire d'une journée dépend de la demi-vie du médicament, et pourrait donc varier avec un autre médicament).

Dans le cas présenté ci-dessous, le chlorure de potassium est administré pendant deux jours consécutifs. La phase 1 du séjour, sur la Figure 11, satisfait les conditions d'alignement entre les causes et l'effet. Ainsi, la pente de la phase 1 est conservée pour étudier la variation de kaliémie en présence de cette séquence d'administration médicamenteuse (motif séquentiel).

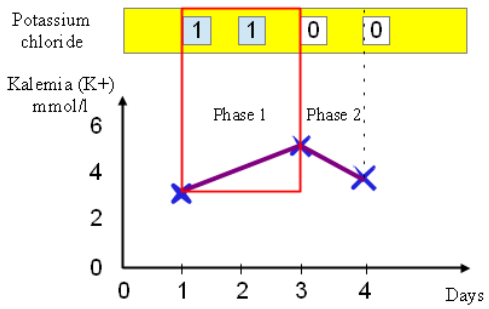


Figure 11 - Liaison entre la cause et la pente

En résumé, les causes définissent une strate (un sous-groupe de phases) à partir de toutes les phases de la base de données et cette strate est ensuite utilisée pour calculer la valeur moyenne de chaque pente d'intérêt ainsi que l'intervalle de confiance à 95% de la moyenne. Le nombre de phases compatibles avec les causes définies est le support observé pour la règle. Enfin, ces calculs sont effectués pour tous les motifs séquentiels obtenus dans la première étape. Ces résultats sont ensuite présentés pour chaque règle. Afin d'illustrer ces résultats, un filtre est appliqué sur le jeu de motifs séquentiels. Ainsi, seules les règles contenant « supplémentation en potassium inférieure à 2g/jour » et « les niveaux de kaliémie au moment de l'administration » parmi leurs causes ont été retenues.

2.1.2.4 Stockage des motifs séquentiels

Le stockage en routine de motifs séquentiels présente plusieurs difficultés, la principale étant que l'imbrication de motifs est d'une profondeur qui n'est pas connue *a priori* et qui n'est en théorie par finie. Ainsi, si l'on considère l'écriture dans une table d'une bi-antibiothérapie sur 2 jours différents, on a 3 informations imbriquées :

1) le code du médicament avec par exemple les seuils pour la binarisation de la dose, la table où celui-ci peut-être trouvé, la terminologie concernée, etc.

2) le motif journalier du jour 1 et celui du jour 2
3) la séquence des motifs définis en 2) sur 2 jours. On peut considérer que le motif en 2) est un cas particulier de séquence avec un écart de 0 jour tel que défini en 3).

On pourrait imaginer des motifs moins profonds comme une séquence de mots ne comprenant que les étapes 1) et 3) mais également des motifs plus profonds comme des séquences d'administration que l'on rechercherait sur plusieurs séjours successifs, voire des séquences de mots ordonnés (n-grams de mots) que l'on rechercherait dans le temps au cours de jours d'hospitalisation successifs, etc.

Nous proposons la solution suivante pour répondre à cette problématique : une table à plat unique interrogée et renseignée par une fonction unique fonctionnant de façon récursive à été proposée. Cette fonction travaille indifféremment avec des données de biologie et médicaments ou des données textuelles, elle ne connaît pas la nature des données qu'elle traite ni les tables où elle doit les rechercher, elle trouve cette information à la volée dans cette table à plat unique. Une fonction analogue a été écrite pour afficher les motifs fréquents et les séjours hospitaliers concernés dans un tableau HTML depuis lequel on peut accéder directement au dossier patient informatisé sur le logiciel ADE-Scorecards [142].

2.1.3 Résultats

2.1.3.1 Motifs séquentiels (causes)

Le nombre total de motifs séquentiels (contenant des supplémentations potassiques à faible dose, moins de 2 g/jour) est égal à 136. Parmi ces 136 motifs séquentiels, 32 apparaissent sur une seule journée et 104 sur 2 jours.

2.1.3.2 Pentes (effet observé)

Le nombre total de pentes de kaliémie générées à partir de la table des résultats de biologie est égal à 4 988 (correspondant à autant de phases de 1 ou 2 jours) et ces pentes sont retrouvées parmi 1 809 patients (hospitalisés pour des séjours différents). Après avoir pris en compte les filtres décrits dans la méthode, il existe six motifs séquentiels pour lesquels sont présentées les variations de kaliémie.

2.1.3.3 Evaluation des règles

Les règles présentées ici ont toutes le même effet, à savoir la variation de la kaliémie. Ces règles ont 3 types de causes différents, c'est à dire qu'elles utilisent trois types de motifs séquentiels :

- un motif séquentiel simple contenant uniquement du chlorure de potassium (sur un jour donné)
- des motifs sur un seul jour, contenant du chlorure de potassium et le contexte du niveau de kaliémie au moment de l'administration médicamenteuse
- deux motifs sur deux jours consécutifs contenant du chlorure de potassium et le contexte du niveau de la kaliémie au moment de l'administration. Ces motifs

contiennent également le fait que du chlorure de potassium a été administré ou non le jour précédent.

2.1.3.3.1 Description de la variation de kaliémie observée consécutive à une prise unique de chlorure de potassium

Le nombre de phases dans ce cas est égal à 587 (support de la règle). La pente moyenne est retrouvée égale à 0,21 [0,17 ; 0,25], comme présenté dans le Tableau 4.

Tableau 4 - Motif unique contenant une supplémentation potassique unique

Cause de la règle	Variation de kaliémie	Support
Chlorure de potassium	0,21 [0,17 ; 0,25]	587

2.1.3.3.2 Présentation des variations de kaliémie consécutives à l'administration d'une dose unique de chlorure de potassium

Le Tableau 5 présente les 4 motifs obtenus avec ces conditions. La valeur moyenne obtenue pour les motifs séquentiels correspondants peut être observée dans la deuxième colonne. La troisième colonne indique le support, à savoir le nombre de motifs à partir desquels la moyenne est calculée.

Tableau 5 - Motifs incluant une supplémentation potassique & une strate de kaliémie (même jour)

Cause de la règle	Variation de kaliémie	Support
{Chlorure de Potassium} ET {2<K+<3.5}	0,44 [0,40 ; 0,48]	268
{Chlorure de Potassium} ET {3.5<K+<3.9}	0,22 [0,12 ; 0,32]	73
{Chlorure de Potassium} ET {3.9<K+<4.2}	0,092 [-0,048 ; 0,232]	51
{Chlorure de Potassium} ET {4.2<K+<4.6}	-0,075 [-0,255 ; 0,105]	44

La distribution de chacune des règles correspondantes est présentée selon une boîte à moustaches sur la Figure 12. On constate que la supplémentation en potassium est associée à une variation différente de kaliémie en fonction de la valeur de la kaliémie au moment de l'administration. L'augmentation de la pente de la kaliémie est de plus en plus petite lorsque la valeur de kaliémie augmente, jusqu'à ce qu'elle soit négative si la kaliémie est comprise entre 4,2 et 4,6 mmol/l au moment de la supplémentation. Il semble qu'il y ait une interaction entre la supplémentation en potassium et la valeur de la kaliémie du jour d'administration.

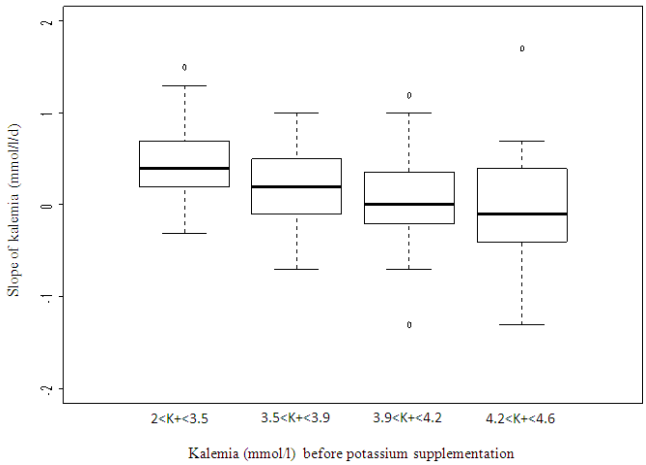


Figure 12 - Distribution des pentes de kaliémie après supplémentation potassique selon la valeur initiale de kaliémie

2.1.3.3.3 Présentation des variations de kaliémie consécutives à l'administration de deux doses de chlorure de potassium sur deux jours consécutifs

Deux motifs séquentiels sur 2 jours consécutifs sont conservés. Le Tableau 6 présente les résultats pour chacun de ces ensembles de causes. Ces motifs se distinguent par la condition du premier jour : dans le premier cas, les phases sélectionnées sont celles dans lesquelles les patients ont reçu du chlorure de potassium sur deux jours consécutifs ; dans le second cas, les phases sélectionnées sont celles dans lesquelles les patients ne recevaient pas de chlorure de potassium le premier jour (puis en ont reçu le jour suivant).

Tableau 6 - Motifs séquentiels incluant une supplémentation potassique et un niveau de kaliémie

Cause de la règle	Variation de kaliémie	Support
[Potassium chloride] THEN [{Potassium chloride} AND {2<K+<3.5}]	0,19 [0,13 ; 0,25]	74
[No Potassium chloride] THEN [{Potassium chloride} AND {2<K+<3.5}]	0,21 [0,15 ; 0,27]	42

Les supports sont de petite taille pour les deux règles. Aucune différence significative entre les moyennes n'est observée dans notre échantillon entre ces deux contextes.

2.1.4 Discussion

Nous avons proposé une méthode pour évaluer l'impact d'un médicament, dans son contexte de prescription (incluant le contexte séquentiel), sur les résultats de biologie médicale. Cette

méthode nous permet de calculer la variation moyenne d'un résultat de biologie médicale associée à une prescription médicamenteuse dans son contexte fréquent de prescription.

Cela a été illustré avec la supplémentation potassique et semble confirmer la nécessité de prendre en compte le contexte clinico-biologique d'une administration de médicament. Ce travail a été réalisé dans un contexte observationnel et permet à un pharmacoépidémiologiste de rechercher un effet attendu ou non survenant secondairement à l'administration d'un médicament. Ce type d'analyse pourrait également permettre d'apprécier l'efficacité d'un médicament dans le cadre d'études observationnelles post-commercialisation. Les contextes d'administration médicamenteuse et de résultat de biologie médicale ont été pris en compte ; ainsi cette méthode pourrait permettre d'explorer l'impact de certaines co-prescriptions sur certains paramètres biologiques.

Plusieurs limites méthodologiques doivent être discutées dans ce contexte observationnel :

- Premièrement, la proportion de phases conservées (par rapport à l'ensemble des résultats de biologie médicale) est petite. Un biais d'observation ou de sélection peut ainsi être évoqué puisque la méthode conserve spécifiquement les phases des patients ayant le suivi biologique le plus régulier.
- Deuxièmement, ce contexte observationnel implique inévitablement un biais d'indication lorsque l'on compare une variation moyenne de kaliémie entre deux traitements ou entre deux contextes. Le score de propension pourrait être utilisé afin de corriger ce biais.
- Troisièmement, l'influence excessive « d'*outliers* » ne peut pas être exclue, en particulier si l'on analyse la kaliémie sans fixer de seuils minimal et maximal.
- Quatrièmement, les résultats de biologie médicale utilisés dans le cadre du séjour hospitalier permettent uniquement de caractériser des effets survenant de façon aiguë après l'administration d'un médicament.

Des méthodes complémentaires devraient être explorées afin de modéliser intégralement la courbe au cours de l'ensemble du séjour hospitalier. De plus, des règles d'association quantitatives pourraient être utiles dans ce contexte. Ensuite, une analyse statistique inférentielle rigoureuse devrait être couplée à ce travail descriptif afin de conserver les éléments de contexte pouvant entrer en interaction avec le médicament vis à vis de la variation de kaliémie. Plusieurs phases successives étant retenues pour un même patient, cette analyse rigoureuse devrait notamment tenir compte du caractère répété de ces mesures.

De plus, la production de motifs séquentiels (utilisés ici comme « causes » de la règle) retourne un nombre très élevé de motifs qui sont parfois similaires. Ainsi, il pourrait être utile de rechercher des « *closed temporal patterns* », c'est à dire des motifs proches (dans l'espace à n dimensions correspondant au nombre de variables implémentées) afin de réduire le nombre de motifs séquentiels par une méthode non-supervisée. Cette réduction du nombre de motifs n'est évidemment pas incompatible avec l'application de filtres par des experts comme cela a été fait dans le cadre de ce travail afin de répondre à une question vis à vis d'un médicament donné.

Enfin, une telle analyse réutilisant des bases de données hospitalières illustre l'intérêt de pouvoir disposer de bases de données de plus grande dimension en construisant des bases de données inter-hospitalières : une condition essentielle pour surmonter cette difficulté est néanmoins de résoudre tout ou partie de la question de l'interopérabilité des données en santé.

2.2 Deuxième publication - Construction et évaluation de règles de détection des effets indésirables médicamenteux à type d'hyperkaliémie

L'étude présentée dans cette partie a porté sur la construction de règles de détection complexes dans le but d'estimer l'incidence hospitalière d'effets indésirables médicamenteux (EIM). Comme expliqué dans l'introduction ci-dessous « 2.2.1 Introduction », la détection automatisée des EIM repose le plus souvent sur l'utilisation de règles utilisant des médicaments ou (exclusif) des résultats de biologie médicale. Plus rarement, on retrouve l'utilisation de règles combinant un médicament et un résultat de biologie médicale mais ces règles ne décrivent pas la chronologie de survenue des évènements qui la composent. Enfin, nous souhaitons construire des règles prenant en compte le contexte de prescription d'un médicament donné.

Les données utilisées pour cette analyse sont obtenues auprès de notre hôpital partenaire. Elles correspondent aux données disponibles dans le dossier hospitalier informatisé et sont présentées en détail dans les parties « 1.2.1.3 Données hospitalières médicales » et « 1.2.1.5 Présentation et comparaison des données réutilisées dans cette thèse ». Plus précisément, nous utilisons pour cette étude :

- les résultats de biologie médicale (codés selon la terminologie C-NPU dite IUPAC)
- les administrations de médicaments (codées selon la classification ATC)
- les caractéristiques démographiques des patients telles que retrouvées parmi les données PMSI
- les codes diagnostiques du PMSI
- les codes d'actes du PMSI

Les données analysées correspondent aux 9 premiers mois de l'année 2010, elles sont structurées selon le modèle de données du projet PSIP présenté dans la partie « 1.2.1.5 Présentation et comparaison des données réutilisées dans cette thèse ».

2.2.1 Introduction

2.2.1.1 Cas des EIM à type d'hyperkaliémie

L'hyperkaliémie induite par un médicament est un problème bien connu [216]. Les complications les plus graves de l'hyperkaliémie sont les troubles du rythme cardiaque (tels que les troubles de conduction, la fibrillation ventriculaire et l'arrêt cardiaque). Les premières modifications électrocardiographiques comprennent une onde T ample et pointue, un allongement de l'espace PR, des troubles de conduction et un élargissement du complexe QRS.

2.2.1.2 Détection des EIM

2.2.1.2.1 Détection rétrospective des EIM

Dans les études post-AMM (phase IV), les données de pharmacovigilance sont classiquement obtenues à partir de déclarations spontanées réalisées par un professionnel de santé confronté à une anomalie qu'il considère être un EIM. Ces déclarations sont loin d'être exhaustives : moins de 5% des EIM seraient ainsi déclarés spontanément [137,138]. En conséquence, les données des bases de pharmacovigilance (i) ne contiennent pas tous les EIM et (ii) fournissent peu d'information sur les erreurs médicamenteuses. De plus, ces données déclaratives sont déjà interprétées et ne contiennent pas de véritables témoins indemnes d'EIM. Les méthodes dites de disproportionnalité sont présentées dans la partie « 1.4.3.1 Transposition dans les bases observationnelles des méthodes dites de disproportionnalité », nous rapportons notamment une étude [163] ayant comparé l'intérêt d'une cohorte de type « nouvel utilisateur » comparativement à une base de pharmacovigilance théorique construite de façon systématique (c'est à dire avec un taux de déclaration de 100%) à partir de cette cohorte de type « nouvel utilisateur » : cette étude conclut que l'analyse de la cohorte de type « nouvel utilisateur » est plus pertinente que l'analyse de la base de pharmacovigilance. Ces limites des bases classiques de pharmacovigilance augmentent l'intérêt de pouvoir disposer de données objectives et non-filtrées telles que celles issues des dossiers patients informatisés. Ainsi, les données issues des systèmes d'information hospitaliers qui ont été collectées en routine et de façon exhaustive semblent être de bonne candidates pour la construction de cohortes rétrospectives. Ces cohortes rétrospectives hospitalières peuvent être explorées par des méthodes de fouille de données [140,217–220].

2.2.1.2.2 Outils informatiques pour la détection des EIM

Plusieurs outils prospectifs de prévention d'EIM ont été développés et évalués [221]. Dans la plupart des cas, ces outils combinent des logiciels de prescription connectée avec un système d'aide à la décision incluant un jeu de règles de prévention, qui détectent des situations à risque. Ces outils fournissent des alertes en temps réel pour éventuellement amener le médecin à modifier une prescription afin de prévenir la survenue d'un EIM. L'intérêt de ces outils dépend directement de la qualité des règles qu'ils implémentent. Chacune de ces règles est composée d'un jeu de causes conduisant à un effet (Équation 3).

$$Cause_1 \cap \ldots \cap Cause_n \rightarrow Effet$$

Équation 3

2.2.1.2.3 Evaluation des outils de détection d'EIM

2.2.1.2.3.1 Métrique pour l'évaluation des règles

Les règles sont classiquement évaluées en termes de niveau de précision (c'est à dire leur valeur prédictive positive, notée VPP) [213]. La précision reflète le niveau de confiance accordé à la règle, c'est à dire la proportion de vrais EIM parmi ceux détectés automatiquement. Dans une perspective de dépistage, la précision devrait toujours être appréciée conjointement avec le rappel (c'est à dire la sensibilité). Le rappel reflète la capacité du système à détecter les EIM qui sont survenus. Rappel et précision sont définis dans la partie « 2.2.2 Méthode ». Le rappel est difficile à évaluer car les EIM sont des évènements rares. L'évaluation du rappel implique donc la revue d'un nombre élevé de séjours hospitaliers par des experts.

Dans les cas les plus extrêmes, une règle qui détecterait un EIM dans chaque séjour hospitalier aurait un rappel de 100% et une précision proche de 0%. Inversement, une règle qui pourrait détecter uniquement les EIM évidents aurait une précision de 100% mais un rappel très faible, égal au taux d'EIM. Ainsi, la qualité globale d'un outil est toujours un compromis entre ces deux mesures, l'équilibrage choisi étant influencé par la finalité d'utilisation de l'outil.

Dans les études traitant de la détection des EIM à type d'hyperkaliémie, le rappel et la précision n'ont pas toujours été évalués. Ainsi par exemple, bien que Dormann [215] ait calculé le rappel et la précision, Brown [222] et Raschke [223] avaient seulement calculé le rappel.

2.2.1.2.3.2 Nature des règles de détection

La prévention prospective des EIM consiste à émettre des alertes lors de la prescription, dès qu'une situation à risque est identifiée par le logiciel de prescription [87,224,225]. Dans ce cadre très précis, l'effet attendu de la situation à risque est décrit en texte libre car il n'est jamais observé, Les règles de prévention sont donc constituées d'un ensemble de conditions structurées, et d'un effet non structuré. Dans le cadre de la prévention prospective d'EIM, Schedlbauer [214] (adapté de Kuperman [226]) distingue deux types d'alertes médicamenteuses. Ces types sont définis non pas d'après la nature de l'effet, mais d'après la nature des causes incriminées. Tout d'abord, les « *basic drug alerts* » sont des règles qui impliquent (notamment) les « interactions médicament-médicament » [227–230] ou les alertes « médicament- allergie » [228]. Ensuite, les « *advanced drug alerts* » impliquent (notamment) les alertes « médicaments et résultat de biologie médicale » [228,230,231], les alertes « âge-médicament » [227] et les *guidelines* relatifs aux doses [227,228,231].

Cette classification peut également être utilisée pour les règles de détection rétrospective des EIM. Dans ce cas, on examine des séjours dans lesquels l'EIM s'est éventuellement déjà produit. Les règles comportent donc nécessairement un effet structuré, de manière à ce que les cas potentiels d'EIM puissent être détectés.

Outre la classification des causes qui constituent la règle, il est également possible de classer la nature des effets. On pourra citer par exemple la survenue d'un résultat anormal de biologie médicale, la suspension d'un médicament, ou la prescription d'un antidote [230]. Cet exemple de classification n'était pas présenté dans la revue de Schedlbauer car cette dernière traitait des méthodes prospectives de prévention et non des méthodes rétrospectives de détection des EIM.

Nous précisons que seules les alertes « médicaments et résultat de biologie médicale » comprennent une cause (administration d'un médicament) et un effet potentiel (un résultat de biologie médicale). Cependant, il n'y a pas d'information sur le lien chronologique entre la validation de la cause et la survenue d'un résultat anormal de laboratoire (puisque l'ordre et l'intervalle de temps ne sont pas spécifiés).

De plus, les experts réalisant une revue de cas doivent toujours évaluer la relation « causale » potentielle entre un médicament et un EIM dans un contexte complexe qui combine des données cliniques et des résultats de biologie médicale. Nous formulons l'hypothèse que ce contexte peut influencer la survenue d'un EIM et devrait ainsi être transcrit dans les règles.

Afin de prendre en compte la chronologie des évènements et le contexte clinique et biologique, nous avons développé un jeu de règles complexes de détection (décrites de façon détaillées dans la partie « 2.2.2.2 Construction des règles et analyse »). Dans le but de lier ces règles aux types d'alertes proposées par Schedlbauer [214], nous incluons :

- des alertes médicamenteuses basiques ou avancées
- des items rendant compte du contexte clinique et biologique du patient
- un contrôle sur la chronologie des évènements (en particulier sur l'intervalle entre l'administration du médicament et la survenue de l'EIM).

