UNIVERSITE BIOLOGIE BRETAGNE SANTE LOIRE

UNIVERSITÉ DE RENNES 1

THESE DE DOCTORAT DE

L'UNIVERSITE DE RENNES 1
COMUE UNIVERSITE BRETAGNE LOIRE

ECOLE DOCTORALE N° 605
Biologie Santé
Spécialité : Génétique, Génomique, Bioinformatique

Par
Yann RIVAULT

Analyse de trajectoires de soins à partir de bases de données médico-administratives : apport d'un enrichissement par des connaissances biomédicales issues du Web des données

Thèse présentée et soutenue à Rennes, le 28 janvier 2019
Unité de recherche : Univ Rennes, EHESP, REPERES - EA 7449
Univ Rennes, CNRS, Inria, IRISA – UMR 6074

Rapporteurs avant soutenance :

Jean-Baptiste Lamy	Maître de conférences, Université Paris 13, LIMICS, INSERM UMRS 1142
Lina Soualmia	Maître de conférences, Université de Rouen, LITIS

Composition du Jury :

Examinateurs :	Anita Burgun	Professeure des Universités, Université Paris Descartes – Praticienne Hospitalière HEGP, INSERM UMR-S 872 équipe 22
	Jean-François Ethier	Professeur agrégé et clinicien-chercheur, Université de Sherbrooke. INSERM UMRS 1138
Dir. de thèse :	Olivier Dameron	Maître de conférences, Univ. Rennes 1, IRISA
	Nolwenn Le Meur	Professeure de l'EHESP, EA 7449 REPERES

Remerciements

Je souhaite tout d'abord exprimer ma gratitude envers Olivier et Nolwenn, les directeurs de cette thèse. Vous m'avez accordé votre confiance très tôt, aussi bien dans les travaux de cette thèse que dans l'encadrement de travaux dirigés, et vous avez ainsi renforcé mon envie d'apprendre et de transmettre. Votre patience, votre écoute, et votre encadrement régulier ont fortement contribué au bon déroulement de cette thèse. Merci.

Je tiens à remercier Madame Lina Soualmia ainsi que Monsieur Jean-Baptiste Lamy pour avoir accepté de juger le travail de cette thèse en tant que rapporteurs. Je remercie également Madame Anita Burgun et Monsieur Jean-François Ethier pour leur participation au jury de cette thèse en tant qu'examinateurs.

Cette thèse s'est déroulée dans le cadre du consortium de Pharmaco-Épidémiologie des Produits de Santé (PEPS). Je souhaite remercier tous les membres du consortium, et particulièrement Emmanuel Oger, son coordinateur.

Accueilli au sein de l'équipe REPERES et du département MéTiS à l'EHESP, je souhaite remercier tous leurs membres. Vous côtoyer m'a permis de mieux comprendre les enjeux de santé publique, la pharmaco-épidémiologie et l'épidémiologie. Vos remarques, avis et commentaires ont fait évoluer mon travail de thèse. Je vous remercie également pour vos encouragements et votre enthousiasme à l'égard du travail mené par les jeunes chercheurs.

Je souhaite également remercier l'équipe Dyliss à l'IRISA. Mes rares présences pour des séminaires ont toujours été très enrichissantes.

J'adresse mes sincères remerciements à Nicolas Jégou. Tu m'as introduit à la recherche lors de mon master, à la fin duquel tu m'as toi aussi accordé très rapidement ta confiance pour encadrer des travaux dirigés en statistiques. Cet enseignement fut pour moi une expérience mémorable, très enrichissante –je crois avoir appris autant que les élèves– que je souhaite bien sûr réitérer.

Je souhaite remercier Nicolas Jay et Aurelie Bannay. Votre bienveillance ont fait de ma mobilité un moment très agréable. Vos commentaires, remarques et propositions ont toujours été très pertinentes, aussi bien dans notre collaboration que pour mon travail de thèse. Je souhaite également remercier les membres du Département d'Information Médicale de l'Hôpital Central de Nancy, ainsi que les membres de l'équipe Orpailleur du LORIA, où j'ai été accueilli lors de cette mobilité. Merci pour l'intérêt que vous avez manifesté à l'égard de mes travaux lorsque j'ai pu vous les présenter.

Je remercie Thomas Guyet et Yann Dauxais. J'ai beaucoup apprécié travailler avec vous. Je repense également à notre virée à Vienne pour AIME 2017, c'était de très bons moments en votre compagnie.

Je tiens à remercier mes collègues proches au sein du département MéTiS, amis, jeunes chercheurs, docteurs, doctorants, stagiaires et internes. J'ai été content de travailler dans un tel environnement, d'accueil, d'écoute et d'entraide. Vous y êtes tous pour beaucoup. Vincent, Jonathan, Mathilde, je suis heureux d'avoir partagé avec vous notre bureau. J'espère que mes monologues ne vous manqueront pas trop.

Je remercie ma famille et mes amis pour leurs encouragements tout au long de ces trois années. Julie, je te remercie particulièrement pour ton enthousiasme et tes encouragements constants. Tu as su me remotiver lorsque j'étais pris de doutes. Merci.

Cette thèse a été financée par l'Agence Nationale de Sécurité du Médicament et des produits de santé (ANSM), au travers du consortium de Pharmaco-Épidémiologie des Produits de Santé (PEPS) coordonné par le Pr. Emmanuel Oger.

Glossaire

ANSM	Agence Nationale de sécurité du médicament et des produits de santé
Afssaps	Agence française de sécurité sanitaire des produits de santé
ASP	Answer Set Progamming
ATC	Classification Anatomique Thérapeutique et Chimique
ATIH	Agence Technique de l'Information sur l'Hospitalisation
ARS	Agence Régionale de Santé
CCAM	Classification Commune des Actes Médicaux
CIM-10	Classification Internationnale des Maladies, 10^{e} révision
CIP	Code Identifiant de Présentation
CISMEF	Catalogue et Index des Sites Médicaux de langue Française
CNIL	Commission Nationale de l'Informatique et des Libertés
CUI	Concept Unique Identifier
DID	Drug Indication Database
DIKB	Drug Interaction Knowledge Base
EGB	Échantillon Généraliste des Bénéficiaires de l'Assurance Maladie
GHM	Groupe Homogène de Maladie
HAS	Haute Autorité de Santé
ICD-10	The 10th revision of the International Statistical of Diseases and Related Health Problems
IDS	Institut des Données de Santé
LCS	Longest Common Subsequence
MCO	Médecine Chirurgie Obstétrique
MSAP	Mise Sous Accord Préalable d'hébergement
NABM	Nomenclature des Actes de Biologie Médicale
NDF-RT	National Drug File - Reference Terminology
NLM	National Library of Medecine
OMS	Organisation Mondiale de la Santé
OWL	Web Ontology Language
PEPS	Consortium de Pharmaco-Épidémiologie des Produits de Santé
PMSI	Programme de Médicalisation des Systèmes d'Information
RDF	Ressource Description Framework
RDFS	RDF schema
RSS	Résumé de Sortie Standardisé
RUM	Résumé d'Unité Médicale
SIFR	Semantic Indexing of French Biomedical Data Resources
SNDS	Système National des Données de Santé

SNIIRAM	Système National d'Information Inter-Régime de l'Assurance Maladie
SOC	Système d'Organisation de la Connaissance
SPARQL	SPARQL Protocol And Query Language
SQL	Structured Query Language
T2A	Tarification À l'Activité
UMLS	Unified Medical Language System
W3C	World Wide Web Consortium

Table des matières

Chapitre 1

Introduction générale

Sommaire

1.1 Contexte et besoins

Le contexte juridique récent [1] a montré que la sécurité des médicaments et des produits de santé reste une préoccupation majeure des autorités de santé publique. En 2011, l'Agence Nationale de Sécurité du Médicament et des produits de santé (ANSM) est officiellement créée et remplace l'Agence française de sécurité sanitaire des produits de santé (Afssaps), dans le cadre de la loi du 29 décembre 2011 relative au renforcement de la sécurité sanitaire du médicament et des produits de santé. Parallèlement, la loi de modernisation de notre système de santé a contribué à l'ouverture du champs des possibles en terme de recherche en santé publique, par la volonté de rendre certaines données de santé plus accessibles aux chercheurs, notamment les bases de données médico-administratives. C'est dans ce contexte que fut lancé en 2015 la plateforme PEPS [2] (consortium de Pharmaco-Épidémiologie des Produits de Santé), financée par l'ANSM. Son objectif principal est la production d'études pharmaco-épidémiologiques par l'exploitation des bases de données médico-administratives françaises. Combiner données de recours à l'offre de soins et données de consommation de soins doit permettre aux chercheurs du consortium d'étudier les risques et bénéfices des produits de santé ainsi que les impacts de leurs conditions d'usage. Cet objectif s'accompagne de celui du développement d'outils et de méthodes génériques facilitant l'exploitation et la ré-utilisation de ces données volumineuses et complexes.

1. LOI n° 2011-2012 du 29 décembre 2011 relative au renforcement de la sécurité sanitaire du médicament et des produits de santé : https://www.legifrance.gouv.fr/affichTexte.do;jsessionid=A81B7C5972601ECB1875340DFD7649A8.tplgfr26s_3?cidTexte=JORFTEXT000025053440&dateTexte=29990101

2. Annonce officielle du lancement de la plateforme PEPS par l'ANSM : https://ansm.sante.fr/S-informer/Communiques-Communiques-Points-presse/Le-CHU-de-Rennes-lance-officiellement-le-consortium-de-Pharmaco-Epidemiologie-des-Produits-de-Sante-PEPS-finance-par-l-ANSM-Communique

Les bases de données médico-administratives françaises recueillent l'exhaustivité des données de consommation et de recours aux soins remboursées par l'Assurance Maladie. De ce fait, il est devenu possible de considérer les données patients comme des trajectoires de soins, traces ou séquences temporelles constituées des événements de santé enregistrés par l'Assurance Maladie. Cette nouvelle vision des données de santé se rapproche étroitement de la notion de parcours de soins, pendant théorique –et non plus observé– de la trajectoire de soins. Les recommandations de bonnes pratiques des autorités et professionnels de santé[3,4] sont un exemple de parcours de soins. Cette vision des données de santé a également amené à la ré-utilisation de méthodes adaptées pour l'analyse de trajectoires. Des méthodes issues de l'informatique, par exemple des méthodes de comparaison de chaîne de caractères ou de séquences biologiques, utilisées avec succès en sociologie pour comparer des trajectoires d'événements de vie (Ritschard et al., 2008), ont ainsi été réutilisées pour comparer des trajectoires d'événements de santé (Le Meur et al., 2015; Roux et al., 2018b). Cette analyse de trajectoires a pu s'illustrer de différentes façons, allant de l'exploration des trajectoires de soins à leurs comparaisons, à la création de groupes homogènes de trajectoires, jusqu'à l'extraction de motifs fréquents à partir de trajectoires. Ces méthodes se sont cependant heurtées aux complexités et limites inhérentes à la réutilisation des bases de données médico-administratives françaises pour la recherche en santé publique. Le caractère massif des données, leur hétérogénéité, et l'importante variabilité des données ont pu rendre ces méthodes difficiles à appliquer sur des trajectoires de soins.