2.2.1.2.3.3 Imputabilité d'un médicament

Plusieurs méthodes existent pour évaluer l'imputabilité d'un médicament et le degré d'accord inter-juges utilisant ces méthodes varie grandement d'une méthode à l'autre. Les trois méthodes principales pour valider des cas individuels [232] sont les approches probabilistes, l'opinion d'expert et les approches basées sur des algorithmes. L'approche probabiliste est la méthode la plus reproductible mais elle n'est pas utilisable en routine car elle implique une modélisation complexe. Inversement, l'opinion d'expert est subjective et donc peu standardisée. Enfin, les méthodes basées sur des algorithmes sont standardisées et communément utilisées. Ces dernières sont basées sur des questionnaires permettant à des experts d'estimer la vraisemblance de survenue d'un EIM. Par exemple, Naranjo [134], Kramer [135] et Bégaud [136] ont proposé des algorithmes pour valider individuellement les cas d'EIM. Nous précisons que l'algorithme de Kramer (utilisé dans cette étude) indique que

la manifestation clinique anormale (affectant le patient et potentiellement causée par un médicament) est un « symptôme anormal » et/ou un « résultat anormal de biologie médicale ».

Dans la littérature, la performance des jeux de règles de détection telle qu'elle est présentée par les auteurs varie grandement. Il semble que, au-delà des qualités intrinsèques d'un jeu de règles, la valeur prédictive positive d'un jeu de règles dépende fortement de la méthode d'imputabilité employée. Cette hétérogénéité a été évaluée par Handler en général et a été confirmée dans le cas de l'hyperkaliémie en particulier [213].

2.2.1.3 Objectif

L'objectif du présent travail est d'évaluer une méthode rétrospective de détection automatisée d'EIM au cours de séjours hospitaliers, avec un focus sur les EIM à type d'hyperkaliémie. La méthode applique un jeu de règles complexes de détection à une base de données hospitalières. La qualité de ce jeu de règles est évaluée en confrontant le résultat à une revue experte.

2.2.2 Méthode

2.2.2.1 Séjours d'hospitalisation utilisés pour l'étude

Le tableau de données comprenait les séjours d'hospitalisation (d'une durée supérieure ou égale à deux jours et présentant au moins une mesure de biologie médicale) dans le centre hospitalier partenaire pour les 9 premiers mois de 2010.

2.2.2.2 Construction des règles et analyse

Les règles de détection sont construites à partir de variables agrégées, c'est à dire de groupes de codes choisis par un comité d'experts. Par exemple, la variable agrégée « diurétiques épargneurs de potassium » fait référence à tous les codes ATC de médicaments ayant cette propriété pharmacodynamique. De façon analogue, la variable « cytolyse hépatique » fait référence à tous les résultats de biologie anormaux compatibles avec une cytolyse hépatique (ALAT ou ASAT supérieures à 3 fois la normale) et la variable « infection urinaire » comprend tous les codes CIM-10 compatibles avec ce diagnostic.

2.2.2.3 Propriétés des règles utilisées pour la revue automatique

Dans cette étude, des règles de détection complexes ont été construites. Les principales propriétés de ces règles sont :

1. Les conditions de cause des règles incluent les deux éléments suivants :
    a. un médicament connu pour être associé à un risque d'hyperkaliémie
    b. une variable de contexte pouvant favoriser la survenue de cette hyperkaliémie. Les médicaments inclus sont les inhibiteurs du système rénine-angiotensine, les bêta-bloquants, les diurétiques épargneurs potassiques, le chlorure de potassium, les anti-inflammatoires non stéroïdiens et les héparines non fractionnées. Comme mentionné ci-avant, ces variables médicaments sont

construites à partir d'une agrégation de codes ATC. Les 3 variables de contexte favorisant la survenue d'une hyperkaliémie sont (i) le diabète (identifié par les codes CIM-10), (ii) l'insuffisance rénale (incluant l'insuffisance rénale fonctionnelle ou organique) ainsi que les principaux contextes associés à une insuffisance rénale fonctionnelle que sont l'insuffisance mitrale et l'insuffisance cardiaque congestive et (iii) l'âge supérieur à 70 ans.

2. L'évènement de la règle est toujours la présence d'une hyperkaliémie définie par la présence d'une kaliémie supérieure à 5,3 mmol/l parmi les résultats de biologie médicale.
3. La chronologie : les conditions de cause de la règle sont vérifiées lorsque l'ensemble des sous-conditions sont réunies. De plus, l'effet doit survenir pendant que les conditions sont réunies, ou au plus tard 5 jours après que les conditions ne sont plus réunies. Cet intervalle de temps (extension du délai d'activité des causes) semble approprié pour prendre en compte les demi-vies les plus longues pour les médicaments utilisés dans les règles (certains inhibiteurs du système rénine-angiotensine et certains diurétiques épargneurs de potassium).

Un exemple de règle complexe de détection est présenté dans l'Équation 4. Ce jeu de règles inclut 18 règles avec le même effet (hyperkaliémie).

$$\underbrace{\underbrace{\textit{Renal failure}}_{\textit{context condition}}\ \textit{AND}\ \underbrace{\textit{Potassium chloride}}_{\textit{drug prescription}}}_{\textit{Cause conditions}}\ \underbrace{\overrightarrow{\textit{AND AFTER}}}_{\text{1–5 days after}}\ \underbrace{\textit{Hyperkalemia}}_{\textit{abnormal lab test}}$$
$$\textit{Outcome=Expected anomaly}$$

Équation 4

Un séjour hospitalier est considéré positif pour une « règle de détection complexe » quand les trois conditions (« conditions de cause » ET « évènement » ET « chronologie compatible ») sont réunies. Nous insistons ici sur le fait que ces règles sont utilisées exclusivement de façon rétrospective puisque l'évènement doit toujours être présent pour qu'un séjour soit considéré positif pour une règle. Après avoir été élaborées, ces règles ont été optimisées sur une table de données de séjours antérieurs à l'année 2009.

2.2.2.4 Revue des séjours

Afin d'évaluer la qualité de la revue des règles, un examen des séjours est réalisé par des experts et selon une analyse automatique par les règles complexes de détection.

Un ensemble de scripts programmés en R est utilisé pour analyser la base de données de l'hôpital avec l'ensemble des règles de détection. Les scripts fonctionnent automatiquement avec les données qui sont conformes au modèle de données commun développé dans le projet PSIP et des fichiers XML sont générés en sortie. Les « ADE-Scorecards » [142] sont utilisés par les experts pour revoir les séjours et évaluer ainsi la performance des règles.

2.2.2.4.1 Examen expert des séjours de patients hospitalisés

L'examen a été effectué par un médecin expert et effectué en aveugle des résultats obtenus par une analyse automatisée des séjours hospitaliers. L'algorithme de Kramer a été utilisé pour évaluer l'imputabilité du médicament [135]. Pour chaque séjour passé en revue, l'expert a dû répondre à une question principale et deux questions conditionnelles :

- Selon l'algorithme de Kramer, cette hospitalisation a présenté un EIM certain (« *definitive ADE* », correspondant à un score de Kramer de 7 ou 6) OU un EIM probable (« *probable ADE* », correspondant à un score de Kramer de 5 ou 4) ? OUI / NON
- Si OUI :
    - Quel(s) médicament(s) est (sont) responsable(s) de cette hyperkaliémie ?
    - Cette hyperkaliémie est-elle associée à un symptôme anormal pour le patient ?

2.2.2.5 Tableau de contingence croisant les résultats des revues expertes et automatisées : calcul des critères de qualité pour la détection automatisée

Le format du tableau de contingence est présenté dans le Tableau 7. Il est important de noter que les vrais positifs sont des séjours qui ne sont pas seulement correctement identifiés comme ayant un EIM, mais également pour lesquels l'hyperkaliémie est notamment due aux causes citées dans la règle.

Tableau 7 - Tableau de contingence

	EIM (hyperkaliémie), revue experte	**Pas d'EIM (à type d'hyperkaliémie), revue experte**
EIM (hyperkaliémie) détection automatisée	Vrais positifs (VP)	Faux positifs (FP)
Pas d'EIM détection automatisée	Faux négatif (FN)	Vrai négatif (VN)

Sur la base du tableau de contingence, les Équation 5 et Équation 6 sont utilisées pour calculer respectivement le rappel et la précision.

$$Rappel = \frac{VP}{VP + FN}$$

Équation 5

$$Précision = \frac{VP}{VP + FP}$$

Équation 6

L'ensemble des règles complexes de détection sont utilisées pour détecter les cas d'EIM. Le système semble *a priori* favoriser davantage le rappel que la précision car il est important de ne manquer aucun EIM lors de la détection automatisée : dans ce contexte, la moyenne harmonique du rappel et de la précision (F-mesure) ne semble pas être un critère pertinent. Il est important de noter que :

1) tout séjour qui déclenche une règle a au moins une hyperkaliémie (puisque cette dernière est l'évènement de la règle)
2) tout séjour choisi par l'expert dispose également nécessairement d'une hyperkaliémie (puisque c'est le type d'EIM d'intérêt dans la présente étude).

Ainsi, le tableau de contingence peut être complété par l'examen uniquement des séjours des patients hospitalisés qui présentent une hyperkaliémie. Il convient de noter que la capacité à détecter une hyperkaliémie en soi n'a pas en général besoin d'être évaluée ; il s'agit d'un élément objectif obtenu simplement en interrogeant la base de données des résultats de biologie. Une évaluation simplifiée amènerait à conclure que la précision est de 100% dans la mesure où tous les séjours sélectionnés présentent nécessairement une hyperkaliémie : ce n'est pas ainsi que nous procèderons. En revanche, il est essentiel d'évaluer la capacité du système automatisé à mettre en évidence la « cause » de l'hyperkaliémie. Autrement dit, les experts ne considèrent qu'un cas positif est un vrai positif que si (en outre) la cause de l'hyperkaliémie apparait dans les conditions de la règle. Enfin, comme mentionné ci-dessus, l'expert est invité à préciser si chaque EIM était grave ou pas.

2.2.3 Résultats

Le système automatisé de détection a pris quelques minutes pour la revue complète de la base de données comprenant 3 444 séjours. La revue experte des séjours a porté sur tous les séjours ayant présenté une hyperkaliémie, soit un total de 120 séjours d'hospitalisation. En moyenne, l'expert a pris 15 minutes pour examiner chaque séjour.

L'examen des séjours par les experts a mis en évidence 57 cas d'EIM avec hyperkaliémie. Le Tableau 8 présente les caractéristiques des patients ayant présenté ou non un EIM. Une forte proportion de patients avec EIM avait une insuffisance rénale aiguë (44%). De même, une forte proportion avait une insuffisance cardiaque (26%). Cela est possiblement dû à une insuffisance rénale fonctionnelle résultant d'une insuffisance cardiaque ou de son traitement.

Tableau 8 - Caractéristiques des patients ayant présenté ou non un EIM selon la revue experte

	EIM (hyperkaliémie en présence ou en absence d'un symptôme anormal) revue experte n1 = 57	Pas d'EIM (à type hyperkaliémie) revue experte n2 = 3387
Age (années)	74,6 ± 2.4	67,4 ± 0.3
Femmes	74%	58%
Insuffisance rénale aiguë	44%	11%
Dommage musculaire	2%	2%
Diabète	21%	11%
Insuffisance cardiaque	26%	6%
Nombre de médicaments administrés	7,7 ± 0,5 (médiane 7,5)	4,9 ± 0,1 (médiane 4,7)
Durée de séjour (jours)	14,6 ± 1,3 (médiane 12)	9,7 ± 0.1 (médiane 8)

Les résultats obtenus pour tous les EIM sont présentés dans le Tableau 9. Le système automatisé de détection retrouve 80 EIM avec hyperkaliémie, y compris 51 des 57 identifiés par la revue experte (donnant un rappel de 89,5%). Sur les 80 EIM identifiés automatiquement, dans 51 cas les règles de détection ont identifié correctement la cause de l'hyperkaliémie (donnant une précision de 63,7%).

Tableau 9 - Evaluation de la détection automatisée des EIM à type d'hyperkaliémie

	EIM (avec hyperkaliémie), revue experte	Pas d'EIM (avec hyperkaliémie), revue experte	Total
EIM (avec hyperkaliémie), détection automatique	**VP=51**	**FP=29**	**80**
Pas d'EIM (avec hyperkaliémie), détection automatique	**FN=6**	VN=3358	**3364**
Total	***57***	***3387***	***3444***

Le Tableau 10 compare la détection automatisée et la revue experte pour chacun des services hospitaliers.

Tableau 10 - Comparaison de la détection automatisée et de la revue experte pour chacun des services hospitaliers

Services hospitaliers	Séjours	Séjours avec hyperkaliémie (HK)	EIM (avec HK), revue experte	EIM (avec HK) détection automatisée	EIM (avec HK), revue experte et détection automatisée
Gériatrie	257	12 (4.6%)	6 (2.3%)	9 (3.5%)	5 (1.9%)
Gastroentérologie et Cardiologie	970	46 (4.7%)	31 (3.1%)	31 (3.1%)	25 (2.5%)
Médecine interne	761	21 (2.7%)	10 (1.3%)	15 (1.9%)	10 (1.3%)
Pneumologie	655	24 (3.6%)	13 (1.9%)	19 (2.9%)	11 (1.6%)
Chirurgie	933	24 (2.5%)	2 (0.2%)	10 (1.0%)	2 (0.2%)
Total (séjours uniques)	**3444**	**120 (3.4%)**	**57 (1.6%)**	**80 (2.3%)**	**51 (1.4%)**

La revue experte a identifié trois EIM graves (tels que définis dans l'introduction) avec une hyperkaliémie. Les trois cas ont été identifiés automatiquement avec une attribution correcte du médicament. Les résultats des EIM et les médicaments concernés sont spécifiés dans le Tableau 11. Bien que deux des trois patients soient décédés, il faut garder à l'esprit que ces deux patients étaient atteints d'une maladie dont l'issue fatale était inévitable.

Tableau 11 - Caractéristiques des patients ayant présenté un EIM grave

Age	Genre	Durée du séjour (jours)	Kaliémie (mmol/l)	Cause de l'EIM	Arythmie sévère	Transfert en unité de soin intensif
50	M	7	8,0	Antagoniste des récepteurs de l'angiotensine (sartan) et chlorure de potassium per os	Oui	Non
97	F	5	6,1	Inhibiteur de l'enzyme conversion	Oui	Non
71	M	24	7,7	Chlorure de potassium intraveineux	Oui	Non

2.2.4 Discussion

Dans cette étude, nous avons utilisé un jeu de règles complexes de détection pour mettre en évidence des EIM à type d'hyperkaliémie par l'analyse d'une base de données hospitalière. Cette détection automatisée a retrouvé des valeurs élevées pour le rappel et la précision lors de la comparaison à la revue experte des séjours. En termes de rappel, 89,5% des EIM à type d'hyperkaliémie ont été retrouvés. De plus, tous les EIM graves ont été détectés ce qui semble un résultat important ; il est néanmoins difficile de généraliser ce résultat compte tenu du

nombre très réduit d'EIM de ce type. En termes de précision, 63,7% des EIM identifiés automatiquement étaient véritablement des EIM à type d'hyperkaliémie.

Ce travail semble confirmer la nécessité de prendre en compte le contexte clinique et biologique du patient pour la détection automatisée d'EIM. Les patients avec une fonction rénale normale sont peu à risque de survenue d'une hyperkaliémie puisque l'excès de potassium est rapidement éliminé par les reins. Inversement, l'insuffisance rénale chronique entraîne rarement une hyperkaliémie à l'exception des cas d'insuffisance rénale terminale. En revanche, une prescription inappropriée en cas d'insuffisance rénale chronique favorise la survenue d'une hyperkaliémie. L'insuffisance rénale (et en particulier l'insuffisance rénale aigüe) apparaît ainsi comme une condition nécessaire mais non suffisante de survenue d'une hyperkaliémie.

De plus, ce travail illustre l'intérêt des résultats de biologie médicale pour la détection automatisée des EIM. Les résultats de biologie médicale ont été utilisés à la fois comme conditions et comme effet au sein de notre jeu de règles complexes de détection. Cela semble appuyer la pertinence de l'utilisation des résultats de biologie médicale pour la détection d'EIM [233] en général et d'EIM à type d'hyperkaliémie en particulier. Ensuite, toujours dans le cas de l'hyperkaliémie, les symptômes cardiaques surviennent après qu'une hyperkaliémie a été observée parmi les résultats de biologie médicale, laissant à penser que cette approche est plus sensible. Ce résultat est en accord avec la revue de Handler [213] qui identifie 36 signaux uniques d'EIM incluant 10 problèmes de dose, 19 valeurs anormales de biologie médicale, et 7 administrations d’antidote. Enfin, les résultats de biologie médicale sont des données structurées disponibles au cours du séjour hospitalier. Ce n'est pas le cas des informations diagnostiques structurées (par exemple les codes diagnostiques) qui sont généralement codées *a posteriori* du séjour et ne contiennent pas les dates d’activité ou de survenue des maladies ou symptômes décrits.

Les règles construites dans cette étude ne prenaient pas en compte l'administration d'antidote. Cela est dû au fait que ces règles étaient construites pour pouvoir fonctionner de façon prospective [234] et rétrospective. Dans un cadre rétrospectif, l'utilisation d'un effet tel que l'administration d'un antidote (par exemple un chélateur du potassium) semble utile.

Notre travail peut être comparé avec trois études similaires ayant également calculé des critères de qualité pour un jeu de règles de détection. Premièrement, Dormann [215] a évalué un système de détection d'EIM automatisé en gastroentérologie. Les règles retenues utilisaient des alertes basées sur les résultats de biologie médicale et un EIM était confirmé si le médecin notait un changement dans la prescription médicamenteuse, un ajout de prélèvement pour un paramètre de biologie médicale ou d'autres actions diagnostiques pouvant être reliées à un EIM. Deux méthodes de détection automatisée ont été employées. La méthode 1 avait une précision de 36% pour les dyskaliémies et un rappel de 91% tout type d'EIM confondus. Les valeurs correspondantes pour la méthode 2 étaient retrouvées à 67% pour la précision (dyskaliémie) et 40% pour le rappel (tout type d'EIM). Deuxièmement, Brown [222] présente les résultats du projet « *Recognizing, Assessing and Documenting Adverse Rx events* (RADARx) ». De nouveau, les règles étaient des alertes basées sur les résultats de biologie

médicale et les EIM étaient validés selon l'algorithme de Naranjo. Brown a rapporté une précision de 11,1% (dans le cas du potassium) mais n'a pas calculé le rappel. Troisièmement, Raschke [223] a présenté un « *Computer Alert System to Prevent Injury From Adverse Drug Events* », avec une règle pour la détection de « hyperkaliémie ET médicaments multiples » (parmi lesquels inhibiteur de l'enzyme de conversion, supplémentation potassique, diurétiques épargneurs de potassium, sulfate de triméthoprime, héparine sodique). L'évaluation conduite par Raschke évalue la détection prospective de situations à risque plutôt que la détection *a posteriori* de la survenue de l'EIM. Parmi les 69 alertes de ce type, 41 alertes étaient des vrais positifs et 10 constituaient une situation à risque potentiel pour le patient (que le médecin n'avait pas identifiée) mais non un EIM prouvé. Il est ainsi difficile de comparer les résultats de Raschke avec nos résultats. La précision trouvée par les auteurs est de 59% et le rappel n'a pas été calculé. Les règles de Raschke ont pris en compte l'hyperkaliémie et la prescription d'un médicament mais pas le contexte clinico-biologique de survenue de cette hyperkaliémie.

En conséquence, notre résultat semble plus satisfaisant que ceux décrits dans les trois études ci-dessus. Cependant, nous ne pouvons pas être certains qu'une même qualité de résultats serait obtenue avec un jeu de règle qui traiterait d'un autre effet, c'est à dire un autre type d'EIM. Ensuite, le faible nombre d'EIM dans cette étude implique qu'il est difficile de généraliser ce résultat. Les EIM sont des évènements rares, on comprend ici l'intérêt de pouvoir construire des bases de données inter-hospitalières.

Enfin, il nous semble que ces règles complexes de détection rétrospective d'EIM pourraient être utiles dans deux situations. Tout d'abord, la revue experte de cas est une tâche fastidieuse. La détection automatique pourrait permettre à un pharmacologue de réaliser une revue sur un plus petit nombre de cas détectés préalablement de façon automatisée. Cette approche réduirait le nombre de séjours hospitaliers à revoir. Ce fonctionnement en deux temps peut être envisagé car le rappel de notre jeu de règles est élevé. Cet outil pourrait ainsi être utilisé pour générer des alertes rétrospectives au cours de l'hospitalisation. Un spécialiste de pharmacovigilance pourrait être alerté, il reverrait alors le cas et contacterait le médecin si nécessaire [235], de la même façon que le biologiste médical peut par exemple décider de contacter sans tarder le médecin en cas d'anomalie grave du bilan biologique sanguin. Ensuite, la performance de notre jeu de règles complexes de détection montre qu'il pourrait être utilisé pour estimer l'incidence hospitalière d'EIM : le nombre d'EIM pourrait être estimé par la formule suivante « # (détecté) * (Précision/Rappel) ». Ce type d’outil informatisé jouerait ainsi un rôle dans la pharmacoépidémiologie hospitalière à travers l’analyse en routine de bases de données inter-hospitalières de grande dimension.

2.3 Troisième publication - Estimation du risque thrombotique secondaire à la pose d'une prothèse totale de hanche

L'étude présentée dans cette partie a porté sur l'estimation du risque thrombotique faisant suite à la pose d'une prothèse totale de hanche. Ce risque est bien décrit et justifie la mise en place systématique d'une anticoagulation afin de prévenir la survenue d'un évènement thromboembolique (les recommandations sont présentées dans l'introduction ci-dessous « 0

Introduction »). L'excès de risque thrombotique résiduel (dans ce contexte d'anticoagulation systématique) en situation réelle gagnerait en revanche à être précisé ainsi que la durée pendant laquelle un excès de risque persiste. Comme présenté dans la partie « 1.4.3 Types d'études permettant la mise en évidence d'évènements indésirables à partir des bases de données observationnelles », les méthodes utilisant le patient comme son propre témoin présentent un grand intérêt à la condition que l'exposition et l'évènement soient limités dans le temps, ce qui est le cas ici. Ces méthodes sont détaillées plus spécifiquement dans la partie « 1.4.3.3 Designs d'études utilisant le patient comme son propre témoin », la méthode retenue pour cette étude est le « *case cross-over* ».