Si les nomenclatures et classifications qui régissent les données de santé ont pu parfois aussi être perçues comme une complexité ou comme une contrainte, notamment lors de l'enregistrement des données, elle permettent de structurer et normaliser les données, et de les lier à des systèmes d'organisation de connaissances (Park and Hardiker, 2009). Baser l'exploration et l'analyse des données sur l'apport de connaissance peuvent alors permettre d'amoindrir les limites et complexités des données (Anjum et al., 2007). Par exemple, la quantité de médicaments différents délivrés ou de diagnostics différents réalisés, créent tellement de variabilité qu'il en devient parfois difficile de traiter statistiquement cette information. La connaissance de la structure hiérarchique des nomenclatures, c'est à dire des familles, classes ou chapitres de médicaments, diagnostics et actes médicaux est dans ce cas indispensable au pharmaco-épidémiologiste dans l'analyse des données. De même, d'autres connaissances médicales, externes aux données, peuvent représenter un réel atout dans la réutilisation des bases médico-administratives. Par exemple, quand on étudie les risques et bénéfices des produits de santé, tenir compte de ceux qui sont déjà connus et répertoriés, en les liant aux nomenclatures et classifications que l'on utilise, est essentiel. L'étude des risques d'interactions entre médicaments est un autre

3. Recommandations de bonnes pratiques de la HAS : https://www.has-sante.fr/portail/jcms/c_1101438/fr/tableau-des-recommandations-ou-travaux-relatifs-a-la-bonne-pratique

4. Recommandations de bonnes pratiques des produits de santé de l'ANSM : https://ansm.sante.fr/Mediatheque/Publications/Recommandations-Medicaments

bon exemple pour démontrer la nécessité de connaissances externes aux données, en l'occurrence d'une bases de connaissances relatives aux interactions médicamenteuses connues (Pathak et al., 2013). L'adaptation des étapes de l'exploration et de l'analyse des données aux connaissances médicales et pharmacologiques disponibles sont dès lors un réel besoin dans le cadre de la réutilisation des données médico-administratives pour la recherche en santé publique, notamment en épidémiologie (Ferreira et al., 2013).

Les systèmes d'organisation des connaissances médicales et pharmacologiques sont cependant de plus en plus nombreux, se chevauchent parfois, avec des schémas de représentation différents. Couplé à une documentation éparse et parfois même incomplète, leur utilisation est loin d'être triviale (Jain et al., 2010). Favoriser la réutilisation de ces systèmes d'organisation des connaissances, en facilitant leur accès et utilisation, est donc un prérequis à la diffusion de méthodes d'exploration et d'analyse de trajectoires de soins enrichie par un apport de connaissances médicales et pharmacologiques.

1.2 Enjeux de la thèse

Les enjeux de cette thèse sont les suivants :

1. Profiter de l'apport de connaissances médicales et pharmacologiques pour la représentation, l'intégration et l'exploration des données médico-administratives/trajectoires de soins ;
2. Adapter l'analyse des trajectoires de soins par des connaissances médicales et pharmacologiques ;
3. Faciliter l'accès des connaissances médicales et pharmacologiques et leurs liens avec les nomenclatures médicales utilisées dans les bases de données médico-administratives.

1.3 Positionnement de la thèse

Afin de poursuivre ces objectifs, nous avons proposé dans cette thèse d'étudier la pertinence des technologies du Web Sémantique et des systèmes d'organisation de connaissances du Web des données, pour la représentation, l'intégration et l'exploration des données issues des bases médico-administratives. Nous proposons également d'utiliser ces technologies et systèmes d'organisation de la connaissance médicale afin d'enrichir des méthodes d'analyse et de fouille de données, pour qu'elles soient plus adaptées aux objets complexes que sont les trajectoires de soins. Particulièrement, nous proposons de prendre en compte les structures hiérarchiques de nomenclatures médicales dans les méthodes de comparaison de trajectoires de soins, et de fouille de motifs à partir de trajectoires de soins. Une exploration des résultats de fouille de données basées sur les technologies du Web Sémantique est-elle aussi menée dans cette thèse. Enfin, dans le but de favoriser la réutilisation des ontologies biomédicales pour la recherche en santé publique et plus spécifiquement dans le domaine de

la pharmaco-épidémiologie, nous avons proposé un package R pour faciliter les liens entre connaissances et données de santé, au sein d'un environnement de statistique que les pharmaco-épidémiologiste utilisent de plus en plus.

Chapitre 2

État de l'art

Sommaire

2.1 Réutiliser les bases de données médico-administratives françaises pour la recherche en santé publique

2.1.1 Santé publique et données médicales

La santé publique couvre aussi bien l'étude des facteurs déterminants des états de santé de la population, que les mesures qui ont pour objectif l'amélioration de

cet état général de santé [1]. De par cette définition, la santé publique se veut pluridisciplinaire, en englobant de nombreuses disciplines différentes, allant ainsi de l'épidémiologie et la pharmaco-épidémiologie, à la sécurité des soins, jusqu'au management des systèmes de santé. Ces disciplines bien différentes, ont cependant en commun au moins un besoin. Pour évoluer dans leurs études, elles doivent comprendre le contexte de santé et de soins d'une population (Century, Institute of Medicine (US) Committee on Assuring the Health of the Public in the 21st, 2002), qu'elles réalisent par le recueil d'informations.