Les données utilisées pour cette analyse sont les données de la base nationale du PMSI. Elles permettent la tarification des séjours hospitaliers dans le champ MCO. Nous les avons présentées en détail dans les parties « 1.2.1.4 Données hospitalières médico-administratives : base nationale du PMSI » et « 1.2.1.5 Présentation et comparaison des données réutilisées dans cette thèse ». Plus précisément, nous utilisons pour cette étude :

- les actes médicaux (codés selon la terminologie CCAM)
- les diagnostics des patients au cours de chaque séjour (codés selon la CIM-10)
- les caractéristiques démographiques des patients

Nous précisons que ces données comprennent le numéro ANO qui est un identifiant unique permanent et anonyme par patient. Cet identifiant permet de reconstruire le parcours de chaque patient en termes de séjours hospitaliers dans le temps et dans l'espace (il est permanent également entre établissements). Les données analysées sont celle des années 2007 à 2013 et sont structurées selon un modèle de données présenté dans la partie « 1.2.1.5 Présentation et comparaison des données réutilisées dans cette thèse ».

.

2.3.1 Introduction

L'*American College of Chest Physicians* recommande l'utilisation d'un antithrombotique après une « chirurgie orthopédique majeure » pour un minimum de 10 à 14 jours [236,237]. On entend par « chirurgie orthopédique majeure » les prothèses totales de hanche et de genou ainsi que la chirurgie d'une fracture du col. Ces types de chirurgie sont à haut risque thrombo-embolique veineux. L'utilisation des héparines de bas poids moléculaire débutées 12h avant ou 12h après la chirurgie est à privilégier devant les autres antithrombotiques. La poursuite des antithrombotiques en ambulatoire est ensuite conseillée pour une durée totale de 35 jours suivant la chirurgie. L'utilisation d'une méthode de compression est conseillée en complément du traitement pharmacologique, cette dernière méthode est de plus la seule recommandée en cas de risque élevé de saignement. Les recommandations de la Société Française d'Anesthésie-Réanimation (SFAR) sont très proches en termes de traitement pharmacologique : la durée de traitement anticoagulant pour la prothèse totale de hanche est dans ce cas de 42 jours.

Sur un plan méthodologique, évaluer le risque d'évènement thrombo-embolique après la pose d'une prothèse totale de hanche requiert un grand volume de données incluant plusieurs centaines de milliers, voire plusieurs millions de prothèses et repose classiquement sur des « *population-based studies* » afin de disposer de la puissance statistique nécessaire. Ensuite, les designs en cross-over tels que les cohortes en cross-over ou les cas-témoins en cross-over (« *case cross-over* ») ont prouvé leur supériorité dans le champ de la pharmacoépidémiologie. Les différents designs d'étude ont en effet été comparés de façon systématique et empirique dans le cadre du projet OMOP [158].

Le but de notre étude est d'évaluer le risque thromboembolique au cours des mois suivant la pose d'une prothèse totale de hanche.

2.3.2 Méthode

2.3.2.1 Données

La base des hospitalisations dans le secteur Médecine Chirurgie Obstétrique (MCO) est utilisée pour cette étude. Elle représente 171 556 421 séjours en France sur la période 2007-2013. Sur cette même période, 983 746 séjours hospitaliers avec pose de prothèse totale de hanche sont retrouvés pour l'ensemble de la France parmi 871 087 patients différents. Ces données ont été obtenues auprès de l'ATIH après obtention de l'autorisation auprès de la CNIL. Un résumé de sortie anonymisé est produit pour chaque hospitalisation dans un hôpital public ou privé français. Ce résumé de sortie anonymisé permet le calcul d'un tarif pour chaque séjour hospitalier et ces données sont recueillies à cet effet par l'assurance maladie et l'ATIH. Le Tableau 12 présente les données disponibles pour chaque séjour hospitalier en France dans le secteur MCO : elles incluent notamment les codes diagnostiques selon la CIM-10, les actes médicaux selon la CCAM, l'âge, le sexe et le numéro anonyme unique par patient. Enfin, l'intervalle en jours entre deux séjours hospitaliers peut être calculé à partir de cette base.

Tableau 12 - Types de données disponibles pour chaque séjour hospitalier

Variables	Terminologie
Diagnostics	CIM-10
Actes	CCAM
Age et genre	-
Identifiant unique anonyme par patient	Code unique « ANO »

Parmi les codes diagnostiques, depuis le 1er mars 2009, le « diagnostic principal » correspond au motif d'hospitalisation tel qu'il est connu à la sortie du patient de l'hôpital.

2.3.2.2 Design d'étude

Une analyse en « *case cross-over* » a été conduite à partir de cette base nationale du PMSI dans le secteur MCO. Une partie de la base de données a été extraite : tous les séjours des patients ayant au moins une fois (au cours d'un séjour) présenté un évènement thromboembolique ont été sélectionnés. Nous allons suivre dans cette étude des expositions et des évènements pouvant être tracés dans cette base d'hospitalisation. C'est le cas de nos expositions (les poses de prothèses totales de hanche) qui font systématiquement l'objet d'un séjour hospitalier, ce sera également le cas des évènements thromboemboliques graves (détaillés ci-après) qui constitueront nos cas.

Dans cette étude, chaque patient sert d'une part comme cas et d'autre part comme son propre témoin à un autre moment dans le temps. Ce type de design adapté pour les expositions et les évènements limités dans le temps permet de contrôler les facteurs de confusion liés au patient et constants dans le temps. Le jeu de séjours ayant présenté un évènement thromboembolique est extrait au cours d'une période de 58 mois débutant au 1er mars 2009 et se terminant au 31 décembre 2013. Les séjours des années 2007 à 2013 sont utilisés pour identifier les prothèses totales de hanche qui sont dans cette étude notre exposition d'intérêt. Nous avons choisi un design en « *case cross-over* » plutôt qu'un design en *« cohort cross-over »* pour deux raisons : tout d'abord, la qualité des données est supérieure à partir du 1er mars 2009 car le diagnostic principal devient le motif d'hospitalisation. Il est ainsi plus facile de disposer des séjours de cas au delà de cette date afin de reconstruire la chronologie entre une prothèse totale de hanche et un évènement thromboembolique survenus au cours du même séjour (c'est à dire le séjour de cas). Ensuite, la base nationale de séjours hospitaliers utilisée pour ce travail ne contient pas les décès survenus en dehors de l'hôpital. Utiliser un design en *« cohort cross-over »* aurait plutôt impliqué le choix de témoins postérieurs au séjour de cas afin de minimiser le biais de survie et nous n'aurions pas pu censurer tous les patients décédés dans cette cohorte rétrospective.

La probabilité d'avoir eu une prothèse totale de hanche au cours de la période précédant l'évènement thromboembolique est comparée à la probabilité d'avoir eu une prothèse totale de hanche au cours d'une période comparable située exactement un an auparavant. Ensuite, 9

périodes successives de 30 jours sont étudiées. Elles s'étalent sur une période de 270 jours avant la pose de la prothèse totale de hanche et sont systématiquement comparées à la période correspondante un an plus tôt.

En cas d'évènements thromboemboliques multiples sur la période d'inclusion, seul le premier évènement est conservé pour l'analyse. De plus, les patients qui avaient été admis à l'hôpital pour un évènement thromboembolique au cours des 27 mois précédant le cas d'inclusion (en utilisant les données depuis le 1er janvier 2007) ne sont pas inclus. Ces patients ne sont pas inclus pour deux raisons : tout d'abord, les patients souffrant d'un antécédent personnel d'évènement thromboembolique sont possiblement sous traitement anticoagulant ce qui pourrait biaiser l'analyse. Ensuite, nous nous assurons ainsi que la période témoin est bien indemne d'évènement thromboembolique.

2.3.2.3 Identification des prothèses totales de hanche

Les séjours hospitaliers avec une prothèse totale de hanche sont identifiés par la présence d'un code du type « NEKA0* » parmi les codes d'acte et la présence d'un code du type « M16.* » ou « S72.* » parmi les codes diagnostiques.

2.3.2.4 Identification des évènements thromboemboliques

Différents algorithmes permettant le suivi des évènements veineux thromboemboliques parmi les données médico-administratives ont été proposés et évalués [113]. La plupart de ces algorithmes se réfèrent à la CIM dans sa 9ème version. Une évaluation d'un algorithme [111] utilisant la CIM dans sa 10è version révèle que les séjours hospitaliers avec une thrombose veineuse profonde (TVP) étaient plus difficiles à identifier que les séjours hospitaliers avec une embolie pulmonaire. Les codes diagnostiques utilisés permettaient l'identification de 88,9% [85,6% ; 92,2%] des embolies pulmonaires alors que la sensibilité de la détection est seulement de 58% [51.9% ; 64.1%] pour la TVP. De plus, la TVP ne requiert pas nécessairement une prise en charge hospitalière ce qui incite également à ne pas les suivre à partir de notre base de données hospitalières.

Ensuite, une TVP a une plus grande probabilité d'être codée au cours du séjour de pose de prothèse par rapport à la période contrôle un an auparavant. Si nous les suivions, nous pourrions être confrontés à un biais d'observation. De façon analogue, il est possible que certaines embolies pulmonaires silencieuses soient détectées au cours du séjour de pose de prothèse totale de hanche alors qu'elles ne le seraient pas au cours de la période contrôle, si cette période ne faisait pas l'objet d'une hospitalisation. On peut néanmoins penser que ce biais potentiel d'observation est moins marqué dans le cas de l'embolie pulmonaire. En synthèse, pour les raisons présentées ci-avant, nous avons choisi de rechercher exclusivement les embolies pulmonaires pour évaluer le risque thromboembolique dans le cadre de cette étude en « *case cross-over* ».

Les codes utilisés pour suivre les embolies pulmonaires sont I26.0 « embolie pulmonaire » et I26.9 « embolie pulmonaire avec mention de cœur pulmonaire aigu », ces deux codes sont donc les seuls utilisés pour identifier les évènements veineux thromboemboliques.

2.3.2.5 Calcul du délai entre l'exposition et l'évènement veineux thromboembolique

Nous présentons maintenant la méthode utilisée pour estimer rétrospectivement le délai entre la pose d'une prothèse totale de hanche et la survenue d'un évènement veineux thromboembolique. Deux situations doivent être différenciées :

1. une première situation où l'évènement est survenu au cours d'un séjour postérieur au séjour de pose de prothèse totale de hanche
2. une seconde situation où l'évènement est survenu au cours du même séjour que celui de pose de prothèse totale de hanche

Dans le premier cas, nous avons utilisé la différence en jours entre les deux dates d'hospitalisation. Dans le second cas qui est moins fréquent, la reconstruction de la chronologie au cours du séjour hospitalier est complexe car nous ne connaissons pas les dates de présence des codes diagnostiques qui sont codés en une fois à la fin du séjour : nous avons utilisé le fait que le diagnostic principal correspond au motif d'hospitalisation. Ainsi, parmi les séjours ayant un acte de pose de prothèse totale de hanche et un diagnostic d'embolie pulmonaire, nous avons considéré que la pose de prothèse totale de hanche était survenue avant l'embolie pulmonaire si le diagnostic principal était un code CIM-10 de fracture (« S72.* ») ou un code CIM-10 de coxarthrose (« M16.* »). Cette condition portant sur le diagnostic principal semble néanmoins une précaution puisqu'un évènement thromboembolique aurait vraisemblablement contre-indiqué la réalisation de l'acte de pose de prothèse totale de hanche. Enfin, les évènements survenus au cours du séjour de pose de la prothèse totale de hanche ont été considérés comme étant survenus au cours du premier mois suivant la pose de la prothèse totale de hanche.

2.3.2.6 Analyse statistique

Nous avons évalué la probabilité d'être exposé à une prothèse totale de hanche au cours des 30 jours précédant l'embolie pulmonaire par rapport à la probabilité d'être exposé à une prothèse totale de hanche au cours de 30 jours situés exactement un an plus tôt. Nous avons réalisé ensuite le même type d'analyse pour chacun des mois depuis le mois « 2 » avant l'embolie pulmonaire jusqu'au mois « 9 » avant l'embolie pulmonaire. Nous avons utilisé une régression logistique conditionnelle pour calculer l'odds ratio et son intervalle de confiance à 95% pour chacun des mois.

Notre hypothèse *a priori* était que le risque diminuerait progressivement au fil des mois et notre but était d'identifier la valeur de ce risque et la durée de la période pendant laquelle il serait significativement augmenté au risque alpha égal à 5%.

2.3.3 Résultats

Du 1er mars 2009 au 31 décembre 2013, 54 268 patients ont été inclus. Le Tableau 13 décrit les séjours hospitaliers des patients ayant eu une pose de prothèse totale de hanche sur la période allant du 1er janvier 2007 au 31 décembre 2013. L'âge moyen des patients au cours de ces séjours est retrouvé à 72,29 ans et la proportion d'hommes est retrouvée à 39,84%. Afin de calculer ultérieurement le taux de complication sur la période d'inclusion, le nombre

de séjours de pose de prothèse totale de hanche entre le 1er mars 2009 et le 31 décembre 2013 est dénombré, il est retrouvé à 673 240.

Tableau 13 - Caractéristiques des séjours hospitaliers avec pose de prothèse totale de hanche de 2007 à 2013

Variables		n=983 746
Age		72,29 ± 12,59
Age>75		45.44%
Sexe (% d'hommes)		39,84%
Année de la pose de prothèse de hanche		
	2007	14.69%
	2008	14.48%
	2009	14.07%
	2010	14.46%
	2011	13.45%
	2012	14.45%
	2013	14.39%

Le Tableau 14 montre ensuite les résultats obtenus pour le risque veineux thrombo-embolique. Ces résultats sont détaillés pour chaque intervalle de 30 jours après la pose d'une prothèse totale de hanche. Le nombre exact d'évènements thrombo-emboliques et le taux correspondant pour 100 000 prothèses totales de hanche sont présentés dans le Tableau 14. Toujours dans ce même tableau, nous présentons les odds ratio et leur intervalle de confiance à 95% pour chaque période de 30 jours.

Du jour 0 au jour 29, 1 264 cas d'embolie pulmonaire sont identifiés ce qui représente un taux de 18,7/10 000 prothèses totales de hanche et l'odds ratio est retrouvé à 18,0 [14,1 ; 22,8]. Le risque est très inférieur pour les périodes suivantes, il est ainsi retrouvé à 3,7 [2,9 ; 4,7] pour la période du jour 30 au jour 59. On observe ensuite une diminution progressive de l'intervalle 30-59 à l'intervalle 150-179 avec un odds ratio passant de 3,7 [2,9 ; 4,7] à 1,3 [1,0 ; 1,7]. Au delà de cette période, l'estimation de l'odds ratio est retrouvée proche de 1 et le risque n'est plus retrouvé significativement augmenté.

Tableau 14 - Risque d'évènement thrombo-embolique suivant une pose de prothèse totale de hanche (selon le délai en jours depuis l'acte de pose)

Intervalle en jours	Période cas Nombre d'évènements* (taux/10 000 prothèses totales de hanche)	Période contrôle Nombre d'évènements* (taux/10 000 prothèses totales de hanche)	Odds Ratio [IC 95%]*
0-29	1264 (18,7)	74 (1,0)	18,0 [14,1 ; 22,8]
30-59	315 (4,6)	85 (1,2)	3,7 [2,9 ; 4,7]
60-89	270 (4,0)	84 (1,2)	3,2 [2,5 ; 4,1]
90-119	175 (2,5)	96 (1,4)	1,8 [1,4 ; 2,3]
120-149	155 (2,3)	78 (1,1)	1,9 [1,5 ; 2,6]
150-179	126 (1,8)	92 (1,3)	1,3 [1,0 ; 1,7]
180-209	98 (1,4)	79 (1,1)	1,2 [0,9 ; 1,6]
210-239	91 (1,3)	82 (1,2)	1,1 [0,8 ; 1,4]
240-269	95 (1,4)	96 (1,4)	0,9 [0,7 ; 1,3]

*les valeurs ont été arrondies à la première décimale

La Figure 13 représente graphiquement l'évolution du risque veineux thrombo-embolique au cours des mois suivant la pose d'une prothèse totale de hanche. Sur cette figure, l'odds ratio pour chaque intervalle de 30 jours ainsi que son intervalle de confiance à 95% sont présentés. La ligne bleue horizontale correspond à un odds ratio égal à 1. On observe une décroissance très rapide du risque après 30 jours puis un excès de risque significatif au cours des 6 premiers mois ; au delà de cette date, nous n'observons plus d'excès de risque.

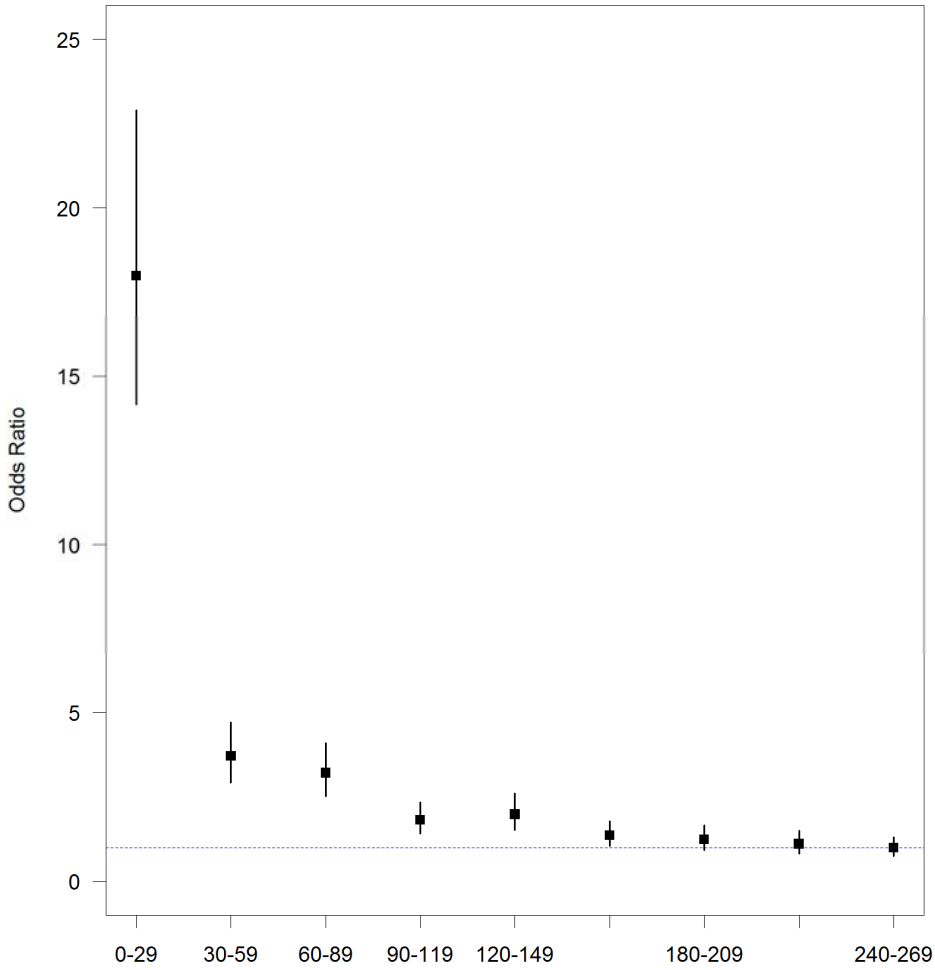


Figure 13 - Risque d'évènement veineux thrombo-embolique selon l'intervalle en jours après la pose d'une prothèse totale de hanche

2.3.4 Discussion

Nous avons évalué le risque veineux thrombo-embolique au décours d'une pose de prothèse totale de hanche à partir des données françaises de la base des hospitalisations dans le secteur MCO. Le risque est retrouvé très élevé au cours du premier mois qui correspond à la période au cours de laquelle l'anticoagulation est recommandée de façon systématique. Cet excès de risque persiste au moins 6 mois après la pose de la prothèse totale de hanche, c'est à dire qu'un excès de risque est retrouvé sur une période qui s'étend largement au delà de la période de recommandation de l'anticoagulation. Si l'on compare le risque retrouvé au cours de la période allant de 0 à 29 jours au risque retrouvé pour la période allant de 30 à 59 jours, on constate néanmoins une forte diminution du risque passant de 18,0 [14,1 ; 22,8] à 3,7 [2,9 ; 4,7]. Il est important de noter que l'on retrouve un risque élevé au cours du premier mois malgré l'anticoagulation systématique en France au cours de cette période.

Il semble pertinent de pouvoir comparer le résultat de notre analyse à celui d'une étude qui s'intéresserait de façon analogue au risque hémorragique suite à une pose de prothèse de hanche.

2.4 Quatrième publication - Proposition d'un outil web permettant le suivi des dispositifs médicaux implantables

L'étude présentée dans cette partie a porté sur la construction d'un outil web interactif réutilisant la base nationale du PMSI afin de permettre une analyse exploratoire des effets indésirables survenant secondairement à la pose d'un dispositif médical implantable (DMI). L'utilisateur cible de cet outil est un pharmacoépidémiologiste (ou un spécialiste de matériovigilance) souhaitant explorer à la volée la relation entre un DMI donné et un effet d'intérêt (identifié par le(s) motif(s) ultérieur(s) d'hospitalisation).

Les données utilisées pour cette analyse sont les données de la base nationale du PMSI. Elles permettent la tarification des séjours hospitaliers dans le champ MCO. Nous les avons présentées en détail dans les parties « 1.2.1.4 Données hospitalières médico-administratives : base nationale du PMSI » et « 1.2.1.5 Présentation et comparaison des données réutilisées dans cette thèse ». Plus précisément, nous utilisons pour cette étude :

- les DMI (codés selon la terminologie LPP) qui sont disponibles pour les seuls hôpitaux publics, dans le fichier FICHCOMP
- les actes médicaux (codés selon la terminologie CCAM)
- les diagnostics des patients au cours de chaque séjour (codés selon la CIM-10)
- les caractéristiques démographiques des patients

Nous précisons que ces données comprennent le numéro ANO qui est un identifiant unique anonyme par patient et qui permet de reconstruire le parcours de chaque patient en termes de séjours hospitaliers. Les données analysées sont celles des années 2007 à 2013 et sont structurées selon un modèle de données présenté dans la partie « 1.2.1.5 Présentation et comparaison des données réutilisées dans cette thèse ».

.

2.4.1 Introduction

Les études épidémiologiques sur les dispositifs médicaux implantables (DMI) sont classiquement basées sur des registres spécifiques ou sur la réutilisation de bases de données administratives. Elles permettent de suivre les complications survenant secondairement à l'implémentation d'un DMI. Ainsi, par exemple dans le cas des prothèses totales de hanche ou de genou, les études s'intéressant aux complications peuvent impliquer des centaines de milliers de patients [238], voire des millions de patients [239].