L'épidémiologie, dans la description et la quantification des maladies, ainsi que dans la recherche de leurs déterminants potentiels, est sans doute la discipline en santé publique la plus habituée au recueil et à l'analyse de données de santé. Les études de cohortes recueillent par exemple des informations concernant les caractéristiques et expositions individuelles, pour suivre dans le temps des groupes d'individus. Du fait de leur caractère longitudinal, elles sont particulièrement utilisées pour étudier des expositions et leurs associations avec le risque de la survenue de maladies. Les études cas-témoins ont également cet objectif, mais sont des études statistiques rétrospectives. Une fois les individus malades sélectionnés (les cas), leur historique est comparé à ceux des individus non malades (les témoins), dans le but de tester une hypothèse sur l'association d'un facteur de risque antérieur à l'étude et la survenue de la maladie étudiée. Néanmoins, la qualité de données rétrospectives peut varier avec l'ancienneté des facteurs potentiels. Une étude prospective type cohorte doit elle débuter assez tôt pour capturer les potentiels facteurs de risque, tout comme durer assez longtemps pour que la maladie se déclare. Aussi, quand de tels résultats épidémiologiques, sur l'association entre facteurs de risque et une maladie, ou même faisant la description et la quantification d'une maladie, sont publiées, leur inférence, c'est à dire leur extension à une population plus générale, est toujours à considérer au regard de la population sélectionnée. Des études sur la totalité de la population française effacerait ce problème. De telles analyses peuvent toutefois être difficilement réalisables si aucun recueil systématique n'est mis en place.

En pharmaco-épidémiologie, les méthodes épidémiologiques sont ré-utilisées pour évaluer l'efficacité, les bénéfices ainsi que les risques d'un médicament. Ces études, nécessitent donc un recueil d'informations sur les caractéristiques sociodémographiques et médicales de la population étudiée en vie réelle, indirectement ciblée par le médicament étudié.

La sécurité des soins, avec par exemple le suivi des recommandations pour une maladie, un état de santé, ou une prise en charge médicalisée, sont des études qui nécessitent des données relatives aux consommations de soins (consommations de médicaments et consultations médicales par exemple) et états de santé. La pharmacovigilance, par l'objectif de surveiller l'usage des médicaments et de prévenir le risque d'effets indésirables, repose notamment sur un recueil d'information basé sur la notification spontanée, et donc non systématique, des effets indésirables par les

1. Code de santé publique, "Politique de santé publique en France" : https://www.legifrance.gouv.fr/affichCode.do;jsessionid=FB58EF5FD19BC12F33BE5FA4FA25E0A4.tpdjo07v_3?idSectionTA=LEGISCTA000006171073&cidTexte=LEGITEXT000006072665&dateTexte=20120804

professionnels de santé, patients, associations de patients et industriels, avec l'appui du réseau de 31 centres régionaux de Pharmacovigilance [2].

Enfin, l'organisation du système de santé doit comprendre –pour organiser– l'ensemble des éléments qui le forment. L'analyse des données émises par tous les éléments du système de santé (établissements hospitaliers, pharmacies, médecine libérale, etc) est essentielle pour son organisation. Cette analyse permet notamment d'élaborer et de piloter des politiques de santé publique, par exemple de prendre des mesures pour une allocation financière plus juste des établissements de soins publics (Jay et al., 2013).

La santé publique et ses disciplines ressentent ainsi de plus en plus le besoin d'un recueil d'informations qui vérifie les exigences suivantes :

— Un recueil d'informations médicales détaillées : maladies et états de santé, offre et consommation de soins, résultats médicaux ;

— Un recueil d'informations socio-démographiques détaillées ;

— Un recueil d'informations administratives relatives au système de santé dans sa globalité, notamment pour son organisation ;

— Un recueil systématique ;

— Un recueil en vie réelle ;

— Un recueil à l'échelle d'une population ;

— Un recueil sur une période de temps importante permettant des études longitudinales.

2.1.2 Les systèmes d'information médico-administratifs français

La France dispose d'un ensemble de systèmes d'information médico-administratifs, souvent appelés bases de données médico-administratives (Tuppin et al., 2017). De façon globale, ces systèmes d'information recueillent les données émises par les remboursements de l'Assurance Maladie pour les bénéficiaires de la quasi totalité des régimes de la sécurité sociale (soit près de 98% des assurés). Si leurs objectifs initiaux ont été comptables et gestionnaires, ces systèmes d'information et leurs données sont elles de plus en plus réutilisées pour la recherche en santé publique, car ils satisfont justement certaines des exigences énumérées en section 2.1.1. Ils sont ainsi de plus en plus ré-utilisés pour des études épidémiologiques, pharmaco-épidémiologiques, de la sécurité des soins ou encore de la pharmacovigilance (Daien et al., 2017).

2.1.2.1 Programme de Médicalisation des Systèmes d'Information

C'est en 1982, avec le projet de médicalisation des systèmes d'information, que le Programme de Médicalisation des Systèmes d'Information (PMSI) voit le jour. Si

2. Organisation de la pharmacovigilance Nationale par l'ANSM : https://ansm.sante.fr/Declarer-un-effet-indesirable/Pharmacovigilance/Organisation-de-la-pharmacovigilance-nationale/(offset)/0

les données recueillies par le PMSI ne sont utilisées dans un premier temps que pour décrire l'activité des hôpitaux et réaliser un suivi épidémiologique de leurs patients, ce système d'information devient peu à peu une base de données à visée financière et gestionnaire. À partir de 1991 avec le code de la santé publique [3], les hôpitaux doivent grâce au PMSI, procéder à l'évaluation et à l'analyse de leurs activités. C'est en France le début d'une gestion des hôpitaux aidée par un système d'information, le PMSI. En 2005, le plan hôpital 2007 [4] instaure la tarification à l'activité (T2A) : les établissements sont financés en fonction de leurs activités. Le PMSI fournit alors un moyen d'évaluer cette activité et trouve ainsi une application purement financière.