Des projets récents ont évalué l'intérêt de grandes bases de données observationnelles pour le suivi des effets indésirables des médicaments (EIM) après leur mise sur le marché. Plusieurs méthodes ont été comparées [158] pour la détection des EIM et un cadre méthodologique mature permet désormais d'envisager l'analyse en routine de ces grandes bases de données pour la surveillance de survenue d'EIM. L'utilisation de grandes bases de données observationnelles apporte la puissance statistique suffisante pour l'identification d'évènements rares. Une première catégorie de méthodes permet l'identification de nouveaux couples « médicament-effet », une seconde catégorie de méthodes permet la surveillance de survenue d'un EIM déjà connu. Il semble que ces méthodes pourraient être « translatées » dans le champ des DMI. De telles approches semblent d'autant plus nécessaires que l'évaluation pré-commercialisation des DMI n'a pas jusqu'ici été aussi détaillée que pour les médicaments en termes d'efficacité et de tolérance.

Les outils de visualisation des données sont utilisés dans l'exploitation qui peut être faite de ces grandes bases de données. Ils jouent un rôle par exemple dans le champ de la surveillance épidémiologique de clusters dans le temps et l'espace [240], en particulier pour les maladies infectieuses [241]. Ils permettent la formulation d'hypothèses qui doivent ensuite être vérifiées selon une méthodologie rigoureuse afin de s'extraire d'un possible biais d'observation. Ces outils peuvent permettre la visualisation d'un résultat agrégé pour une ou plusieurs strates d'intérêt. C'est le cas de l'outil développé dans le cadre du projet « LIBRA » [242] qui permet de construire des cohortes rétrospectives de patients recevant un médicament donné à partir d'une base de données médicale et administrative. Ces cohortes sont ensuite suivies pendant par exemple 12 mois et un taux de survenue d'un évènement d'intérêt peut alors être calculé. Un outil similaire pourrait être utilisé pour suivre les patients ayant eu un séjour de pose de DMI. De plus, les évènements d'intérêt étant temps-dépendants, il semble utile de les traiter comme tels dans le contexte de l'analyse exploratoire en permettant la construction dynamique de courbes de Kaplan-Meier.

L'objectif de ce travail est de construire un outil web interactif réutilisant la base nationale du PMSI dans le secteur Médecine Chirurgie Obstétrique (MCO) afin de permettre la réalisation d'une analyse exploratoire des motifs de réhospitalisation après la pose d'un DMI. Cet outil pourrait permettre à un pharmacoépidémiologiste (spécialiste de matériovigilance) d'explorer à la volée la relation entre une pose de DMI et un évènement indésirable potentiel.

2.4.2 Méthode

La base des hospitalisations dans le secteur MCO est utilisée pour cette étude. Elle représente 150 355 319 séjours en France sur la période 2008-2013. Ces données ont été obtenues auprès de l'ATIH après obtention de l'autorisation auprès de la CNIL. Un résumé de sortie anonymisé est produit pour chaque hospitalisation dans un hôpital public ou privé français. Ce résumé de sortie anonymisé permet le calcul d'un tarif pour chaque séjour hospitalier et ces données sont recueillies à cet effet par l'assurance maladie et l'ATIH. Le Tableau 15 présente les données disponibles pour chaque séjour hospitalier en France dans le secteur MCO : elles incluent notamment les codes diagnostiques selon la CIM-10, les actes médicaux selon la CCAM, les DMI posés selon la classification LPP, l'âge, le sexe et le numéro anonyme unique par patient. La pertinence de cette base a été évaluée sur la période 2007 à 2010 et le chaînage était correct en moyenne dans 95,43% des séjours sur cette période [96]. Enfin, l'intervalle en jours entre deux séjours hospitaliers peut être calculé exactement à partir de cette base.

Tableau 15 - Types de données disponibles pour chaque séjour hospitalier

Variables	Terminologie
Diagnostics	CIM-10
Actes	CCAM
Age et genre	-
Identifiant unique anonyme par patient	Code unique « ANO »
Dispositif médicale implantable	LPP
Hôpital	FINESS
Durée de séjour, année et mois d'hospitalisation	-
Lieu de résidence du patient	Code géographique

Toutes ces données ont été initialement collectées pour des raisons médico-administratives incluant la facturation du séjour hospitalier. L'identifiant unique anonyme par patient est utilisé pour suivre les réadmissions suivant un séjour comprenant une pose de DMI. Parmi les codes diagnostiques, depuis le 1er mars 2009, le « diagnostic principal » correspond au motif d'hospitalisation tel qu'il est connu à la sortie du patient de l'hôpital.

Tous les séjours hospitaliers des patients ayant au moins une fois un séjour comprenant un code de DMI sont extraits. Cela représente environ 400 000 séjours hospitaliers par an. On retrouve en moyenne environ 4 DMI par séjour hospitalier puisqu'il y a un code différent pour chaque élément composant le DMI. Par exemple, dans le cas de la prothèse totale de hanche, on peut retrouver un code pour la tête, un code pour l'insert, un code pour la tige, etc.

Un outil web est développé en utilisant PHP, MySQL et Javascript. Les graphiques présentés à l'utilisateur sont construits en utilisant des librairies Javascript gratuites ou programmées *de novo*, en particulier pour la construction à la volée d'une courbe de Kaplan-Meier. Les requêtes sont optimisées par le pré-traitement des données contenues dans les tables : les données sont agrégées pour chaque graphique selon des critères compatibles avec la granularité proposée à l'utilisateur pour sa requête.

Deux dénombrements sont réalisés :

1. un premier permettant de connaître le nombre total de DMI
2. un second permettant de connaître le nombre total de séjours hospitaliers ayant au moins un des DMI choisis par l'utilisateur

Les terminologies et leur hiérarchie sont également incluses dans la base de données.

2.4.3 Résultats

Le résultat consiste en une application web utilisable avec un navigateur web standard. Après s'être identifié, l'utilisateur sélectionne un DMI ou un groupe de DMI à partir d'un arbre contenant la classification hiérarchique des DMI (la classification LPP) ; cet arbre est présenté sur la Figure 14. Le fait de cliquer sur des codes finaux ou des nœuds de cet arbre ajoute les codes correspondants à un premier champ nommé "*Medical device identifier*" sur la Figure 15. L'utilisateur peut ensuite filtrer sa requête en spécifiant la (ou les) année(s) et l'établissement (ou les établissements) d'intérêt. Ensuite, quatre boutons de soumission du formulaire sont disponibles et redirigent l'utilisateur vers quatre rapports standardisés. Ces boutons sont « *Demography* », « *Geography* », « *Origin of the patient* », et "*Rehospitalisation*" (les pages correspondantes sont présentées ci-dessous). Le temps d'exécution de la requête la plus volumineuse incluant toutes les années, tous les hôpitaux et tous les DMI ne prend que quelques secondes : le pré-traitement des données a permis de diviser par un facteur 500 ce temps d'exécution.

Figure 14 - Arbre contenant la classification hiérarchique des Dispositifs Médicaux Implantables

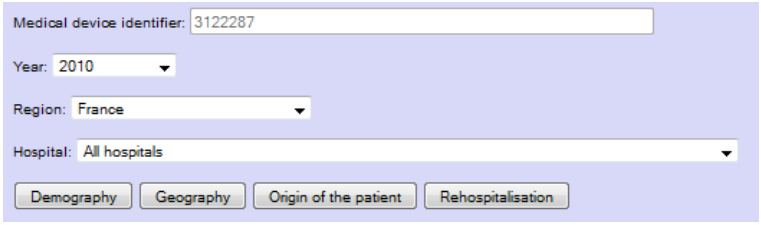


Figure 15 - Ecran proposé à l'utilisateur pour définir la requête

Nous détaillons maintenant le résultat obtenu pour chaque bouton :

- Bouton « *Demography* » : les séjours hospitaliers comprenant une pose de DMI sont décrits selon un histogramme montrant la répartition des séjours dans le temps (Figure 16), un histogramme de la durée de séjour, un histogramme de l'âge du patient au cours de ce séjour (Figure 20), un diagramme circulaire du sexe, un diagramme en barres des diagnostics les plus fréquents (Figure 18), un diagramme en barres des actes thérapeutiques les plus fréquents (Figure 17) et un diagramme en barres des codes de DMI les plus fréquents (Figure 19). Il ne s'agit pas du seul DMI sélectionné, mais de tous les DMI référencés dans les séjours comportant au moins un des codes de DMI de la requête de sélection. Nous précisons que ce qui est habituellement appelé un dispositif médical (par exemple une prothèse de hanche) correspond dans les bases de données et donc dans cette application à un ensemble de matériels référencés par des codes (par exemple la tige, la tête, l'insert, le cotyle, les vis, etc.).

Figure 16 - Illustration de la description temporelle des séjours

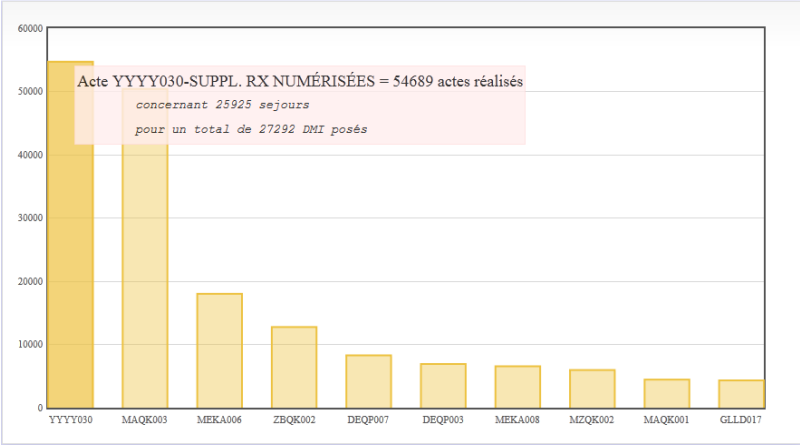


Figure 17 - Illustration de la description des actes fréquents des séjours

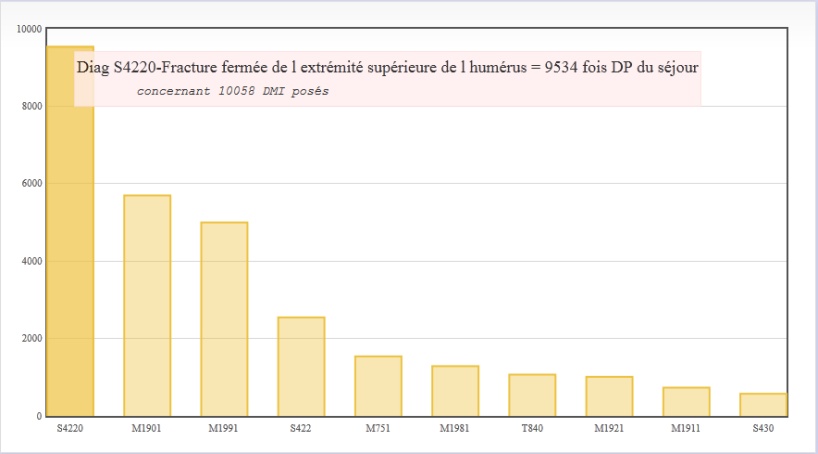


Figure 18 - Illustration de la description des diagnostics principaux fréquents au cours de ces séjours

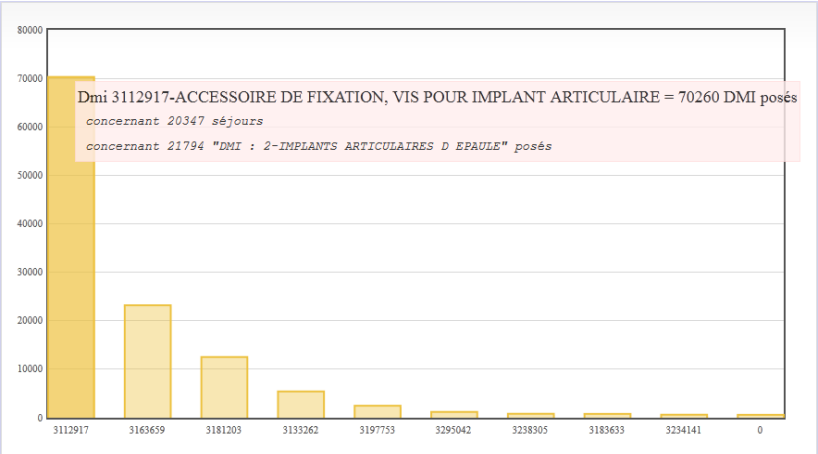


Figure 19 - Illustration de la description des DMI fréquemment associés au cours des séjours

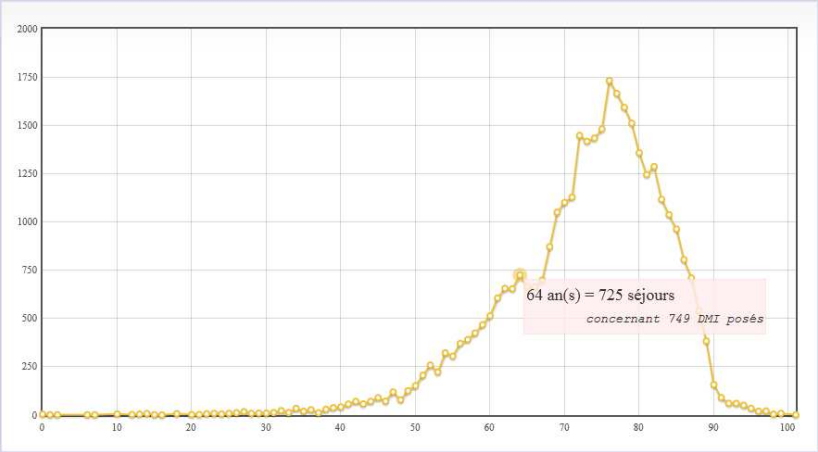


Figure 20 - Illustration de l'histogramme de l'âge des patients pour les séjours sélectionnés

- Bouton « *Geography* » : les séjours hospitaliers sont présentés selon une carte de France interactive telle qu'illustrée sur la Figure 21. En passant la souris au dessus de chaque région, l'utilisateur voit s'afficher sur la partie droite de l'écran le nombre de séjours hospitaliers et de DMI posés dans les hôpitaux de la zone correspondante.

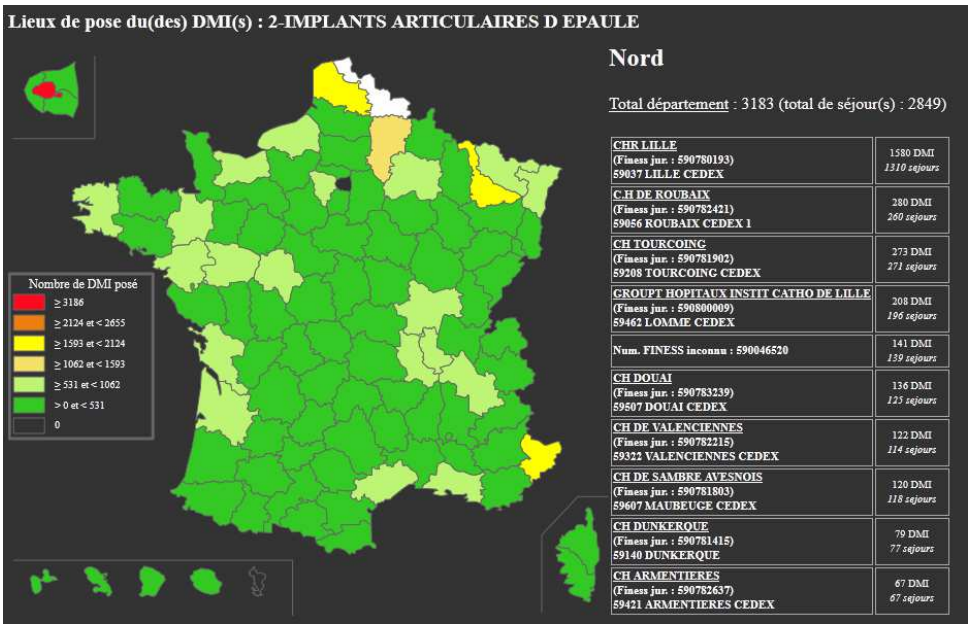


Figure 21 - Répartition géographique des séjours. Dans cet exemple, le curseur de la souris est positionné sur le département du Nord (59), le faisant apparaître en blanc et révélant sur la droite les effectifs par établissement de ce département.

- Bouton « *Origin of the patient* » : l'origine géographique des patients ayant eu un séjour de pose de DMI est présentée sur une carte dynamique analogue à celle présentée ci-avant sur la Figure 21. Si l'utilisateur filtre une requête sur un hôpital (ou zone) spécifique alors la carte permet de visualiser la zone d'attraction de l'hôpital (ou de sa zone).
- Bouton « *Rehospitalisation* » : un tableau de contingence contenant les motifs de réhospitalisation est affiché. On s'intéresse dans ce cas au diagnostic principal (DP) des séjours de réhospitalisation faisant immédiatement suite au séjour présentant la pose du DMI d'intérêt. Les DP sont dénombrés par défaut pour les 12 mois suivant le séjour de pose. Ce tableau est trié par ordre décroissant de fréquence de survenue de séjours de réhospitalisation pour ce DP. Une ou plusieurs lignes peuvent être sélectionnées, il est alors possible de construire la courbe de Kaplan-Meier correspondante. Dans cette courbe, l'événement est la réhospitalisation et le délai est le délai jusqu'à cette réhospitalisation. Un exemple de sortie est affiché sur la Figure 22 : la partie de gauche montre l'ensemble des motifs les plus fréquents, et la courbe de Kaplan-Meier à droite concerne les seules réhospitalisations pour complications mécaniques des prothèses totales de hanche (comme par exemple un descellement de prothèse).

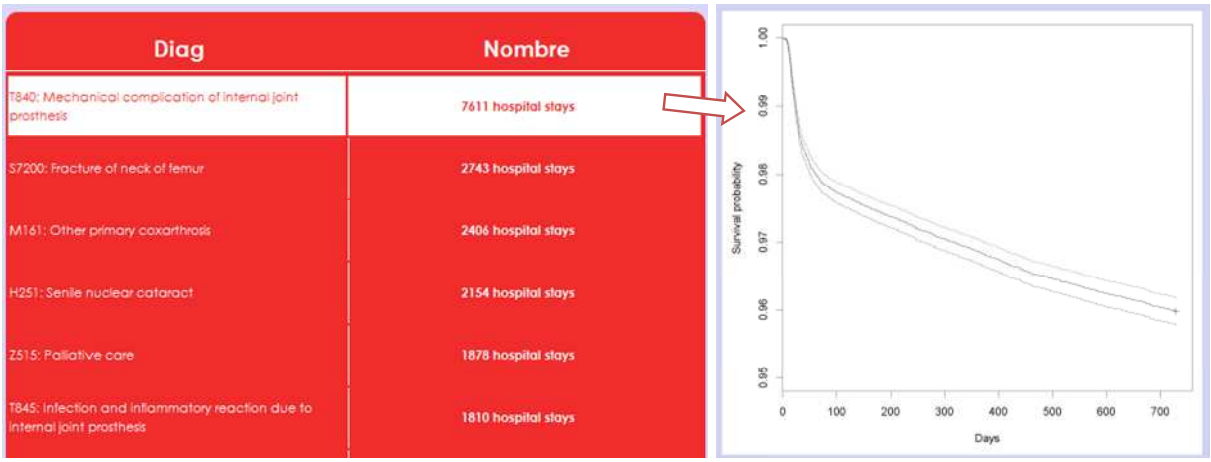


Figure 22 - Motifs de réhospitalisation et courbe de Kaplan Meier correspondant aux réhospitalisations pour complications mécaniques.

2.4.4 Discussion

Un outil exploratoire dédié au suivi des patients ayant eu une pose de DMI a été construit. Il permet à l'utilisateur de visualiser des descriptions temporelles, démographiques et géographiques des séjours hospitaliers avec pose de DMI. Il permet également de suivre la survenue de réhospitalisations (qui sont des évènements temps-dépendants) par la construction dynamique de courbes de Kaplan-Meier. Ce travail nous semble confirmer l'intérêt de la réutilisation de bases de données médico-administratives pour la mise en évidence d'effets indésirables pouvant survenir secondairement à l'implémentation d'un DMI.

Plusieurs limites liées à la réutilisation de la base nationale du PMSI doivent être discutées. Premièrement, cette base ne contient pas d'identifiant unique de DMI, contrairement aux registres dédiés par exemple aux Etats-Unis [243]. De plus, les séjours hospitaliers avec une pose de DMI ne sont accessibles que dans les hôpitaux publics donc l'information présentée dans notre outil ne concerne pas les cliniques privées. Cela n'est néanmoins pas un obstacle pour le suivi des complications car les informations sur les diagnostics (ou les actes médicaux) sont disponibles pour l'ensemble des séjours hospitaliers en France. Enfin, un travail conséquent sur la qualité des données a été réalisé et doit être poursuivi.

La construction de la courbe de Kaplan-Meier est très utile pour le suivi épidémiologique. Néanmoins, elle ne tient pas compte des décès dont une forte proportion n'est pas visible dans le PMSI, les décès pouvant survenir en-dehors de toute hospitalisation. Si des décès surviennent dans les 12 mois suivant l'intervention, cela amène à surestimer la durée de suivi des données censurées et donc à légèrement surestimer la fonction de survie.

L'utilisabilité de cet outil est actuellement évaluée. Cet outil permettant le suivi des patients ayant eu une pose de DMI pourrait être généralisé à d'autres types de données pour suivre, par exemple, les complications survenant secondairement à la réalisation de certains actes chirurgicaux. Le suivi des maladies chroniques pourrait également bénéficier de ce type d'outil. Enfin, des analyses statistiques utilisant des méthodes de fouille de données sont envisagées afin de connecter cet outil avec le logiciel d'analyse statistique R [244].

3 Discussion

L'objectif de ce travail était d'étudier la réutilisation de bases de données hospitalières de grande dimension pour la mise en évidence d'effets indésirables. Cet objectif principal s'est décliné en 2 objectifs opérationnels réalisés sur deux bases de données hospitalières que sont d'une part la base nationale des données du PMSI de 2007 à 2013 et d'autre part une base d'un centre hospitalier (CH) partenaire pour la même période. Afin de répondre à ces objectifs opérationnels, nous avons réalisé quatre études : deux ont porté sur la réutilisation de données hospitalières pour la recherche d'effets indésirables médicamenteux et deux ont porté sur la réutilisation de données hospitalières pour le suivi des dispositifs médicaux implantables.

Dans une première partie « 3.1 Synthèse générale des résultats », nous rappellerons le contenu de chaque étude ainsi que les principaux résultats puis, dans une seconde partie « 3.2 Intérêt et limites de la réutilisation de données » nous comparerons la réutilisation de données aux études plus traditionnelles organisant un recueil actif des données et pour lesquelles le recueil comporte d'emblée une finalité épidémiologique. Toujours dans cette seconde partie, les avantages et les inconvénients de la réutilisation de données seront présentés en suivant en guise de plan le cours naturel de réalisation d'une étude.

3.1 Synthèse générale des résultats

Nous rappelons ici le contenu de chacune des quatre études ainsi que leurs principaux résultats.

Dans une première étude [210], nous avons proposé une méthode pour évaluer l'impact d'un médicament, dans son contexte de prescription (incluant le contexte séquentiel), sur des résultats de biologie médicale. Cette méthode nous a permis de calculer la variation moyenne d'un résultat de biologie médicale (en l'occurrence la kaliémie) associée à une prescription médicamenteuse dans son contexte fréquent de prescription. Cela a été illustré avec la supplémentation potassique et semble confirmer la nécessité de prendre en compte le contexte clinico-biologique d'une administration de médicament. Ce travail a été réalisé dans un contexte observationnel et permet à un pharmacoépidémiologiste de rechercher un effet attendu ou non survenant secondairement à l'administration d'un médicament. Ce type d'analyse pourrait également permettre d'apprécier l'efficacité d'un médicament dans le cadre d'études observationnelles post-commercialisation. Les contextes d'administration médicamenteuse et de résultat de biologie médicale ont été pris en compte ; ainsi cette méthode pourrait permettre d'explorer l'impact de certaines co-prescriptions sur certains paramètres biologiques.

Dans une deuxième étude [211], nous avons construit et évalué un jeu de règles complexes de détection pour mettre en évidence des effets indésirables médicamenteux à type d'hyperkaliémie par l'analyse d'une base de données hospitalière. Cette détection automatisée a retrouvé des valeurs élevées pour le rappel et la précision lors de la comparaison à la revue experte des séjours. En termes de rappel, 89,5% des EIM à type d'hyperkaliémie ont été retrouvés. De plus, tous les EIM graves ont été détectés ce qui semble un résultat important ;

il est néanmoins difficile de généraliser ce résultat compte tenu du nombre très réduit d'EIM de ce type. En termes de précision, 63,7% des EIM identifiés automatiquement étaient véritablement des EIM à type d'hyperkaliémie. Ce travail qui s'inscrit pleinement dans le prolongement du projet PSIP [139,234,245–247] semble confirmer la nécessité de prendre en compte le contexte clinique et biologique du patient pour la détection automatisée d'EIM [248–250].

Ensuite, dans une troisième étude, nous avons évalué le risque thromboembolique au décours d'une pose de prothèse totale de hanche à partir des données françaises de la base des hospitalisations dans le secteur MCO. Le risque est retrouvé très élevé au cours du premier mois qui correspond à la période au cours de laquelle l'anticoagulation est recommandée de façon systématique. Cet excès de risque persiste au moins 6 mois après la pose de la prothèse totale de hanche, c'est à dire qu'un excès de risque est retrouvé sur une période qui s'étend largement au delà de la période de recommandation de l'anticoagulation. Si l'on compare le risque au cours de la période allant de 0 à 29 jours au risque de la période allant de 30 à 59 jours, on retrouve néanmoins une forte diminution du risque passant de 18,0 [14,1 ; 22,8] à 3,7 [2,9 ; 4,7]. Il nous semble important de noter que l'on retrouve un risque élevé au cours du premier mois malgré l'anticoagulation systématique en France au cours de cette période.

Enfin, dans une quatrième étude, un outil web dédié au suivi des patients ayant eu une pose de DMI a été construit. Il permet à l'utilisateur de visualiser des descriptions temporelles, démographiques et géographiques des séjours hospitaliers avec pose de DMI. Il permet également de suivre la survenue de réhospitalisations (qui sont des évènements temps-dépendants) par la construction dynamique de courbes de Kaplan-Meier. Ce travail confirme l'intérêt de la réutilisation de bases de données médico-administratives pour la mise en évidence d'effets indésirables pouvant survenir secondairement à l'implémentation d'un DMI.

Ces quatre études montrent l'intérêt de la réutilisation de données hospitalières pour la mise en évidence d'effets indésirables liés aux médicaments et à la pose de DMI. La faisabilité incluant la méthodologie informatique et statistique a également été étudiée. Ces bases de données hospitalières rendent possible l'exploration de couples exposition-évènement de natures variées : ils peuvent l'un et l'autre être aigu ou chronique ; ils peuvent l'un et l'autre être un diagnostic ou une prise en charge au sens large, qu'elle soit diagnostique, chirurgicale ou médicamenteuse.

Plus généralement, la réutilisation de ces bases de données peut s'envisager selon au moins trois axes :

- Le premier axe est celui de surveillance épidémiologique (épidémiologie descriptive) vis-à-vis de la survenue d'évènements [154,188,190,235,251]. Ainsi, de véritables protocoles de surveillance systématique peuvent être envisagés vis-à-vis de diagnostics d'intérêt et selon une stratification qui pourrait s'envisager à des niveaux de granularité allant du médecin à un hôpital dans sa globalité. Ce premier axe rencontre néanmoins plusieurs limites que nous discuterons par la suite.

- Le deuxième axe est celui de l'épidémiologie analytique dont un cas particulier est la mise en évidence de certains effets indésirables secondaires à une prise en charge [16,69,218,232]. Il est exclu de renoncer à la pharmacovigilance sous sa forme actuelle, c'est-à-dire fondée sur les déclarations spontanées, comme cela avait été évoqué préalablement [252]. Néanmoins, la réutilisation de ces données telle qu'elle a été envisagée par la FDA, selon une méthodologie éprouvée, constitue un progrès certain. Notre travail met en œuvre les méthodes en cross-over [160,192,197] utilisant le patient comme son propre témoin pour la mise en évidence de couples « médicament-évènement ». On peut considérer que les mises en évidence de tels couples seront plus efficientes lorsque des listes exhaustives d'évènements aigus et de prises en charge aiguës auront pu être préalablement proposées afin d'être explorés en routine. Une évaluation rigoureuse de la puissance statistique disponible par l'exploitation de ces bases pour chaque couple « médicament-évènement » d'intérêt pourrait inciter dans certains cas à la définition d'évènements intermédiaires plus fréquents tels que les fluctuations de paramètres de biologie médicale explorées dans ce travail. Il semble que le nombre de données recueillies en routine pour chaque patient poursuivra son développement, néanmoins le nombre de patients analysés dépassera sans doute plus difficilement les volumes évoqués par la FDA. Un nombre plus élevé de patients pourrait être obtenu en colligeant des données issues de pays différents ce qui pose naturellement des contraintes non triviales d'interopérabilité, y compris pour la part administrative de ces données hospitalières. Enfin, l'exploration de ces bases de données pourrait dans certains cas suggérer des pistes thérapeutiques fondées sur des médicaments déjà utilisés en routine dans d'autres indications.
- Le troisième axe est celui des études écologiques cherchant à mettre en évidence l'association entre certaines expositions environnementales et la survenue de pathologies. Ces bases administratives contiennent le lieu de résidence des patients ainsi que leur lieu d'hospitalisation, on peut ainsi envisager des études écologiques à partir de cette base, cette dernière piste n'a toutefois pas été explorée dans le cadre de notre travail.

La qualité des résultats pouvant être obtenus à partir de ces données observationnelles doit être discutée. Tout d'abord, la qualité de ces résultats est intriquée avec la nature du design d'étude retenu : les designs d'études utilisant le patient comme son propre témoin et traitant les expositions comme des variables dépendantes du temps ont prouvé leur intérêt méthodologique lorsqu'ils sont applicables [159,160,197]. Les méthodes exploratoires d'évènements indésirables, dont on peut souligner ici la relative complexité sur un plan statistique, semblent avoir atteint un niveau de maturité [158] compatible avec leur utilisation en routine.

Toujours concernant la qualité des résultats issus d'études observationnelles, Schneeweiss [253] a comparé les résultats obtenus par une étude observationnelle à ceux rendus par des essais randomisés sur le même sujet et a montré la similarité des résultats en particulier lorsque des restrictions successives sont appliquées sur les données. Il semble ainsi utile de ne garder parfois qu'un sous-échantillon réduit mais homogène à partir de l'ensemble

des séjours potentiellement compatibles dans une base de données administratives. Golder [254] a de même montré que les études observationnelles pouvaient dans certains cas obtenir des résultats « similaires » aux essais randomisés. Il serait intéressant de comparer dans les essais en cross-over la différence obtenue entre les deux bras au cours de la première période à celle obtenue dans chaque bras sur l'ensemble de l'étude. Cela permettrait de comparer les études quasi-expérimentales aux études randomisées. Idéalement, cela devrait être fait pour les quelques essais en cross-over ayant pu montrer une différence significative dans la survenue d'un effet indésirable. En effet, le fait de s'intéresser à un évènement non associé à la mesure expérimentale du critère de jugement principal expose moins au problème d'interprétation que pose la régression vers la moyenne [191].

3.2 Intérêt et limites de la réutilisation de données

La réutilisation de données présente des intérêts particuliers, que nous présenterons ici en la comparant aux études plus traditionnelles organisant un recueil actif des données et pour lesquelles le recueil comporte d'emblée une finalité épidémiologique. Les avantages et les inconvénients de la réutilisation de données [255] seront présentés ci-après en suivant en guise de plan le cours naturel de réalisation d'une étude [256]. Nous discuterons successivement :

- la nature de la question posée
- le recueil de données
- l'agrégation des données
- l'analyse statistique
- les validités interne et externe des résultats

Nous examinerons ces différents points dans les paragraphes suivants.

3.2.1 Nature de la question posée

Cette section montrera en quoi la réutilisation de données se distingue des études classiques du point de vue de la question scientifique posée.

Tout d'abord, la réutilisation de données permet de se positionner dans un contexte exploratoire [140,217,257]. Ainsi, il est possible de générer des hypothèses à partir des données et de ne pas figer d'emblée la question posée par l'étude. Nous détaillons plus loin la nécessité d'être vigilant néanmoins vis à vis de cette attitude exploratoire en terme d'inflation du risque alpha [258].

Ensuite, le fait d'étudier ce que le médecin a choisi de prescrire et selon les modalités qu'il a souhaitées peut entrainer un biais d'observation [166,167,169]. Par ailleurs, les données les plus nombreuses pouvant être analysées à partir des bases médico-administratives correspondent à des paramètres d'usage fréquent : en conséquence, la réutilisation de données ne permet pas, par exemple, l'étude d'expositions expérimentales tels que des paramètres de biologie médicale innovants. Inversement, le fait de travailler sur des données réelles apporte une information nouvelle par rapport à celle issue d'essais cliniques comportant le plus

souvent un profil de patients particuliers suivis dans un contexte protocolisé. Cette caractéristique de la réutilisation de données est souvent énoncée à travers l'expression « population-based ». Elle constitue un avantage, car elle correspond plus à la réalité, mais aussi un inconvénient, car elle perturbe l'interprétation de l'effet propre d'une exposition.

De même, la réutilisation de bases de données observationnelles permet la mise en évidence d'effets indésirables uniquement pour les médicaments suffisamment fréquents dans la base de données, donc des médicaments de routine et donc pas des médicaments très rares ou récents. Il y a ainsi de nombreux couples médicament-effet que l'on ne peut pas étudier [259] à ce jour, de façon inférentielle, par la réutilisation de ces données. Il en est de même pour les mesures de biologie rares ou très récentes.

A ce jour, les données hospitalières médicales permettent de rechercher des effets indésirables médicamenteux survenant en aigu (elles concernent un séjour hospitalier) alors que les données médico-administratives hospitalières permettent également la recherche d'effets indésirables médicamenteux survenant de façon chronique (elles peuvent concerner une succession de séjours hospitaliers). Une voie intermédiaire nous semble de pouvoir connecter des données intra-hospitalières (voire des bases de données inter-hospitalières) aux données médico-administratives [260]. On pourrait ainsi imaginer rechercher parmi les paramètres de biologie médicale (parmi les données hospitalières médicales) ceux pouvant être des déterminants de la survenue d'un accident vasculaire cérébral (parmi les données hospitalières médico-administratives).

La réutilisation de données impose de plus un périmètre d'inclusion limité. Par exemple, dans le cas des thromboses [111], il n'est pas possible à partir d'une base de données hospitalière de suivre rigoureusement les phlébites du membre inférieur puisque ces dernières ne font pas systématiquement l'objet d'une hospitalisation. Pour cette raison, nous avons choisi dans notre troisième étude de suivre les embolies pulmonaires, car *a priori* la grande majorité des embolies pulmonaires ont une manifestation clinique conduisant à l'hospitalisation. Néanmoins, là encore, il est certain que certaines embolies pulmonaires massives entraînant un décès ambulatoire sont de fait exclues de l'étude, et font penser à tort que le patient survit en ambulatoire sans complication.

Enfin, l'originalité de ces données ne nous semble pas résider dans leur nature mais plutôt dans le fait qu'elles soient recueillies de façon systématique en routine permettant ainsi l'exploration d'associations originales entre des expositions et des évènements, fussent-ils de natures différentes (par exemple DMI et diagnostics).

3.2.2 Recueil de données

Cette section montrera en quoi la réutilisation de données se distingue des études classiques du point de vue du recueil de données.

Tout d'abord, la réutilisation de données concerne presque toujours des cohortes historiques [72] (rétrospectives) : on peut ainsi considérer *a priori* que le niveau de preuve

issu de ces études est supérieur à celui des études cas-témoins mais en général moins bon que celui issu de cohortes prospectives ou d'un véritable essai randomisé [179].

Ensuite, la réutilisation de données présente plusieurs avantages. Elle permet des études à bas coût marginal : si on considère que le coût du recrutement des patients et du recueil de données doit être affecté à l'activité transactionnelle (par exemple le soin) qui a de toute manière eu lieu, alors le coût de la réutilisation de données est un coût marginal, qui ne tient qu'à l'extraction des données et à son analyse. La réutilisation de données permet en outre d'obtenir des résultats très rapidement : on peut appliquer le même raisonnement que précédemment, et considérer que la durée marginale de l'étude est donc limitée à l'extraction et l'analyse. Elle permet de plus de générer immédiatement des données historiques sur une longue période (depuis 2007 par exemple pour les données du PMSI MCO chaînables), et de reconstruire cet historique immédiatement, ce qui est très utile lorsque la méthode d'investigation change, et n'est inversement pas possible dans les études prospectives traditionnelles.

Par ailleurs, concernant l'interopérabilité des données [261], un vrai point à surmonter concerne la fusion de données issues d'hôpitaux utilisant des systèmes d'information différents (dans l'espace) et à des moments différents (dans le temps). Aujourd'hui, par exemple, les données du PMSI sont entièrement interopérables mais cette situation reste relativement rare. Elles le sont car elles font l'objet d'une contrainte financière. L'ajout de données nouvelles dans le champ du PMSI permet l'extension de ce champ de l'interopérabilité car elle rend de fait leur recueil obligatoire. L'utilisation d'ontologies [262] semble dans ce contexte un point clé de l'interopérabilité sémantique.

L'extension des bases de données hospitalières peut, d'après nous, s'anticiper selon deux axes :

1) Tout d'abord, on peut imaginer la construction de bases de données comprenant davantage d'individus, par la construction de bases internationales [263]. Cette première piste ne nous semble pas le point le plus déterminant tant la puissance statistique des bases actuelles est d'ores et déjà satisfaisante.
2) Ensuite, on peut anticiper l'avènement d'un plus grand nombre de données numérisées. Il pourrait tout d'abord s'agir de données « omiques », ce terme désignant les données de la génomique, de la protéomique ou issues de l'étude du métabolisme. Par exemple, l'apport de la génomique dans la pratique clinique ne concerne pour l'instant qu'un nombre limité de disciplines telles que l'hématologie [264]. Cependant, le séquençage systématique et intégral du génome ne devrait plus tarder, tant les coûts sont devenus raisonnables, de l'ordre de quelques milliers d'euros. La génétique translationnelle telle qu'elle est initiée aujourd'hui de façon ponctuelle en France, pourrait changer quelque peu le cadre méthodologique des projets réutilisant les données hospitalières. Certains dispositifs médicaux génèrent des données qui sont enregistrées en routine dans les bases de données. Cela permet également d'anticiper un développement du volume d'information associé à ces dispositifs. Ainsi, la réalisation de dosages portatifs systématiques ou l'utilisation de

systèmes de surveillance tels qu'ils existent déjà en télécardiologie, sont des pistes permettant d'anticiper l'entrée des données hospitalières dans le champ du « big data » [2].

Nous insistons sur le fait que ces données sont pertinentes pour la recherche d'effets indésirables à la condition que le numéro permettant le chaînage des séjours de chaque patient (le numéro ANO) continue à être rendu accessible aux analystes de données par la CNIL.

Concernant la qualité des données, il est communément admis qu'elle est guidée par la finalité du recueil. Ainsi, la qualité des données du PMSI dans le secteur MCO a été croissante depuis l'avènement de la tarification à l'activité et cela est particulièrement vrai pour les codes d'actes ou de diagnostics impactant directement la valorisation d'un séjour hospitalier : les actes chirurgicaux sont par exemple très bien codés car ils peuvent classer un séjour dans un GHM rémunérateur mais également car ils sont payés en direct au praticien dans le cadre de l'hospitalisation privée. Inversement, l'hypertension artérielle n'influe que modérément sur la nature du GHM du séjour et ne fait pas partie des complications ou morbidités associées pouvant fortement influencer la valorisation d'un séjour : en conséquence, cette pathologie n'est pas retrouvée codée avec la même rigueur que l'acte chirurgical décrit ci-avant. Il semble, pour ces raisons, indispensable de pouvoir systématiquement évaluer la qualité du codage en le confrontant à des registres de référence. C'est également dans ce contexte que sont proposés de nombreux algorithmes [112,113,184,265] comprenant des listes de codes d'acte ou de diagnostics dans la perspective d'identifier une pathologie au sein d'une base médico-administrative. Ces algorithmes font pleinement partie de la problématique d'agrégation des données que nous évoquons dans le paragraphe suivant.

3.2.3 Agrégation des données

Cette étape complexe est une caractéristique des études de réutilisation de données, qui est classiquement absente des protocoles classiques.

La réutilisation de données met les chercheurs face à la complexité des données à traiter. Cette complexité dépasse largement le nombre d'individus et de variables [199]. On peut citer notamment le nombre élevé de modalités possibles pour les variables qualitatives (plus de 32 000 codes possibles pour coder un diagnostic en CIM10 version française), le nombre élevé de tables et de relations dans une base de données, et le fait que certaines mesures soient répétées dans le temps un grand nombre de fois (comme par exemple les mesures de biologie médicale). Or les méthodes statistiques, pour leur grande majorité, doivent être alimentées par un simple tableau contenant une ligne par patient et une colonne par variable, et les variables qualitatives doivent avoir un faible nombre de modalités. Certaines méthodes sont capables de traiter des mesures répétées, à condition qu'elles ne concernent qu'un seul paramètre, qui est alors l'unique objet de l'étude. L'obtention d'une forme de données constituée par un tableau unique tel que décrit ci-dessus passe par une étape d'agrégation de données. Cette étape a généralement les effets suivants :

- elle conserve le nombre d'individus
- elle réduit le nombre de variables

- elle réduit le jeu de données à une seule table, sans aucune relation
- elle réduit considérablement le nombre de modalités des variables qualitatives, et peut éventuellement les transformer en néo-variables binaires simplifiées
- elle supprime généralement la notion de mesures répétées

De notre expérience [199], la qualité de cette étape détermine la possibilité d'obtenir des résultats probants. Cette agrégation nécessite l'association de plusieurs compétences, par ordre chronologique :

- Tout d'abord, il faut connaître la nature des données médicales et leur signification, afin d'en réduire la complexité sans perdre de sens, ou parfois de faire apparaître un sens caché dans les données (par exemple, ne pas simplement lister les médicaments mais faire apparaître le fait qu'un patient prenne un inhibiteur enzymatique, ce qui est une propriété latente de certaines molécules et ne correspond à aucune classe thérapeutique en particulier).
- Ensuite, il faut comprendre quelle forme de données les méthodes statistiques sont capables de traiter.
- L'étape d'agrégation de données est également fortement influencée par la finalité du traitement. Ainsi, des codes de chirurgie seront mappés différemment selon qu'on s'intéresse à la reprise de la marche (on discriminera alors la localisation de l'intervention, la présence de fixateurs externes, la taille des plaies opératoires, etc.), qu'on s'intéresse aux effets indésirables du médicament (alors schématiquement seule la présence d'une anesthésie générale, ou une atteinte du foie ou du rein seront importantes), ou au risque de thrombose (on fera alors en sorte de mettre en évidence l'atteinte des axes vasculaires, de la fonction cardiaque, ou encore l'immobilisation généralement associée à tel acte chirurgical, sachant qu'elle n'est pas directement tracée dans les données).
- Enfin, ces transformations de données, une fois définies, nécessiteront le recours à des compétences de programmation, car le *data management* nécessaire pour les mettre en œuvre pourra souvent dépasser les opérations couramment implémentées dans des fonctions standard.

En synthèse, il nous parait indispensable de réunir des médecins, des programmeurs et des statisticiens pour réussir une bonne réutilisation de données médicales. Cela peut passer par la réalisation de cursus complémentaires ou la constitution d'équipes multidisciplinaires. Ce besoin de multidisciplinarité ne se limite pas à ces trois orientations : lorsque la réutilisation de données permet la découverte de connaissances, leur mise en œuvre nécessite une intégration beaucoup plus large dans l'organisation actuelle des pratiques de soins, et par exemple les spécialistes des facteurs humains ont montré combien les connaissances ou les logiciels ne suffisaient pas à améliorer les pratiques de soins. Enfin, la réutilisation de données permet de valoriser l'effort des acteurs de terrain qui réalisent ce codage et ainsi de donner un sens supplémentaire à leur travail.

3.2.4 Analyse statistique

Cette section montrera en quoi la réutilisation de données se distingue des études classiques du point de vue de l'analyse statistique.

Tout d'abord, la réutilisation de données permet un gain de puissance statistique permettant la mise en évidence d'évènements rares tels que les effets indésirables des médicaments.