Données du PMSI Chaque hospitalisation en médecine, chirurgie ou obstétrique (MCO) dans un établissement de santé public ou privé, mène à la création d'un résumé de sortie standardisé (RSS). Ce RSS est lui-même constitué d'un résumé d'unité médicale (RUM) pour chaque unité médicale fréquentée durant le séjour hospitalier. Les RUM contiennent des informations administratives et médicales [5] :

- Informations administratives :
  - Identifiants RSS ;
  - Numéro de l'établissement ;
  - Date de naissance ;
  - Sexe ;
  - Code postal de résidence ;
  - Numéro de l'unité médicale d'hospitalisation ;
  - Dates et modes d'entrée et de sortie, provenance et destination ;
  - Nombre de séances.
- Informations médicales :
  - Diagnostic principal : le problème de santé qui a motivé l'admission du patient dans l'unité médicale ;
  - Diagnostic relié : maladie à l'origine du problème de santé, s'il y en a une ;
  - Diagnostics associés ;
  - Actes médicaux ;

3. Article L710-5 du Code de la santé publique créé par la Loi n°91-748 du 31 juillet 1991 : https://www.legifrance.gouv.fr/affichCodeArticle.do;jsessionid=BFA18C7C84E4EEAAD3C0BF7FD5186F51.tplgfr41s_1?idArticle=LEGIARTI000006694595&cidTexte=LEGITEXT000006072665&categorieLien=id&dateTexte=19930129

4. Ordonnance n° 2005-406 du 2 mai 2005 simplifiant le régime juridique des établissements de santé : https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000606537&categorieLien=cid#

5. Arrêté du 22 février 2008 répertoriant les informations administratives et médicales présentes dans le PMSI : https://www.legifrance.gouv.fr/jo_pdf.do?numJO=0&dateJO=20080229&numTexte=60&pageDebut=03577&pageFin=03579

— Types de dosimétrie et de machine en radiothérapie ;

— Poids à l'entrée dans l'unité médicale pour le nouveau-né ;

— Âge gestationnel de la mère et du nouveau-né, dates des dernières règles de la mère ;

— Indice de gravité simplifié ;

— Données à visée documentaire.

Les diagnostics, principaux, reliés et associés sont codés selon la Classification Internationale des Maladies-10e révision [6] (CIM-10), en anglais *the 10th revision of the International Statistical of Diseases and Related Health Problems* (ICD-10). Les actes médicaux sont eux codés selon la Classification Commune des Actes Médicaux [7] (CCAM). Le RSS est en plus classé dans un groupe homogène de maladie (GHM), une classification française adaptée des *Diagnosis Related Groups* (Fetter et al., 1980). Cette procédure classifie chaque RSS dans un groupe, sur la base des données médicales et administratives qui lui sont associées. Les établissements sont ensuite financés au regard des GHM transmis à l'assurance maladie via les Agences Régionales de Santé (ARS).

2.1.2.2 Système National d'Information Inter-Régime de l'Assurance Maladie

En 1999 est créé par la loi de financement de l'Assurance Maladie, le Système National d'Information Inter-Régime de l'Assurance Maladie (SNIIRAM). Défini par le code de la sécurité sociale [8], le SNIIRAM contribue :

1. « À la connaissance des dépenses de l'ensemble des régimes d'assurance maladie par circonscription géographique, par nature de dépenses, par catégorie de professionnels responsables de ces dépenses et par professionnel ou établissement » ;
2. « À la transmission en retour aux prestataires de soins d'informations pertinentes relatives à leur activité et leurs recettes, et s'il y a lieu à leurs prescriptions » ;
3. « À la définition, à la mise en œuvre et à l'évaluation de politiques de santé publique ».

Cette définition souligne bien les objectifs initiaux d'une meilleure gestion et évaluation de l'Assurance Maladie ainsi que des politiques de santé publique.

6. La CIM-10 sur le site de l'Organisation Mondiale de la Santé (OMS) : http://www.who.int/classifications/icd/icd10updates/en/

7. La CCAM sur le site de l'Assurance Maladie : https://www.ameli.fr/accueil-de-la-ccam/index.php

8. Code de la sécurité sociale - Article L161-28-1 : https://www.legifrance.gouv.fr/affichCodeArticle.do;jsessionid=D94E263FD2ECC58BB1E2E4A6A3EFAFB4.tpdjo02v_1?idArticle=LEGIARTI000006741267&cidTexte=LEGITEXT000006073189&categorieLien=id&dateTexte=20140121

Données du SNIIRAM En plus de contenir des données du PMSI pour l'ensemble des établissements de santé, le SNIIRAM contient des données liées au remboursement de l'Assurance Maladie pour les consommations de soins en ville, pour exemple :

— Actes médicaux réalisés en ville, codés selon la CCAM ;

— Actes de biologie médicale, codés selon la nomenclature des actes de biologie médicale (NABM) ;

— Dispositif médicaux ;

— Délivrance des médicaments prescrits, codés selon les codes identifiants de présentation (CIP).