Ensuite, l'analyse de grandes bases de données médico-administratives [266] peut se faire à l'aide de méthodes de fouille de données qui permettent de répondre à la fois à la problématique du grand nombre de variables disponibles et à celle du grand nombre d'individus statistiques. Ces analyses par fouille de données peuvent tout à fait s'envisager dans une étape d'analyse exploratoire, c'est-à-dire non centrée sur un objectif principal précis, ou encore « data-driven ».

Il faut toutefois rester vigilant vis à vis du goût naturel de chacun pour la mise en évidence d'association statistique. La philosophie même des tests statistiques illustre cette volonté de contrôler le risque de première espèce, au détriment parfois du risque de seconde espèce et donc de la mise en évidence de « découvertes ». Le contrôle du risque de première espèce implique de prendre en compte sa possible inflation dans les contextes de fouille de données et de recherche systématique de couples « médicament-évènement » décrits ci-avant. Cette tendance humaine à la trop grande confiance dans les résultats peut être décrite d'un point de vue comportemental [267]. Ainsi, la conservation du risque alpha implique parfois la correction de ce dernier lors de la réalisation de tests multiples dans un contexte confirmatoire.

Ensuite, nous insistons sur l'erreur méthodologique qui consiste à considérer la plausibilité physiopathologique d'un résultat comme un argument de validité de celui-ci. Il semble en effet simple de proposer des mécanismes biologiques et leur contraire. Cette rationalisation *a posteriori* devrait ainsi être crainte autant que l'évocation de la causalité en épidémiologie [268,269].

Par ailleurs, une hypothèse implicite de la réutilisation de ces bases de données médico-administratives est que les erreurs de codage constitueraient un biais non différentiel. En effet, si on s'intéresse à l'exploitation en épidémiologie descriptive d'une telle base de données, il est fort probable que les résultats soient faussés notamment par le sous-codage de pathologies non valorisantes. Néanmoins, lorsque l'on s'intéresse à l'association statistique entre deux paramètres contenus dans cette base, on suppose qu'il peut exister des erreurs de codage mais que ces erreurs de codage sont uniformément réparties et ne modifient ainsi que peu la valeur d'un facteur d'association. On considère implicitement que ces bases de données présentent des limites pour réaliser des analyses descriptives mais beaucoup moins de limites pour réaliser des analyses inférentielles. Cette hypothèse implicite n'a pas été rigoureusement explorée dans le cadre de notre travail et ce point constitue en conséquence une limite.

Enfin, plus généralement, et comme cela a été présenté en introduction dans la partie « 1.1.2 Définition des Big Data ou données massives », le data reuse implique souvent l'analyse de

grands volumes de données et/ou de variables. Dans ce contexte, les méthodes de fouille de données présentées en introduction dans la partie « 1.4.3.4 Construction de modèles prédictifs par fouille de données » peuvent être très utiles, en particulier pour la construction de modèles prédictifs. Parmi ces méthodes, il est important de noter que la très grande qualité prédictive des réseaux de neurones n'est pas forcément compatible avec leur utilisation en routine dans le champ médical contrairement à d'autres domaines tels que par exemple la publicité ciblée en ligne ou encore la détection de fraude bancaire.

En effet, les réseaux de neurone permettent d'obtenir des modèles probabilistes très performants (et robustes) en termes de prédiction mais ces mêmes modèles sont le plus souvent inintelligibles pour un humain car d'une trop grande complexité. Ce fonctionnement en « boîte noire » peut déstabiliser le professionnel de santé et inciter le statisticien à utiliser des méthodes moins performantes mais dont les modèles de prédiction obtenus peuvent être interprétés facilement par un expert telles que les arbres de décision et les règles d'association [1,202].

3.2.5 Validités interne et externe des résultats

Nous présentons enfin dans cette section l'intérêt de l'accessibilité à ces bases de données réutilisées dans une perspective de réplication des résultats.

L'utilisation de telles bases de données permet d'imaginer une plus grande transparence méthodologique et une meilleure reproductibilité des analyses. En effet, toutes les équipes de recherche qui le souhaitent peuvent, sous réserve d'obtention de l'autorisation CNIL, accéder à ces données médico-administratives. Cette autorisation est rendue nécessaire face au risque de réidentification des personnes, et ce même dans des données qui ne sont pas directement nominatives. La réplication de résultats nécessite la réalisation *de novo* d'une analyse sur une base identique (que l'on peut considérer comme une validation interne) avant même la réplication sur une autre base de données. On pourrait même imaginer que les journaux scientifiques puissent être détenteurs de certaines de ces bases et ré-exécutent les scripts fournis par l'auteur ayant soumis un article pour revue. On estime en effet le pourcentage d'études observationnelles dont le résultat n'est pas reproductible à 10% ou 20% [270].

Ainsi, dans les années 1980, le traitement hormonal substitutif était considéré comme protégeant les femmes contre la survenue de cardiopathies [271,272] même si certains effet indésirables étaient également décrits [273]. Puis, un essai randomisé de grande envergure [274] fut réalisé et démontra une élévation du risque de cardiopathies dans le groupe traité par traitement hormonal substitutif. Plus récemment, deux études s'intéressant à l'association entre biphosphonates et cancer de l'œsophage sur la même base de données médico-administratives (du Royaume-Uni) ont obtenu des résultats contradictoires : la première n'a pas retrouvé d'augmentation du risque de cancer alors que la seconde a retrouvé un risque doublé de survenue de cancer de l'œsophage [275,276]. Cela illustre l'absolue nécessité de pouvoir répliquer les résultats, en particulier dans le champ épidémiologique, idéalement sur des bases analogues et selon une démarche empirique qui est la seule compatible avec une démarche scientifique. La réplication entière des études et des résultats

est la seule façon de procéder tant chacun sait combien des détails apparemment insignifiants, connus de la seule personne ayant conduit l'étude, peuvent conduire à des résultats différents.

Ensuite, l'accès facilité aux bases de données médico-administratives se déroule de façon parallèle à l'ouverture des bases de données des essais randomisés aux Etats-Unis, apportant un gain évident de transparence méthodologique [83,277].

Enfin, nous considérons valable la généralisation des approches empiriques en épidémiologie, ce qui revient à considérer la biostatistique et l'épidémiologie comme des sciences expérimentales.

4 Conclusion

Depuis 20 ans, de grandes bases de données médicales ou médico-économiques ont été collectées. Elles sont prêtes à être analysées tant du point de vue de la qualité des données que de la maturité des méthodes pouvant être appliquées sur ces bases de données observationnelles.

En nous appuyant sur quatre études réutilisant deux types de bases de données hospitalières de grande dimension, nous avons montré l'intérêt de la réutilisation des bases de données administratives dans la détection des effets indésirables médicamenteux et des effets indésirables faisant suite à la pose d'un dispositif médical implantable.

La réutilisation de données permet d'obtenir rapidement des résultats originaux venant confirmer ou infirmer sur un grand nombre de cas une hypothèse pharmacoépidémiologique ce qui constitue un enjeu essentiel en santé publique. De plus, la facilitation de l'accès aux bases de données médico-administratives permet d'envisager une plus grande réplication des analyses réalisées à partir des bases de données concernées ce qui constitue un gage de qualité des résultats.

Ce travail préliminaire mériterait d'être poursuivi sur des thématiques similaires.

5 Références

[1] Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. ACM SIGMOD Rec 1993.

[2] Baro E, Degoul S, Beuscart R, Chazard E. Toward a Literature-Driven Definition of Big Data in Healthcare. BioMed Res Int 2015.

[3] Albertsen PC, Hanley JA, Fine J. 20-year outcomes following conservative management of clinically localized prostate cancer. JAMA 2005;293:2095–101. doi:10.1001/jama.293.17.2095.

[4] Sheffield KM, Riall TS, Han Y, Kuo Y, Townsend CM, Jr, et al. Association between cholecystectomy with vs without intraoperative cholangiography and risk of common duct injury. JAMA 2013;310:812–20. doi:10.1001/jama.2013.276205.

[5] Cram P, Rosenthal GE, Vaughan-Sarrazin MS. Cardiac Revascularization in Specialty and General Hospitals. N Engl J Med 2005;352:1454–62. doi:10.1056/NEJMsa042325.

[6] Hu JC, Gu X, Lipsitz SR, Barry MJ, D'Amico AV, Weinberg AC, et al. Comparative effectiveness of minimally invasive vs open radical prostatectomy. JAMA 2009;302:1557–64.

[7] Heit JA. Estimating the incidence of symptomatic postoperative venous thromboembolism: the importance of perspective. JAMA 2012;307:306–7.

[8] Cavallaro P, Rhee AJ, Chiang Y, Itagaki S, Seigerman M, Chikwe J. In-hospital Mortality and Morbidity After Robotic Coronary Artery Surgery. J Cardiothorac Vasc Anesth 2014.

[9] Poulose BK, Griffin MR, Yuwei Z, Walter S, Richards WO, Wright JK, et al. National analysis of adverse patient safety events in bariatric surgery. Am Surg 2005;71:406–13.

[10] Eappen S, Lane BH, Rosenberg B, Lipsitz SA, Sadoff D, Matheson D, et al. Relationship between occurrence of surgical complications and hospital finances. JAMA 2013;309:1599–606.

[11] Giles KA, Schermerhorn ML, O'Malley A, Cotterill P, Jhaveri A, Pomposelli FB, et al. Risk prediction for perioperative mortality of endovascular vs open repair of abdominal aortic aneurysms using the Medicare population. J Vasc Surg 2009;50:256–62.

[12] Ghaferi AA, Birkmeyer JD, Dimick JB. Variation in hospital mortality associated with inpatient surgery. N Engl J Med 2009;361:1368–75.

[13] Horton D, Mehta P, Antao VC. Quantifying a non notifiable disease in the united states: The national amyotrophic lateral sclerosis registry model. JAMA 2014. doi:10.1001/jama.2014.9799.

[14] Hanly J, Thompson K, Skedgel C. Identification of patients with systemic lupus erythematosus in administrative healthcare databases. Lupus 2014. doi:10.1177/0961203314543917.

[15] Muzaale AD, Massie AB, Wang M-C, Montgomery RA, McBride MA, Wainright JL, et al. Risk of end-stage renal disease following live kidney donation. JAMA 2014;311:579–86. doi:10.1001/jama.2013.285141.

[16] Chan EW, Liu KQ, Chui CS, Sing C, Wong LY, Wong IC. Adverse Drug Reactions – Examples of detection of rare events using databases. Br J Clin Pharmacol 2014

[17] Chu SY, Bachman DJ, Callaghan WM, Whitlock EP, Dietz PM, Berg CJ, et al. Association between Obesity during Pregnancy and Increased Use of Health Care. N Engl J Med 2008;358:1444–53. doi:10.1056/NEJMoa0706786.

[18] Kamel H, Navi BB, Sriram N, Hovsepian DA, Devereux RB, Elkind MSV. Risk of a Thrombotic Event after the 6-Week Postpartum Period. N Engl J Med 2014;370:1307–15. doi:10.1056/NEJMoa1311485.

[19] Bateman BT, Olbrecht VA, Berman MF, Minehart RD, Schwamm LH, Leffert LR. Peripartum subarachnoid hemorrhage: nationwide data and institutional experience. Anesthesiology 2012;116:324–33.
[20] Mølgaard-Nielsen D, Pasternak B, Hviid A. Use of Oral Fluconazole during Pregnancy and the Risk of Birth Defects. N Engl J Med 2013;369:830–9. doi:10.1056/NEJMoa1301066.
[21] Hunt LP, Ben-Shlomo Y, Clark EM, Dieppe P, Judge A, MacGregor AJ, et al. 45-day mortality after 467 779 knee replacements for osteoarthritis from the National Joint Registry for England and Wales: an observational study. The Lancet 2014.
[22] Curtis JP, Luebbert JJ, Wang Y, et al. Association of physician certification and outcomes among patients receiving an implantable cardioverter-defibrillator. JAMA 2009;301:1661–70. doi:10.1001/jama.2009.547.
[23] Redberg RF. Disparities in use of implantable cardioverter-defibrillators: Moving beyond process measures to outcomes data. JAMA 2007;298:1564–6. doi:10.1001/jama.298.13.1564.
[24] Slover J, Zuckerman JD. Increasing use of total knee replacement and revision surgery. JAMA 2012;308:1266–8.
[25] Odum SM, Springer BD. In-Hospital complication rates and associated factors after simultaneous bilateral versus unilateral total knee arthroplasty. J Bone Jt Surg 2014;96:1058–65.
[26] Sims DB, Naka Y, Jorde UP. Outcomes of Medicare beneficiaries with ventricular assist devices. JAMA 2009;301:1656–8.
[27] Griffin JW, Hadeed MM, Novicoff WM, Browne JA, Brockmeier SF. Patient age is a factor in early outcomes after shoulder arthroplasty. J Shoulder Elbow Surg 2014.
[28] Memtsoudis SG, Ma Y, Gonzalez Della Valle A, Mazumdar M, Gaber-Baylis LK, MacKenzie CR, et al. Perioperative outcomes after unilateral and bilateral total knee arthroplasty. Anesthesiology 2009;111:1206–16.
[29] Zmistowski B, Hozack WJ, Parvizi J. Readmission rates after total hip arthroplasty. JAMA 2011;306:825–6.
[30] Curtis LH, Al-Khatib SM, Shea AM, Hammill BG, Hernandez AF, Schulman KA. Sex differences in the use of implantable cardioverter-defibrillators for primary and secondary prevention of sudden cardiac death. JAMA 2007;298:1517–24. doi:10.1001/jama.298.13.1517.
[31] Dieterich JD, Fields AC, Moucha CS. Short Term Outcomes of Revision Total Knee Arthroplasty. J Arthroplasty 2014.
[32] Cram P, Lu X, Kates SL, Singh JA, Li Y, Wolf BR. Total knee arthroplasty volume, utilization, and outcomes among Medicare beneficiaries, 1991-2010. JAMA 2012;308:1227–36. doi:10.1001/2012.jama.11153.
[33] Canto JG, Rogers WJ, Goldberg RJ, Peterson ED, Wenger NK, Vaccarino V, et al. Association of age and sex with myocardial infarction symptom presentation and in-hospital mortality. JAMA 2012;307:813–22.
[34] Dharmarajan K, Hsieh AF, Lin Z, et al. Diagnoses and timing of 30-day readmissions after hospitalization for heart failure, acute myocardial infarction, or pneumonia. JAMA 2013;309:355–63. doi:10.1001/jama.2012.216476.
[35] Hernandez AF, Greiner MA, Fonarow GC, et al. Relationship between early physician follow-up and 30-day readmission among Medicare beneficiaries hospitalized for heart failure. JAMA 2010;303:1716–22. doi:10.1001/jama.2010.533.
[36] Fonarow GC, Adams KF, Abraham WT, Yancy CW, Boscardin W, ADHERE Scientific Advisory Committee, et al. Risk stratification for in-hospital mortality in

acutely decompensated heart failure: Classification and regression tree analysis. JAMA 2005;293:572–80. doi:10.1001/jama.293.5.572.
[37] Stensland J, Pettengill J, Winter A, Miller M. Specialty cardiac hospitals and coronary revascularization rates. JAMA 2007;297:2696–2696. doi:10.1001/jama.297.24.2696-a.
[38] Olesen JB, Lip GYH, Kamper A-L, Hommel K, Køber L, Lane DA, et al. Stroke and Bleeding in Atrial Fibrillation with Chronic Kidney Disease. N Engl J Med 2012;367:625–35. doi:10.1056/NEJMoa1105594.
[39] Washington CW, Derdeyn CP, Dacey Jr RG, Dhar R, Zipfel GJ. Analysis of subarachnoid hemorrhage using the Nationwide Inpatient Sample: the NIS-SAH Severity Score and Outcome Measure: Clinical article. J Neurosurg 2014;121:482–9.
[40] Claassen J, Bateman BT, Willey JZ, Inati S, Hirsch LJ, Mayer SA, et al. Generalized convulsive status epilepticus after nontraumatic subarachnoid hemorrhage: the nationwide inpatient sample. Neurosurgery 2007;61:60–5.
[41] Grossman R, Mukherjee D, Chang DC, Purtell M, Lim M, Brem H, et al. Predictors of inpatient death and complications among postoperative elderly patients with metastatic brain tumors. Ann Surg Oncol 2011;18:521–8.
[42] Ananthakrishnan AN, McGinley EL, Binion DG. Inflammatory bowel disease in the elderly is associated with worse outcomes: A national study of hospitalizations. Inflamm Bowel Dis 2009;15:182–9. doi:10.1002/ibd.20628.
[43] Wu H, Nguyen GC. Liver cirrhosis is associated with venous thromboembolism among hospitalized patients in a nationwide US study. Clin Gastroenterol Hepatol 2010;8:800–5.
[44] Kuy S, Dua A, Chappidi R, Seabrook G, Brown KR, Lewis B, et al. The increasing incidence of thromboembolic events among hospitalized patients with inflammatory bowel disease. Vascular 2014:1708538114541799.
[45] Tseng VL, Yu F, Lum F, Coleman AL. Risk of fractures following cataract surgery in medicare beneficiaries. JAMA 2012;308:493–501. doi:10.1001/jama.2012.9014.
[46] Baxter NN, Habermann EB, Tepper JE, Durham SB, Virnig BA. Risk of pelvic fractures in older women following pelvic irradiation. JAMA 2005;294:2587–93. doi:10.1001/jama.294.20.2587.
[47] Jencks SF, Williams MV, Coleman EA. Rehospitalizations among Patients in the Medicare Fee-for-Service Program. N Engl J Med 2009;360:1418–28. doi:10.1056/NEJMsa0803563.
[48] Kansagara D, Englander H, Salanitro A, et al. Risk prediction models for hospital readmission: A systematic review. JAMA 2011;306:1688–98. doi:10.1001/jama.2011.1515.
[49] Epstein AM, Jha AK, Orav EJ. The Relationship between Hospital Admission Rates and Rehospitalizations. N Engl J Med 2011;365:2287–95. doi:10.1056/NEJMsa1101942.
[50] Joynt KE, Orav E, Jha AK. Thirty-day readmission rates for Medicare beneficiaries by race and site of care. JAMA 2011;305:675–81. doi:10.1001/jama.2011.123.
[51] Ghaferi AA, Birkmeyer JD, Dimick JB. Variation in Hospital Mortality Associated with Inpatient Surgery. N Engl J Med 2009;361:1368–75.
[52] Tsai TC, Joynt KE, Orav EJ, Gawande AA, Jha AK. Variation in surgical-readmission rates and quality of hospital care. N Engl J Med 2013;369:1134–42. doi:10.1056/NEJMsa1303118.
[53] Williams CL, Bunch KJ, Stiller CA, Murphy MFG, Botting BJ, Wallace WH, et al. Cancer Risk among Children Born after Assisted Conception. N Engl J Med 2013;369:1819–27. doi:10.1056/NEJMoa1301675.

[54] Kulkarni AD, Jamieson DJ, Jones HW, Kissin DM, Gallo MF, Macaluso M, et al. Fertility Treatments and Multiple Births in the United States. N Engl J Med 2013;369:2218–25. doi:10.1056/NEJMoa1301467.
[55] Hviid A, Melbye M, Pasternak B. Use of Selective Serotonin Reuptake Inhibitors during Pregnancy and Risk of Autism. N Engl J Med 2013;369:2406–15. doi:10.1056/NEJMoa1301449.
[56] Griffin MR, Zhu Y, Moore MR, Whitney CG, Grijalva CG. U.S. Hospitalizations for Pneumonia after a Decade of Pneumococcal Vaccination. N Engl J Med 2013;369:155–63. doi:10.1056/NEJMoa1209165.
[57] Lapi F, Azoulay L, Niazi M, Yin H, Benayoun S, Suissa S. Androgen deprivation therapy and risk of acute kidney injury in patients with prostate cancer. JAMA 2013;310:289–96. doi:10.1001/jama.2013.8638.
[58] Pasternak B, Svanström H, Melbye M, Hviid A. Association between oral fluoroquinolone use and retinal detachment. JAMA 2013;310:2184–90. doi:10.1001/jama.2013.280500.
[59] Vigen R, O'Donnell CI, Barón AE, et al. Association of testosterone therapy with mortality, myocardial infarction, and stroke in men with low testosterone levels. JAMA 2013;310:1829–36. doi:10.1001/jama.2013.280386.
[60] Gandhi S, Fleet JL, Bailey DG, et al. Calcium-channel blocker–clarithromycin drug interactions and acute kidney injury. JAMA 2013;310:2544–53. doi:10.1001/jama.2013.282426.
[61] McWilliams J, Landon BE, Chernew ME. Changes in health care spending and quality for medicare beneficiaries associated with a commercial aco contract. JAMA 2013;310:829–36. doi:10.1001/jama.2013.276302.
[62] Silber JH, Rosenbaum PR, Clark AS, et al. Characteristics associated with differences in survival among black and white women with breast cancer. JAMA 2013;310:389–97. doi:10.1001/jama.2013.8272.
[63] Vinden C, Nash DM, Rangrej J, et al. Complications of daytime elective laparoscopic cholecystectomies performed by surgeons who operated the night before. JAMA 2013;310:1837–41. doi:10.1001/jama.2013.280372.
[64] Bilimoria KY, Chung J, Ju MH, et al. Evaluation of surveillance bias and the validity of the venous thromboembolism quality measure. JAMA 2013;310:1482–9. doi:10.1001/jama.2013.280048.
[65] Matlock DD, Groeneveld PW, Sidney S, et al. Geographic variation in cardiovascular procedure use among medicare fee-for-service vs medicare advantage beneficiaries. JAMA 2013;310:155–61. doi:10.1001/jama.2013.7837.
[66] Reed M, Huang J, Brand R, et al. Implementation of an outpatient electronic health record and emergency department visits, hospitalizations, and office visits among patients with diabetes. JAMA 2013;310:1060–5. doi:10.1001/jama.2013.276733.
[67] Mack MJ, Brennan J, Brindis R, et al. Outcomes following transcatheter aortic valve replacement in the united states. JAMA 2013;310:2069–77. doi:10.1001/jama.2013.282043.
[68] Peterson PN, Greiner MA, Qualls LG, et al. Qrs duration, bundle-branch block morphology, and outcomes among older patients with heart failure receiving cardiac resynchronization therapy. JAMA 2013;310:617–26. doi:10.1001/jama.2013.8641.
[69] Hawn MT, Graham LA, Richman JS, Itani KF, Henderson WG, Maddox TM. Risk of major adverse cardiac events following non cardiac surgery in patients with coronary stents. JAMA 2013;310:1462–72. doi:10.1001/jama.2013.278787.

[70] Warner DO, Berge K, Sun H, Harman A, Hanson A, Schroeder DR. Substance use disorder among anesthesiology residents, 1975-2009. JAMA 2013;310:2289–96. doi:10.1001/jama.2013.281954.
[71] Barreto-Filho J, Wang Y, Dodson JA, et al. Trends in aortic valve replacement for elderly patients in the united states, 1999-2011. JAMA 2013;310:2078–84. doi:10.1001/jama.2013.282437.
[72] Hunt LP, Ben-Shlomo Y, Clark EM, Dieppe P, Judge A, MacGregor AJ, et al. 90-day mortality after 409 096 total hip replacements for osteoarthritis, from the National Joint Registry for England and Wales: a retrospective analysis. The Lancet 2013;382:1097–104. doi:10.1016/S0140-6736(13)61749-3.
[73] Sitas F, Egger S, Bradshaw D, Groenewald P, Laubscher R, Kielkowski D, et al. Differences among the coloured, white, black, and other South African populations in smoking-attributed mortality at ages 35–74 years: a case-control study of 481 640 deaths. The Lancet 2013;382:685–93. doi:10.1016/S0140-6736(13)61610-4.
[74] Rasella D, Aquino R, Santos CA, Paes-Sousa R, Barreto ML. Effect of a conditional cash transfer programme on childhood mortality: a nationwide analysis of Brazilian municipalities. The Lancet 2013;382:57–64. doi:10.1016/S0140-6736(13)60715-1.
[75] Fazel S, Wolf A, Långström N, Newton CR, Lichtenstein P. Premature mortality in epilepsy and the role of psychiatric comorbidity: a total population study. The Lancet 2013;382:1646–54. doi:10.1016/S0140-6736(13)60899-5.
[76] Massie AB, Kuricka LM, Segev DL. Big Data in Organ Transplantation: Registries and Administrative Claims. Am J Transplant 2014;14:1723–30. doi:10.1111/ajt.12777.
[77] Roder D m., Fong K m., Brown M p., Zalcberg J, Wainwright C e. Realising opportunities for evidence-based cancer service delivery and research: linking cancer registry and administrative data in Australia. Eur J Cancer Care (Engl) 2014
[78] Nigwekar SU, Solid CA, Ankers E, Malhotra R, Eggert W, Turchin A, et al. Quantifying a Rare Disease in Administrative Data: The Example of Calciphylaxis. J Gen Intern Med 2014;29:724–31. doi:10.1007/s11606-014-2910-1.
[79] Royer JA, Hardin JW, McDermott S, Ouyang L, Mann JR, Ozturk OD, et al. Use of State Administrative Data Sources to Study Adolescents and Young Adults with Rare Conditions. J Gen Intern Med 2014;29:732–8. doi:10.1007/s11606-014-2925-7.
[80] WHO | International Classification of Diseases (ICD). WHO http://www.who.int/classifications/icd/en/ (accessed October 21, 2014).
[81] Données ouvertes en France. Wikipédia 2015.
[82] Nisen P, Rockhold F. Access to Patient-Level Data from GlaxoSmithKline Clinical Trials. N Engl J Med 2013;369:475–8. doi:10.1056/NEJMsr1302541.
[83] Strom BL, Buyse M, Hughes J, Knoppers BM. Data Sharing, Year 1 — Access to Data from Industry-Sponsored Clinical Trials. N Engl J Med 2014;371:2052–4. doi:10.1056/NEJMp1411794.
[84] Arrêté du 11 juillet 2012 relatif à la mise en œuvre du système national d'information interrégimes de l'assurance maladie.
[85] Rapport commission open data 2014.
[86] Loi n° 2002-303 du 4 mars 2002 relative aux droits des malades et à la qualité du système de santé.
[87] Hartmann HA, Anhøj J, Hellebek A, Egebart J, Bjørn B, Lilja B. Computerised Physician Order Entry (CPOE). Stud Health Technol Inform 2008;148:159–62.
[88] Loi n° 91-748 du 31 juillet 1991 portant réforme hospitalière.
[89] Chazard E, Merlin B, Ficheur G, Sarfati JC, Beuscart R. Detection of adverse drug events: proposal of a data model. Stud Health Technol Inf 2009;148:63–74.

[90] Chazard E, Mouret C, Ficheur G, Schaffar A, Beuscart J-B, Beuscart R. Proposal and evaluation of FASDIM, a Fast And Simple De-Identification Method for unstructured free-text clinical records. Int J Med Inf 2014;83:303–12. doi:10.1016/j.ijmedinf.2013.11.005.
[91] Beeler GW. HL7 Version 3—An object-oriented methodology for collaborative standards development. Int J Med Inf 1998;48:151–61.
[92] Muñoz P, Trigo JD, Martínez I, Muñoz A, Escayola J, García J. The ISO/EN 13606 standard for the interoperable exchange of electronic health records. J Healthc Eng 2011;2:1–24.
[93] Schloeffel P, Beale T, Hayworth G, Heard S, Leslie H. The relationship between CEN 13606, HL7, and openEHR. HIC 2006 HINZ 2006 Proc 2006:24.
[94] Reisinger SJ, Ryan PB, O'Hara DJ, Powell GE, Painter JL, Pattishall EN, et al. Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases. J Am Med Inform Assoc JAMIA 2010;17:652–62. doi:10.1136/jamia.2009.002477.
[95] Dan B. Challenges in using hospital billing databases for epidemiology. Dev Med Child Neurol 2014.
[96] EMOIS Nancy 2011 - Données PMSI chaînées : attention un patient peut en cacher un autre !
[97] EMOIS Nancy 2011 - Taux de mortalité hospitalier : les données du PMSI sont-elles utilisables ?
[98] Dhalwani NN, Tata LJ, Coleman T, Fiaschi L, Szatkowski L. A comparison of UK primary care data with other national data sources for monitoring the prevalence of smoking during pregnancy. J Public Health 2014:fdu060. doi:10.1093/pubmed/fdu060.
[99] Larsen TB, Johnsen SP, Møller CI, Larsen H, Sørensen HT. A review of medical records and discharge summary data found moderate to high predictive values of discharge diagnoses of venous thromboembolism during pregnancy and postpartum. J Clin Epidemiol 2005;58:316–9. doi:10.1016/j.jclinepi.2004.07.004.
[100] Enomoto LM, Hollenbeak CS, Bhayani NH, Dillon PW, Gusani NJ. Measuring surgical quality: a national clinical registry versus administrative claims data. J Gastrointest Surg 2014;18:1416–22. doi:10.1007/s11605-014-2569-2.
[101] Quantin C, Cottenet J, Vuagnat A, Prunet C, Mouquet M-C, Fresson J, et al. Qualité des données périnatales issues du PMSI : comparaison avec l'état civil et l'enquête nationale périnatale 2010. J Gynécologie Obstétrique Biol Reprod.
[102] Espehaug B, Furnes O, Havelin LI, Engesæter LB, Vollset SE, Kindseth O. Registration completeness in the Norwegian Arthroplasty Register. Acta Orthop 2006;77:49–56. doi:10.1080/17453670610045696.
[103] Pedersen A, Johnsen S, Overgaard S, Søballe K, Sørensen H, Lucht U. Registration in the Danish Hip Arthroplasty Registry Completeness of total hip arthroplasties and positive predictive value of registered diagnosis and postoperative complications. Acta Orthop 2004;75:434–41. doi:10.1080/00016470410001213-1.
[104] Penberthy L, Petkov V, McClish D, Peace S, Overton S, Radhakrishnan S, et al. The value of billing data from oncology practice to supplement treatment information for cancer surveillance. J Regist Manag 2014;41:57–64.
[105] Arthursson AJ, Furnes O, Espehaug B, Havelin LI, Söreide JA. Validation of data in the Norwegian Arthroplasty Register and the Norwegian Patient Register : 5,134 primary total hip arthroplasties and revisions operated at a single hospital between 1987 and 2003. Acta Orthop 2005;76:823–8. doi:10.1080/17453670510045435.

[106] Marder E, Garman K, Jones TF, Dunn J, Jones S. Assessment of administrative claims data for public health reporting of Salmonella in Tennessee. J Am Med Inform Assoc 2014:amiajnl – 2014–002909. doi:10.1136/amiajnl-2014-002909.
[107] Viboud C, Charu V, Olson D, Ballesteros S, Gog J, Khan F, et al. Demonstrating the Use of High-Volume Electronic Medical Claims Data to Monitor Local and Regional Influenza Activity in the US. PLoS ONE 2014;9:e102429. doi:10.1371/journal.pone.0102429.
[108] Bekkers S, Bot AGJ, Makarawung D, Neuhaus V, Ring D. The National Hospital Discharge Survey and Nationwide Inpatient Sample: The Databases Used Affect Results in THA Research. Clin Orthop Relat Res 2014;472:3441–9. doi:10.1007/s11999-014-3836-y.
[109] Reich CG, Ryan PB, Schuemie MJ. Alternative outcome definitions and their effect on the performance of methods for observational outcome studies. Drug Saf 2013;36:181–93. doi:10.1007/s40264-013-0111-1.
[110] Funk MJ, Landi SN. Misclassification in Administrative Claims Data: Quantifying the Impact on Treatment Effect Estimates. Curr Epidemiol Rep 2014;1:175–85. doi:10.1007/s40471-014-0027-z.
[111] Casez P, Labarère J, Sevestre M-A, Haddouche M, Courtois X, Mercier S, et al. ICD-10 hospital discharge diagnosis codes were sensitive for identifying pulmonary embolism but not deep vein thrombosis. J Clin Epidemiol 2010;63:790–7. doi:10.1016/j.jclinepi.2009.09.002.
[112] Andrade SE, Harrold LR, Tjia J, Cutrona SL, Saczynski JS, Dodd KS, et al. A systematic review of validated methods for identifying cerebrovascular accident or transient ischemic attack using administrative data. Pharmacoepidemiol Drug Saf 2012;21:100–28. doi:10.1002/pds.2312.
[113] Tamariz L, Harkins T, Nair V. A systematic review of validated methods for identifying venous thromboembolism using administrative and claims data. Pharmacoepidemiol Drug Saf 2012;21:154–62. doi:10.1002/pds.2341.
[114] Tagalakis V, Kahn SR. Determining the test characteristics of claims-based diagnostic codes for the diagnosis of venous thromboembolism in a medical service claims database. Pharmacoepidemiol Drug Saf 2011;20:304–7. doi:10.1002/pds.2061.
[115] Tirschwell DL, Longstreth WT. Validating Administrative Data in Stroke Research. Stroke 2002;33:2465–70. doi:10.1161/01.STR.0000032240.28636.BD.
[116] Sundararajan V, Henderson T, Perry C, Muggivan A, Quan H, Ghali WA. New ICD-10 version of the Charlson comorbidity index predicted in-hospital mortality. J Clin Epidemiol 2004;57:1288–94. doi:10.1016/j.jclinepi.2004.03.012.
[117] Quan H, Eastwood C, Cunningham CT, Liu M, Flemons W, Coster CD, et al. Validity of AHRQ patient safety indicators derived from ICD-10 hospital discharge abstract data (chart review study). BMJ Open 2013;3:e003716. doi:10.1136/bmjopen-2013-003716.
[118] Hippisley-Cox J, Coupland C. Predicting risk of upper gastrointestinal bleed and intracranial bleed with anticoagulants: cohort study to derive and validate the QBleed scores. BMJ 2014;349:g4606–g4606. doi:10.1136/bmj.g4606.
[119] Friberg L, Rosenqvist M, Lip GYH. Evaluation of risk stratification schemes for ischaemic stroke and bleeding in 182 678 patients with atrial fibrillation: the Swedish Atrial Fibrillation cohort study. Eur Heart J 2012;33:1500–10. doi:10.1093/eurheartj/ehr488.
[120] Hippisley-Cox J, Coupland C. Development and validation of risk prediction algorithm (QThrombosis) to estimate future risk of venous thromboembolism: prospective cohort study. BMJ 2011;343:d4656–d4656. doi:10.1136/bmj.d4656.

[121] Hartzema AG, Reich CG, Ryan PB, Stang PE, Madigan D, Welebob E, et al. Managing Data Quality for a Drug Safety Surveillance System. Drug Saf 2013;36:49–58. doi:10.1007/s40264-013-0098-7.
[122] Arrêté du 19 juillet 2013 relatif à la mise en œuvre du Système national d'information interrégimes de l'assurance maladie.
[123] Rapport au Parlement Institut des données de santé - 2014.
[124] Kohn LT, Corrigan JM, Donaldson MS. To Err Is Human:: Building a Safer Health System. vol. 627. National Academies Press; 2000.
[125] Lazarou J, Pomeranz BH, Corey PN. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. JAMA 1998;279:1200–5.
[126] Juntti-Patinen L, Neuvonen P. Drug-related deaths in a university central hospital. Eur J Clin Pharmacol 2002;58:479–82. doi:10.1007/s00228-002-0501-2.
[127] Michel P, Lathelize M, Bru-Sonnet R, Domecq S, Kret M, Quenon JL. Enquête Nationale sur les Evénements Indésirables graves liés aux Soins 2009 (ENEIS2): description des résultats 2009. Rapport final à la DREES (Ministère du travail, de l'emploi et de la Santé)–Février 2011, Bordeaux.
[128] International drug monitoring. The role of the hospital. World Health Organ Tech Rep Ser 1969;425:5–24.
[129] Bates DW, Cullen DJ, Laird N, Petersen LA, Small SD, Servi D, et al. Incidence of adverse drug events and potential adverse drug events: implications for prevention. JAMA 1995;274:29–34.
[130] Reporting Serious Problems to FDA - What is a Serious Adverse Event? 2013. http://www.fda.gov/safety/medwatch/howtoreport/ucm053087.htm (accessed December 27, 2013).
[131] Directive 2007/47/CE du Parlement européen et du Conseil modifiant la directive 90/385/CEE du Conseil concernant le rapprochement des législations des États membres relatives aux dispositifs médicaux implantables actifs, la directive 93/42/CEE du Conseil relative aux dispositifs médicaux et la directive 98/8/CE concernant la mise sur le marché des produits biocides (Texte présentant de l'intérêt pour l'EEE).
[132] Arrêté du 15 mars 2010 fixant les conditions de mise en œuvre des exigences essentielles applicables aux dispositifs médicaux, pris en application de l'article R. 5211-24 du code de la santé publique.
[133] Code de la santé publique - Article L5212-1. vol. L5212-1.
[134] Naranjo CA, Busto U, Sellers EM, Sandor P, Ruiz I, Roberts EA, et al. A method for estimating the probability of adverse drug reactions. Clin Pharmacol Ther 1981;30:239–45.
[135] Kramer MS, Leventhal JM, Hutchinson TA, Feinstein AR. An Algorithm for the Operational Assessment of Adverse Drug Reactions: I. Background, Description, and Instructions for Use. JAMA J Am Med Assoc 1979;242:623–32. doi:10.1001/jama.1979.03300070019017.
[136] Bégaud B, Evreux JC, Jouglard J, Lagier G. [Imputation of the unexpected or toxic effects of drugs. Actualization of the method used in France]. Thérapie 1985;40:111–8.
[137] Alvarez-Requejo A, Carvajal A, Bégaud B, Moride Y, Vega T, Arias LHM. Under-reporting of adverse drug reactions Estimate based on a spontaneous reporting scheme and a sentinel system. Eur J Clin Pharmacol 1998;54:483–8. doi:10.1007/s002280050498.
[138] Hazell L, Shakir SAW. Under-Reporting of Adverse Drug Reactions: A Systematic Review. Drug Saf 2006;29:385–96.
[139] Beuscart R, McNair P, Brender J, others. Patient safety through intelligent procedures in medication: the PSIP project. Stud Health Technol Inf 2009;148:6–13.

[140] Chazard E, Ficheur G, Bernonville S, Luyckx M, Beuscart R. Data mining to generate adverse drug events detection rules. IEEE Trans Inf Technol Biomed Publ IEEE Eng Med Biol Soc 2011;15:823–30. doi:10.1109/TITB.2011.2165727.
[141] Chazard E, Preda C, Merlin B, Ficheur G, Beuscart R. Data-mining-based detection of adverse drug events. Stud Health Technol Inf 2009;150:552–6.
[142] Chazard E, Băceanu A, Ferret L, Ficheur G. The ADE scorecards: a tool for adverse drug event detection in electronic health records. Stud Health Technol Inform 2011;166:169–79.
[143] Aronson JK, Hauben M. Anecdotes that provide definitive evidence. BMJ 2006;333:1267–9. doi:10.1136/bmj.39036.666389.94.
[144] Murray MD. Use of Data from Electronic Health Records for Pharmacoepidemiology. Curr Epidemiol Rep 2014;1:186–93. doi:10.1007/s40471-014-0020-6.
[145] Carnahan RM. Mini-Sentinel's systematic reviews of validated methods for identifying health outcomes using administrative data: summary of findings and suggestions for future research. Pharmacoepidemiol Drug Saf 2012;21:90–9. doi:10.1002/pds.2318.
[146] Platt R, Carnahan R. The US Food and Drug Administration's Mini-Sentinel Program. Pharmacoepidemiol Drug Saf 2012;21:1–303.
[147] CNAM. Risque de cancer de la vessie chez les personnes diabétiques traitées par pioglitazone en France : une étude de cohorte sur les données du SNIIRAM et du PMSI 2011.
[148] Neumann A, Weill A, Ricordeau P, Fagot JP, Alla F, Allemand H. Pioglitazone and risk of bladder cancer among diabetic patients in France: a population-based cohort study. Diabetologia 2012;55:1953–62.
[149] Glasziou P, Meats E, Heneghan C, Shepperd S. What is missing from descriptions of treatment in trials and reviews? BMJ 2008;336:1472–4. doi:10.1136/bmj.39590.732037.47.
[150] McCulloch P, Taylor I, Sasako M, Lovett B, Griffin D. Randomised trials in surgery: problems and possible solutions. BMJ 2002;324:1448–51. doi:10.1136/bmj.324.7351.1448.
[151] Boutron I, Moher D, Altman DG, Schulz KF, Ravaud P. Extending the CONSORT Statement to Randomized Trials of Nonpharmacologic Treatment: Explanation and Elaboration. Ann Intern Med 2008;148:295–309. doi:10.7326/0003-4819-148-4-200802190-00008.
[152] Lilford RJ, Braunholtz DA, Greenhalgh R, Edwards SJL. Trials and fast changing technologies: the case for tracker studies. BMJ 2000;320:43–6.
[153] MCI portant sur les dispositifs médicaux implantables : compte rendu de la semaine du 26/03/12 http://www.senat.fr/compte-rendu-commissions/20120326/mci_implants.html (accessed November 10, 2014).
[154] Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. Ann Intern Med 2010;153:600–6.
[155] Ryan PB, Schuemie MJ. Evaluating Performance of Risk Identification Methods Through a Large-Scale Simulation of Observational Data. Drug Saf 2013;36:171–80. doi:10.1007/s40264-013-0110-2.
[156] Ryan PB, Schuemie MJ, Welebob E, Duke J, Valentine S, Hartzema AG. Defining a Reference Set to Support Methodological Research in. Drug Saf 2013;36:33–47. doi:10.1007/s40264-013-0097-8.
[157] Schuemie MJ, Gini R, Coloma PM, Straatman H, Herings RMC, Pedersen L, et al. Replication of the OMOP experiment in Europe: evaluating methods for risk

identification in electronic health record databases. Drug Saf Int J Med Toxicol Drug Exp 2013;36 Suppl 1:S159–69. doi:10.1007/s40264-013-0109-8.
[158] 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 Int J Med Toxicol Drug Exp 2013;36 Suppl 1:S143–58. doi:10.1007/s40264-013-0108-9.
[159] Ryan PB, Schuemie MJ, Madigan D. Empirical Performance of a Self-Controlled Cohort Method: Lessons for Developing a Risk Identification and Analysis System. Drug Saf 2013;36:95–106. doi:10.1007/s40264-013-0101-3.
[160] Suchard MA, Zorych I, Simpson SE, Schuemie MJ, Ryan PB, Madigan D. Empirical Performance of the Self-Controlled Case Series Design: Lessons for Developing a Risk Identification and Analysis System. Drug Saf 2013;36:83–93. doi:10.1007/s40264-013-0100-4.
[161] Ryan PB, Schuemie MJ, Gruber S, Zorych I, Madigan D. Empirical Performance of a New User Cohort Method: Lessons for Developing a Risk Identification and Analysis System. Drug Saf 2013;36:59–72. doi:10.1007/s40264-013-0099-6.
[162] Madigan D, Schuemie MJ, Ryan PB. Empirical Performance of the Case–Control Method: Lessons for Developing a Risk Identification and Analysis System. Drug Saf 2013;36:73–82. doi:10.1007/s40264-013-0105-z.
[163] DuMouchel W, Ryan PB, Schuemie MJ, Madigan D. Evaluation of Disproportionality Safety Signaling Applied to Healthcare Databases. Drug Saf 2013;36:123–32. doi:10.1007/s40264-013-0106-y.
[164] Schuemie MJ, Madigan D, Ryan PB. Empirical Performance of LGPS and LEOPARD: Lessons for Developing a Risk Identification and Analysis System. Drug Saf 2013;36:133–42. doi:10.1007/s40264-013-0107-x.
[165] Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for reporting observational studies. Prev Med 2007;45:247–51. doi:10.1016/j.ypmed.2007.08.012.
[166] Lévesque LE, Hanley JA, Kezouh A, Suissa S. Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. Bmj 2010;340.
[167] Suissa S. Immortal time bias in observational studies of drug effects. Pharmacoepidemiol Drug Saf 2007;16:241–9.
[168] Suissa S. Immortal time bias in pharmacoepidemiology. Am J Epidemiol 2008;167:492–9.
[169] Matok I, Azoulay L, Yin H, Suissa S. Immortal time bias in observational studies of drug effects in pregnancy. Birt Defects Res A Clin Mol Teratol 2014;100:658–62. doi:10.1002/bdra.23271.
[170] Koutkias V, Jaulent M-C. An Agent-based Approach for Integrated Pharmacovigilance Signal Detection.
[171] Bate A, Lindquist M, Edwards IR, Olsson S, Orre R, Lansner A, et al. A Bayesian neural network method for adverse drug reaction signal generation. Eur J Clin Pharmacol 1998;54:315–21.
[172] Dumouchel W. Bayesian Data Mining in Large Frequency Tables, with an Application to the FDA Spontaneous Reporting System. Am Stat 1999;53:177–90. doi:10.1080/00031305.1999.10474456.
[173] McMahon AD, Evans JM, McGilchrist MM, McDevitt DG, Macdonald TM. Drug exposure risk windows and unexposed comparator groups for cohort studies in pharmacoepidemiology. Pharmacoepidemiol Drug Saf 1998;7:275–80.