2.1.2.3 Échantillon Généraliste des Bénéficiaires de l'Assurance Maladie

L'arrêté du 20 juin 2005 [9] crée l'Échantillon Généraliste des Bénéficiaires de l'Assurance Maladie (EGB). L'EGB est un échantillon du SNIIRAM, qui résulte d'un sondage au $1/97^{e}$ sur le numéro de sécurité sociale des bénéficiaires de l'Assurance Maladie. Les répartitions de l'âge, du sexe ou encore des dépenses moyennes de remboursements de soins sont proches de la population totale (Tuppin et al., 2010). De ce fait, il est qualité d'échantillon représentatif de la population française. Par la création de cet échantillon, les autorités de santé publique confirment leur volonté de réutiliser les bases de données médico-administratives pour des finalités autres que managériales et financières. L'objectif de l'EGB est en effet de permettre à des chercheurs de réaliser des études longitudinales sur les trajectoires de soins (de ville et d'hôpital) de patients bénéficiaires de l'Assurance Maladie. Il trouve ainsi particulièrement des applications en pharmaco-épidémiologie (Maura et al., 2018), en épidémiologie (Roux et al., 2018b; Le Meur et al., 2015), ainsi qu'en pharmacovigilance (Létinier et al., 2018).

2.1.2.4 Système National des Données de Santé

En janvier 2016, la loi de modernisation des systèmes de santé [10] annonce et définit le Système National des Données de Santé (SNDS). Il rassemble des bases de données déjà existantes, séparées jusqu'alors, le SNIIRAM, le PMSI et la base de données du Centre d'épidémiologie sur les causes médicales de Décès (CépiDc), base de données recueillant les causes de décès, gérée par l'Institut National de la Santé et de la Recherche Médicale (INSERM). La Caisse Nationale de l'Assurance

9. Arrêté du 20 juin 2005 relatif à la mise en œuvre du système national d'information interrégimes de l'assurance maladie : https://www.legifrance.gouv.fr/affichTexteArticle.do;jsessionid=7F1794E87F02C896F567EB773B6E62A7.tplgfr21s_3?idArticle=JORFARTI000002275726&cidTexte=JORFTEXT000000808427&dateTexte=29990101&categorieLien=id

10. Loi de modernisation du système de santé français : https://www.legifrance.gouv.fr/affichTexteArticle.do;jsessionid=8C3901DF701DE7FCC51453E8D105A11D.tpdila20v_1?idArticle=JORFARTI000031914480&cidTexte=JORFTEXT000031912641&dateTexte=29990101&categorieLien=id

Maladie des Travailleurs Salariés (CNAMTS) est responsable du traitement et de la mise en place du SNIIRAM. C'est l'Institut National des Données de Santé (INDS), qui remplace l'Institut des Données de Santé (IDS) par la même loi, qui veille à la qualité des données, leur mise à disposition, ainsi qu'au respect de leur confidentialité. Pour favoriser l'ouverture des données de santé publique, afin que « leurs potentialités soient utilisées au mieux dans l'intérêt de la collectivité », notamment pour la recherche en santé publique, des accès permanents au SNDS sont accordés à certains organismes publics tels que l'ANSM, l'INSERM ou encore les ARS. Des accès à des fin de recherche, étude ou évaluation dans le domaine de la santé, soumis à l'autorisation de la Commission Nationale de l'Informatique et des Libertés (CNIL), peuvent également mener à l'extraction de données du SNDS.

2.1.3 Complexités des données

Si on a pu voir que le contexte juridique récent a favorisé la réutilisation des bases de données médico-administratives pour la recherche en santé publique, les chercheurs doivent faire face à des complexités qui peuvent parfois limiter cette réutilisation.

La complexité principale est sans doute la volumétrie des données. Puisque recueillies à chaque remboursement d'un soin ou d'une consommation de soin, ces données sont massives. Pour donner un ordre de grandeur, chaque année, près de 2,5 milliards de boîtes de médicaments sont remboursées par l'Assurance Maladie. Ce sont autant d'enregistrements de données dans le SNIIRAM.

En plus d'être massives, ces données sont très hétérogènes, de par leurs sources de recueil notamment. Les données peuvent provenir des hôpitaux, de centres de soins de villes, de pharmacies, ou de tout autre source qui conduirait à un remboursement par L'assurance Maladie. Cette hétérogénéité des sources, va *in fine* mener à une hétérogénéité de la structure des données, de l'Architecture du système d'information. Il se construit en accumulant plusieurs tables de données, qui elles-même regroupent chacune un type de sources de recueil.

Le SNIIRAM, et donc l'EGB, constituent un bon exemple de système d'information à architecture complexe en étoile (figure 2.1). Plusieurs tables de données gravitent autour d'une table principale, celle des prestations de santé. Et pour chaque type de recueil de données (par exemple le recueil de données en pharmacie) est associé une table détaillant la prestation dans ce contexte (la table des prestations affinées en pharmacie, pour suivre cet exemple). Si certaines variables sont partagées par ces nombreuses tables de données (plus de 250), chaque table dispose de variables spécifiques.

Enfin, l'information médicale, que ce soit les actes ou les médicaments délivrés, sont codifiés selon des nomenclatures médicales, nationales ou internationales. Par exemple, les actes médicaux à l'hôpital sont codifiés selon la CIM-10. Cette codification systématique est parfois vue par les épidémiologistes comme une contrainte de plus, une complexité. C'est surtout le nombre de nomenclatures utilisées et leurs profondeur qui peuvent en faire une complexité.

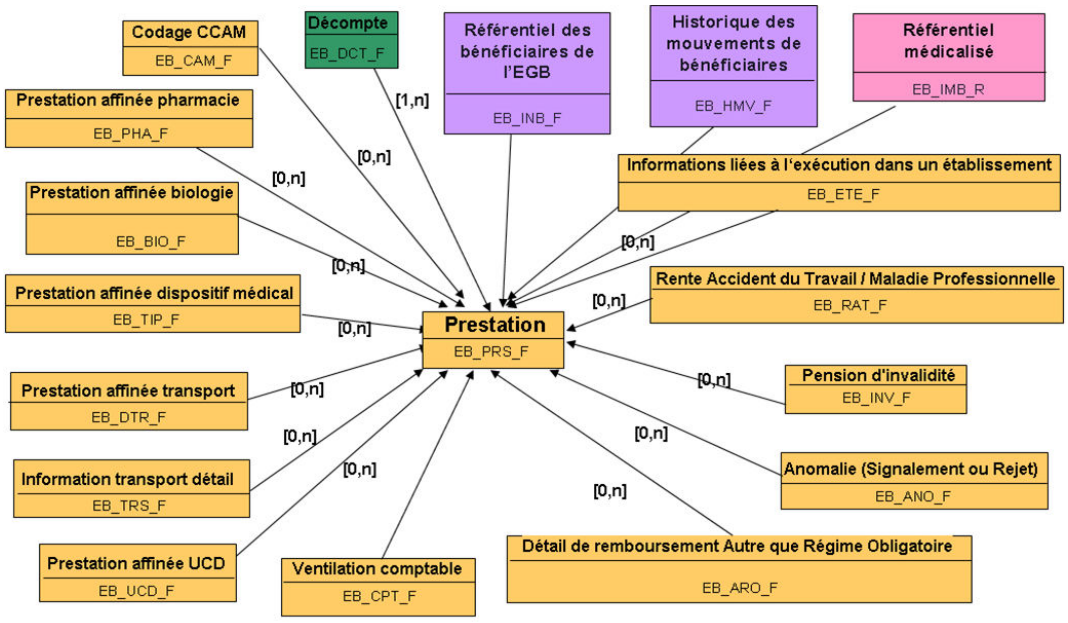

FIGURE 2.1 – Architecture complexe de l'EGB (équivalente au SNIIRAM) en étoile. Une table principale des prestations est reliée à plusieurs tables par type de sources de recueil.

Ces complexités peuvent mener à des limites dans la réutilisation des bases de données médico-administratives pour la recherche en santé publique.

2.1.4 Limites pour la réutilisation des données médico-administratives en santé publique

Volumétrie des données Disposer de données sur toute la population bénéficiaire de l'Assurance Maladie est évidemment d'un grand intérêt pour la recherche en santé publique. Néanmoins, des calculs pour l'exploration ou l'analyse de telles données sont souvent compliqués à réaliser sur une population de plus de 60 millions d'individus. Certaines méthodes statistiques peuvent alors se voir restreintes à des échantillons. L'exploration peut elle aussi être impactée par cet aspect massif, avec par exemple des temps d'exécution pour de la recherche de patient et la constitution de cohortes sur des critères détaillées pouvant dépasser quelques heures.

Variabilité des données Des données déjà très diverses et hétérogènes –car décrivant le domaine complexe médico-administratif– et l'utilisation de plusieurs nomenclatures à vastes vocabulaires mènent à une grande variabilité des données. Il n'est pas rare, même en disposant d'échantillons de grandes tailles, d'avoir plusieurs soins ou consommations de soins dont la fréquence observée ne dépasse pas un cas. Des méthodes d'analyse statistique ou de fouille de données sur des événements si rares sont alors très limitées.

D'autres limites sont elles directement liées à la nature des données.

Nature comptable et managériale des données Les données ont été recueillies dans une optique de mieux gérer et financer l'Assurance Maladie et les établissements de santé. Il en résulte que ces systèmes d'information ne contiennent pas de résultats médicaux, mais uniquement ce qui est remboursé (le remboursement d'un acte médical, ou encore d'une délivrance de médicament). De même, aucune information sur les médicaments non remboursés n'est recueillie dans les bases de données médico-administratives. Ces données peuvent ainsi parfois être jugées de données pauvres, en tout cas dans le cadre de leur réutilisation en épidémiologie et pharmaco-épidémiologie, quand bien même elles couvrent l'ensemble de la population bénéficiaire de l'Assurance Maladie.

Codage des données et financement des établissements La nature comptable et financière mène à une autre limite, liée au mode de financement des établissements de santé. Il est parfois souligné que les hôpitaux, motivés par le financement selon leurs activités, peuvent avoir tendance à adopter des habitudes de sur-codage des séjours (Georgescu and Hartmann, 2013). De la même façon, certains services d'hôpitaux sous-codent certains événements médicaux qui ne rentreraient pas en compte dans le calcul de leur financement. Par exemple, le code "Y95" de la CIM-10 codant une infection nosocomiale, est très rarement utilisé à l'hôpital, car n'étant pas pris en compte dans le calcul du financement des établissements, et car en plus reflétant une infection acquise à l'hôpital, pouvant être liée aux pratiques de soins (Fourquet et al., 2003). Cette limite encore une fois va à l'encontre de la qualité des données. En l'occurrence, l'étude des infections nosocomiales identifiées sur des données médico-administratives traite donc plus souvent des infections qui seraient très probablement acquises en établissement. De telles études requièrent la connaissance d'experts du codage et des infections nosocomiales pour identifier tous les codes CIM-10 correspondant à une telle infection, et ainsi pour contourner cette limite(Grammatico-Guillon et al., 2014).