[174] Zhang Z, Ni H, Xu X. Observational studies using propensity score analysis underestimated the effect sizes in critical care medicine. J Clin Epidemiol 2014. doi:10.1016/j.jclinepi.2014.02.018.
[175] Kitsios GD. Propensity score studies are unlikely to underestimate treatment effects in critical care medicine: a critical reanalysis. J Clin Epidemiol 2015;68:467–9. doi:10.1016/j.jclinepi.2014.10.012.
[176] 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. Epidemiol Camb Mass 2009;20:512–22. doi:10.1097/EDE.0b013e3181a663cc.
[177] Brookhart M, Sturmer T, Glynn R, Rassen J, Schneeweiss S. Confounding control in healthcare database research: challenges and potential approaches. Med Care 2010;48:S114–20. doi:10.1097/MLR.0b013e3181dbebe3.
[178] Genkin A, Lewis DD, Madigan D. Large-scale Bayesian logistic regression for text categorization. Technometrics 2007;49:291–304.
[179] Ryan PB, Madigan D, Stang PE, Marc Overhage J, 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.
[180] Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis 1987;40:373–83.
[181] Boogaarts HD, Conde MPD, Janssen E, Nuenen WFM van, Vries J de, Donders R, et al. The value of the Charlson Co-morbidity Index in aneurysmal subarachnoid haemorrhage. Acta Neurochir (Wien) 2014:1–5. doi:10.1007/s00701-014-2160-3.
[182] Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. J Clin Epidemiol 1992;45:613–9.
[183] Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. Med Care 1998;36:8–27.
[184] Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi JC, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Med Care 2005;43:1130–9. doi:00005650-200511000-00010.
[185] Schuemie MJ. Methods for drug safety signal detection in longitudinal observational databases: LGPS and LEOPARD. Pharmacoepidemiol Drug Saf 2011;20:292–9. doi:10.1002/pds.2051.
[186] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Methodol 1996:267–88.
[187] Kulldorff M. A spatial scan statistic. Commun Stat-Theory Methods 1997;26:1481–96.
[188] Kulldorff M, Fang Z, Walsh SJ. A Tree-Based Scan Statistic for Database Disease Surveillance. Biometrics 2003;59:323–31.
[189] Brown JS, Petronis KR, Bate A, Zhang F, Dashevsky I, Kulldorff M, et al. Drug Adverse Event Detection in Health Plan Data Using the Gamma Poisson Shrinker and Comparison to the Tree-based Scan Statistic. Pharmaceutics 2013;5:179–200. doi:10.3390/pharmaceutics5010179.
[190] Kulldorff M, Davis RL, Kolczak† M, Lewis E, Lieu T, Platt R. A Maximized Sequential Probability Ratio Test for Drug and Vaccine Safety Surveillance. Seq Anal 2011;30:58–78. doi:10.1080/07474946.2011.539924.
[191] Bland JM, Altman DG. Statistics Notes: Some examples of regression towards the mean. BMJ 1994;309:780. doi:10.1136/bmj.309.6957.780.

[192] 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.
[193] Farrington CP. Relative incidence estimation from case series for vaccine safety evaluation. Biometrics 1995:228–35.
[194] Dalgaard P. Introductory statistics with R. Springer; 2008.
[195] Therneau TM, Lumley T. survival: Survival Analysis. 2014.
[196] Elff M. mclogit: Mixed Conditional Logit. 2014.
[197] 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. doi:10.1111/biom.12078.
[198] Chazard E, Ficheur G, Merlin B, Serrot E, Beuscart R. Adverse drug events prevention rules: multi-site evaluation of rules from various sources. Stud Health Technol Inf 2009;148:102–11.
[199] Chazard E, Ficheur G, Merlin B, Genin M, Preda C, Beuscart R. Detection of adverse drug events detection: data aggregation and data mining. Stud Health Technol Inf 2009;148:75–84.
[200] Kass GV. An exploratory technique for investigating large quantities of categorical data. Appl Stat 1980:119–27.
[201] Breiman L, Friedman JH, Olshen R, Stone CJ. Classification and Regression Trees 1984.
[202] Wilkinson L. Tree Structured Data Analysis: AID, CHAID and CART 1992.
[203] Van Diepen M, Franses PH. Evaluating chi-squared automatic interaction detection. Inf Syst 2006;31:814–31.
[204] Breiman L. Heuristics of instability and stabilization in model selection. Ann Stat 1996.
[205] Breiman L. Random forests. Mach Learn 2001;45:5–32.
[206] CART and Random Forest Models. 2013.
[207] Piatetsky-Shapiro G, Frawley W. Knowledge discovery in databases. Menlo Park, Calif.: AAAI Press : MIT Press; 1991.
[208] Zaki MJ. Generating non-redundant association rules. Proc. Sixth ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., ACM; 2000, p. 34–43.
[209] Zaki MJ. SPADE: An Efficient Algorithm for Mining Frequent Sequences. Mach Learn 2001;42:31–60. doi:10.1023/A:1007652502315.
[210] Ficheur G, Chazard E, Merlin B, Ferret L, Luyckx M, Beuscart R. Supervised analysis of drug prescription sequences. Stud Health Technol Inform 2012;192:293–7.
[211] Ficheur G, Chazard E, Beuscart J-B, Merlin B, Luyckx M, Beuscart R. Adverse drug events with hyperkalaemia during inpatient stays: evaluation of an automated method for retrospective detection in hospital databases. BMC Med Inform Decis Mak 2014;14:83.
[212] Jha AK, DesRoches CM, Campbell EG, Donelan K, Rao SR, Ferris TG, et al. Use of Electronic Health Records in U.S. Hospitals. N Engl J Med 2009;360:1628–38. doi:10.1056/NEJMsa0900592.
[213] Handler SM, Altman RL, Perera S, Hanlon JT, Studenski SA, Bost JE, et al. A systematic review of the performance characteristics of clinical event monitor signals used to detect adverse drug events in the hospital setting. J Am Med Inf Assoc 2007;14:451–8. doi:M2369 - 10.1197/jamia.M2369.
[214] Schedlbauer A, Prasad V, Mulvaney C, Phansalkar S, Stanton W, Bates DW, et al. What evidence supports the use of computerized alerts and prompts to improve clinicians' prescribing behavior? J Am Med Inf Assoc 2009;16:531–8. doi:M2910 - 10.1197/jamia.M2910.

[215] Dormann H, Criegee-Rieck M, Neubert A, Egger T, Levy M, Hahn EG, et al. Implementation of a computer-assisted monitoring system for the detection of adverse drug reactions in gastroenterology. Aliment Pharmacol Ther 2004;19:303–9. doi:10.1111/j.1365-2036.2004.01854.x.
[216] Ponce SP, Jennings AE, Madias NE, Harrington JT. Drug-induced hyperkalemia. Medicine (Baltimore) 1985;64:357–70.
[217] 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. doi:10.1038/clpt.2012.50.
[218] Brown JS, Kulldorff M, Chan KA, Davis RL, Graham D, Pettus PT, et al. Early detection of adverse drug events within population-based health networks: application of sequential testing methods. Pharmacoepidemiol Drug Saf 2007;16:1275–84. doi:10.1002/pds.1509.
[219] Berlowitz DR, Miller DR, Oliveria SA, Cunningham F, Gomez-Caminero A, Rothendler JA. Differential associations of beta-blockers with hemorrhagic events for chronic heart failure patients on warfarin. Pharmacoepidemiol Drug Saf 2006;15:799–807. doi:10.1002/pds.1301.
[220] Schildcrout JS, Haneuse S, Peterson JF, Denny JC, Matheny ME, Waitman LR, et al. Analyses of longitudinal, hospital clinical laboratory data with application to blood glucose concentrations. Stat Med 2011;30:3208–20. doi:10.1002/sim.4352.
[221] Bates DW, Evans RS, Murff H, Stetson PD, Pizziferri L, Hripcsak G. Detecting Adverse Events Using Information Technology. J Am Med Inform Assoc 2003;10:115–28. doi:10.1197/jamia.M1074.
[222] Brown S, Black K, Mrochek S, Wood A, Bess T, Cobb J, et al. RADARx: Recognizing, Assessing, and Documenting Adverse Rx events. Proc AMIA Symp 2000:101–5.
[223] Raschke RA, Gollihare B, Wunderlich TA, Guidry JR, Leibowitz AI, Peirce JC, et al. A Computer Alert System to Prevent Injury From Adverse Drug Events. JAMA J Am Med Assoc 1998;280:1317–20. doi:10.1001/jama.280.15.1317.
[224] Aarts J, Van Der Sijs H. CPOE, alerts and workflow: taking stock of ten years research at Erasmus MC. Stud Health Technol Inform 2008;148:165–9.
[225] Beuscart-Zephir M-C. Contribution of human factors for the review of automatically detected ADE. Detect Prev Adverse Drug Events Inf Technol Hum Factors 2009;148:170.
[226] Kuperman GJ, Bobb A, Payne TH, Avery AJ, Gandhi TK, Burns G, et al. Medication-related clinical decision support in computerized provider order entry systems: a review. J Am Med Inform Assoc 2007;14:29–40.
[227] Gandhi TK, Weingart SN, Seger AC, Borus J, Burdick E, Poon EG, et al. Outpatient prescribing errors and the impact of computerized prescribing. J Gen Intern Med 2005;20:837–41. doi:JGI05414 - 10.1111/j.1525-1497.2005.0194.x.
[228] Bates DW, O'Neil AC, Boyle D, Teich J, Chertow GM, Komaroff AL, et al. Potential identifiability and preventability of adverse events using information systems. J Am Med Inf Assoc 1994;1:404–11.
[229] Kuperman GJ, Bates DW, Teich JM, Schneider JR, Cheiman D. A new knowledge structure for drug-drug interactions. Proc Annu Symp Comput Appl Med Care 1994:836–40.
[230] Jha AK, Kuperman GJ, Teich JM, Leape L, Shea B, Rittenberg E, et al. Identifying adverse drug events: development of a computer-based monitor and comparison with chart review and stimulated voluntary report. J Am Med Inf Assoc 1998;5:305–14.

[231] Field TS, Gurwitz JH, Harrold LR, Rothschild JM, Debellis K, Seger AC, et al. Strategies for detecting adverse drug events among older persons in the ambulatory setting. J Am Med Inf Assoc 2004;11:492–8. doi:10.1197/jamia.M1586.
[232] Agbabiaka TB, Savović J, Ernst E. Methods for causality assessment of adverse drug reactions: a systematic review. Drug Saf Int J Med Toxicol Drug Exp 2008;31:21–37.
[233] Yoon D, Park MY, Choi NK, Park BJ, Kim JH, Park RW. Detection of Adverse Drug Reaction Signals Using an Electronic Health Records Database: Comparison of the Laboratory Extreme Abnormality Ratio (CLEAR) Algorithm. Clin Pharmacol Ther 2012;91:467–74. doi:10.1038/clpt.2011.248.
[234] Koutkias V, Kilintzis V, Stalidis G, Lazou K, Collyda C, Chazard E, et al. Constructing Clinical Decision Support Systems for Adverse Drug Event Prevention: A Knowledge-based Approach.
[235] Jha AK, Laguette J, Seger A, Bates DW. Can Surveillance Systems Identify and Avert Adverse Drug Events? A Prospective Evaluation of a Commercial Application. J Am Med Inform Assoc 2008;15:647–53. doi:10.1197/jamia.M2634.
[236] Geerts WH, Bergqvist D, Pineo GF, Heit JA, Samama CM, Lassen MR, et al. Prevention of venous thromboembolism: American college of chest physicians evidence-based clinical practice guidelines (8th edition). Chest 2008;133:381S – 453S. doi:10.1378/chest.08-0656.
[237] Guyatt GH, Akl EA, Crowther M, Gutterman DD, Schuünemann HJ. Executive summary: Antithrombotic therapy and prevention of thrombosis, 9th ed: american college of chest physicians evidence-based clinical practice guidelines. Chest 2012;141:7S – 47S. doi:10.1378/chest.1412S3.
[238] Smith AJ, Dieppe P, Vernon K, Porter M, Blom AW. Failure rates of stemmed metal-on-metal hip replacements: analysis of data from the National Joint Registry of England and Wales. The Lancet 2012;379:1199–204. doi:10.1016/S0140-6736(12)60353-5.
[239] Kurtz SM, Ong KL, Lau E, Widmer M, Maravic M, Gómez-Barrena E, et al. International survey of primary and revision total knee replacement. Int Orthop 2011;35:1783–9. doi:10.1007/s00264-011-1235-5.
[240] Boscoe FP, McLaughlin C, Schymura MJ, Kielb CL. Visualization of the spatial scan statistic using nested circles. Health Place 2003;9:273–7.
[241] Carroll LN, Au AP, Detwiler LT, Fu T, Painter IS, Abernethy NF. Visualization and analytics tools for infectious disease epidemiology: A systematic review. J Biomed Inform 2014;51:287–98. doi:10.1016/j.jbi.2014.04.006.
[242] Mark T, Pepitone A, Hatzmann M, Navathe A, Goodrich K, Chang S. CO3 Project LIBRA: A new analytic tool for comparative effectiveness analyses of multipayer claims databases. Value Health 2011;14:A2. doi:10.1016/j.jval.2011.02.012.
[243] Rising J, Moscovitch B. The Food and Drug Administration's Unique Device Identification System: Better Postmarket Data on the Safety and Effectiveness of Medical Devices. JAMA Intern Med 2014;174:1719–20. doi:10.1001/jamainternmed.2014.4195.
[244] RStudio, Inc. shiny: Web Application Framework for R. 2014.
[245] Riccioli C, Leroy N, Pelayo S. The PSIP approach to account for human factors in Adverse Drug Events: Preliminary field studies. Stud Health Technol Inf 2009;148:197–205.
[246] Beuscart-Zéphir M-C, Bernonville S, Leroy N, Marcilly R, Pelayo S. Final recommendations and specifications for the design of the PSIP CDSS modules 2009.

[247] Beuscart R, others. Human factors engineering for computer-supported identification and prevention of adverse drug events. Detect Prev Adverse Drug Events Inf Technol Hum Factors 2009;148:14.
[248] Bates DW. Using information technology to reduce rates of medication errors in hospitals. BMJ 2000;320:788–91. doi:10.1136/bmj.320.7237.788.
[249] Bates DW, Leape LL, Cullen DJ, Laird N, Petersen LA, Teich JM, et al. Effect of computerized physician order entry and a team intervention on prevention of serious medication errors. JAMA 1998;280:1311–6. doi:joc80319.
[250] Bates DW. Measuring Patient Safety: the Need for Prospective Detection of Adverse Events. Stud Health Tech Inf 2009;148:3.
[251] Leal J, Laupland KB. Validity of electronic surveillance systems: a systematic review. J Hosp Infect 2008;69:220–9.
[252] Waller PC, Evans SJ. A model for the future conduct of pharmacovigilance. Pharmacoepidemiol Drug Saf 2003;12:17–29.
[253] Schneeweiss S, Patrick AR, Sturmer T, Brookhart MA, Avorn J, Maclure M, et al. Increasing Levels of Restriction in Pharmacoepidemiologic Database Studies of Elderly and Comparison With Randomized Trial Results. Med Care 2007;45:S131–42. doi:10.1097/MLR.0b013e318070c08e.
[254] Golder S, Loke YK, Bland M. Meta-analyses of Adverse Effects Data Derived from Randomised Controlled Trials as Compared to Observational Studies: Methodological Overview. PLoS Med 2011;8:e1001026. doi:10.1371/journal.pmed.1001026.
[255] Schneeweiss S. Learning from Big Health Care Data. N Engl J Med 2014;370:2161–3. doi:10.1056/NEJMp1401111.
[256] Harper E. Can big data transform electronic health records into learning health systems? Stud Health Technol Inform 2014;201:470–5.
[257] Bate A, Lindquist M, Edwards IR, Orre R. A data mining approach for signal detection and analysis. Drug Saf 2002;25:393–7. doi:250602.
[258] Hsu J. Multiple comparisons: theory and methods. CRC Press; 1996.
[259] Reich CG, Ryan PB, Suchard MA. The Impact of Drug and Outcome Prevalence on the Feasibility and Performance of Analytical Methods for a Risk Identification and Analysis System. Drug Saf 2013;36:195–204. doi:10.1007/s40264-013-0112-0.
[260] West SL, Johnson W, Visscher W, Kluckman M, Qin Y, Larsen A. The challenges of linking health insurer claims with electronic medical records. Health Informatics J 2014;20:22–34. doi:10.1177/1460458213476506.
[261] Ficheur G, Chazard E, Schaffar A, Genty M, Beuscart R. Interoperability of medical databases: construction of mapping between hospitals laboratory results assisted by automated comparison of their distributions. AMIA Annu Symp Proc AMIA Symp AMIA Symp 2011;2011:392–401.
[262] Sedano FJF, Cuadrado MT, Rebolledo EMG, Clemente YC, Balazote PS, Delgado ÁG. Implementation of SNOMED CT to the Medicines Database of a General Hospital.
[263] Kurtz SM, Ong KL, Lau E, Widmer M, Maravic M, Gómez-Barrena E, et al. International survey of primary and revision total knee replacement. Int Orthop 2011;35:1783–9. doi:10.1007/s00264-011-1235-5.
[264] Hagner PR, Schneider A, Gartenhaus RB. Targeting the translational machinery as a novel treatment strategy for hematologic malignancies. Blood 2010;115:2127–35. doi:10.1182/blood-2009-09-220020.
[265] Quan H. ICD-10-CA/CCI coding algorithms for defining clinical variables to assess outcome after aortic and mitral valve replacement surgery. Can J Cardiol 2006;22:153–4.
[266] Quality Guidelines for statistical processes using administrative data

[267] Mannes A, Moore D. I know I’m right! A behavioural view of overconfidence. Significance 2013;10:10–4.
[268] Avorn J. In defense of pharmacoepidemiology--embracing the yin and yang of drug research. N Engl J Med 2007;357:2219–21. doi:10.1056/NEJMp0706892.
[269] Taleb NN. The Black Swan:: The Impact of the Highly Improbable Fragility. Random House LLC; 2010.
[270] Young SS, Karr A. Deming, data and observational studies. Significance 2011;8:116–20.
[271] Grodstein F, Stampfer MJ, Manson JE, Colditz GA, Willett WC, Rosner B, et al. Postmenopausal estrogen and progestin use and the risk of cardiovascular disease. N Engl J Med 1996;335:453–61.
[272] Varas-Lorenzo C, García-Rodríguez LA, Perez-Gutthann S, Duque-Oliart A. Hormone replacement therapy and incidence of acute myocardial infarction A population-based nested case-control study. Circulation 2000;101:2572–8.
[273] Wilson PW, Garrison RJ, Castelli WP. Postmenopausal estrogen use, cigarette smoking, and cardiovascular morbidity in women over 50: the Framingham Study. N Engl J Med 1985;313:1038–43.
[274] Manson JE, Hsia J, Johnson KC, Rossouw JE, Assaf AR, Lasser NL, et al. Estrogen plus progestin and the risk of coronary heart disease. N Engl J Med 2003;349:523–34.
[275] Green J, Czanner G, Reeves G, Watson J, Wise L, Beral V. Oral bisphosphonates and risk of cancer of oesophagus, stomach, and colorectum: case-control analysis within a UK primary care cohort. BMJ 2010;341.
[276] Cardwell CR, Abnet CC, Cantwell MM, Murray LJ. Exposure to oral bisphosphonates and risk of esophageal cancer. JAMA 2010;304:657–63.
[277] Preparing for Responsible Sharing of Clinical Trial Data. N Engl J Med 2014;370:484–5. doi:10.1056/NEJMc1314515.

Auteur : Docteur Grégoire Ficheur

Date de soutenance : jeudi 11 juin 2015

Titre de la thèse : « Réutilisation de données hospitalières pour la recherche d'effets indésirables liés à la prise d'un médicament ou à la pose d'un dispositif médical implantable »

Directeur de thèse : Professeur Régis Beuscart

Mots clés : données massives, réutilisation de données, pharmaco-épidémiologie, évènement indésirable, cas-témoin en cross-over

Introduction : les effets indésirables associés à un traitement médicamenteux ou à la pose d'un dispositif médical implantable doivent être recherchés systématiquement après le début de leur commercialisation. Les études réalisées pendant cette phase sont des études observationnelles qui peuvent s'envisager à partir des bases de données hospitalières. L'objectif de ce travail est d'étudier l'intérêt de la ré-utilisation de données hospitalières pour la mise en évidence de tels effets indésirables.

Matériel et méthodes : deux bases de données hospitalières sont ré-utilisées pour les années 2007 à 2013 : une première contenant 171 000 000 de séjours hospitaliers incluant les codes diagnostiques, les codes d'actes et des données démographiques, ces données étant chaînées selon un identifiant unique de patient ; une seconde issue d'un centre hospitalier contenant les mêmes types d'informations pour 80 000 séjours ainsi que les résultats de biologie médicale, les administrations médicamenteuses et les courriers hospitaliers pour chacun des séjours. Quatre études sont conduites sur ces données afin d'identifier d'une part des évènements indésirables médicamenteux et d'autre part des évènements indésirables faisant suite à la pose d'un dispositif médical implantable.

Résultats : la première étude démontre l'aptitude d'un jeu de règles de détection à identifier automatiquement les effets indésirables à type d'hyperkaliémie. Une deuxième étude décrit la variation d'un paramètre de biologie médicale associée à la présence d'un motif séquentiel fréquent composé d'administrations de médicaments et de résultats de biologie médicale. Un troisième travail a permis la construction d'un outil web permettant d'explorer à la volée les motifs de réhospitalisation des patients ayant eu une pose de dispositif médical implantable. Une quatrième et dernière étude a permis l'estimation du risque thrombotique et hémorragique faisant suite à la pose d'une prothèse totale de hanche.

Conclusion : la ré-utilisation de données hospitalières dans une perspective pharmacoépidémiologique permet l'identification d'effets indésirables associés à une administration de médicament ou à la pose d'un dispositif médical implantable. L'intérêt de ces données réside dans la puissance statistique qu'elles apportent ainsi que dans la multiplicité des types de recherches d'association qu'elles permettent.

Adresse de l'auteur : 27 rue de la Halloterie, 59000 LILLE - gregoire.ficheur@univ-lille2.fr