Confidentialité des données De manière générale, les disciplines de la santé publique, et notamment l'épidémiologie, sont particulièrement encadrées du fait du caractère sensible et confidentiel des données utilisées (Goldberg et al., 2008). Les données issues des systèmes d'information médico-administratifs français sont des données à caractère personnel, très sensibles et confidentielles. Néanmoins –nous l'avons vu en section 2.1.2.4– l'obtention des données du SNDS peut être facilitée pour certains établissements publics de recherche par un accès permanent au portail du SNDS. En revanche, des recherches plus expérimentales ou moins conventionnelles, nécessitant un extraction de données, doivent obtenir l'accord de la CNIL. Les procédures à suivre pour aboutir à l'obtention des données peuvent alors s'avérer assez longues dans ce cas. Elles peuvent constituer une limite à l'avancée de la recherche dans le cadre de la réutilisation des bases de données médico-administratives

françaises à des fins de santé publique.

2.1.5 Les trajectoires et parcours de soins

Si la distinction entre les deux termes n'est parfois pas très claire en français, les chercheurs anglophones y voient une différence essentielle. Dans cette thèse, nous verrons les parcours de soins comme la traduction de *care pathways*, *clinical pathways* ou encore *integrated care pathways*. Les parcours de soins sont ainsi des plans interdisciplinaires de santé qui définissent les étapes importantes dans l'accompagnement d'un patient, pour un contexte clinique et une période données (Campbell et al., 1998). On parle aussi de bonnes pratiques ou de recommandations de santé. Ces plans de recommandations ont montré que leur mise en place pouvait permettre de réduire la variabilité des pratiques de soins (Panella et al., 2003), et ainsi le coût de la prise en charge interdisciplinaire des patients grâce à une meilleure organisation des soins (Deneckere et al., 2012). L'amélioration de la qualité des soins a aussi pu être démontrée (Panella et al., 2003), avec par exemple une réduction des risques de complications à l'hôpital (Rotter et al., 2010).

Si la notion de parcours de soins a été théorisée dès les années 1950 (Schrijvers et al., 2012), celle des trajectoires de soins est bien plus récente. Elle n'en ai cependant sans doute pas moins utilisée, comme le montre la revue systématique de Pinaire et al. (2017b). Les trajectoires de soins sont des traces des soins reçus et états de santé d'un patient sur une période donnée. Les trajectoires de soins sont ainsi des successions d'événements de santé observées, pouvant se rattacher ou non à un ou plusieurs parcours de soins théorique. L'analyse des trajectoires de soins contribue à la compréhension du contexte d'offre et de consommation des soins et produits de santé d'une population. Cette compréhension peut alors permettre d'améliorer la prise en charge des patients pour une maladie ou un état de santé donné, et donc potentiellement amener à de meilleurs résultats cliniques pour les patients (Adeyemi et al., 2013), ainsi qu'à une meilleure planification et gestion des ressources d'un système de santé (Jay et al., 2013). Les bases de données médico-administratives, et particulièrement en France, ont montré qu'elles pouvaient servir à construire ces traces, constituées d'événements médicaux, des prescriptions de médicaments, des actes médicaux en ville ou à l'hôpital, ou encore des diagnostics et états de santé des patients (Defossez et al., 2014; Le Meur et al., 2015).

2.2 Intégrer les données et les lier à des connaissances grâce au Web Sémantique

La réutilisation des bases de données médico-administratives pour la recherche en santé publique, consiste en grande partie à l'analyse statistique de ces données. La connaissance médicale est alors essentielle à de nombreuses étapes lors des études statistiques, pour sélectionner des patients et leurs données, pour les analyses statistiques, ou pour l'interprétation des résultats. Les épidémiologistes utilisent par exemple les connaissances à leur disposition sur une maladie étudiée, pour la constitution d'une cohorte, pour l'étude des potentiels facteurs de risque, jusqu'à l'analyse des résultats statistiques. Des formalisations de cette connaissance ne sont elles que rarement utilisées pour automatiser ces étapes, ce qui est pourtant rendu nécessaire du fait de la volumétrie et la complexité des données médico-administratives. La codification systématique des données médico-administratives contribue pourtant à leur interopérabilité sémantique, en partageant auprès des chercheurs de santé publique une même définition et un même identifiant pour chaque concept médical. L'interopérabilité sémantique des données médico-administratives rend alors possible le lien entre données et formalisations de la connaissance, bien qu'il ne soit en pratique que peu réalisé.

2.2.1 Les systèmes d'organisation des connaissances médicales et pharmacologiques

En matière de formalisation de la connaissance d'un domaine, on parle souvent d'ontologies, quelque fois de systèmes d'organisation ou de représentation de la connaissance (parfois abrévié par SOC). La définition de l'ontologie la plus souvent citée est sans doute celle de Gruber (1995) : « *une ontologie est la spécification d'une conceptualisation. [...] Une conceptualisation est une vue abstraite et simplifiée du monde que l'on veut représenter.* » Bard and Rhee (2004) proposent une définition plus appliquée : « *Une ontologie est une façon formelle de représenter la connaissance dans laquelle des concepts sont décrits à la fois par leur signification et par les relations qui les lient* ». Les ontologies prennent ainsi souvent la forme d'un graphe de concepts. Il existe bien sur d'autre types de SOCs, souvent vues comme des nuances des ontologies. Ainsi, les taxonomies sont des systèmes de classification dans lesquels les concepts définis sont organisés hiérarchiquement. Ils prennent ainsi la forme d'arbres, dont les feuilles et nœuds du tronc sont des concepts. Dans le domaine médical, et notamment dans le cadre de la réutilisation des systèmes d'information médico-administratifs, on peut par exemple évoquer la CIM-10, utilisée pour codifier les maladies et états de santé, la Classification Anatomique Thérapeutique et Chimique (ATC) utilisée pour codifier les médicaments (voir figure 2.2), ou encore la Classification Commune des Actes Médicaux (CCAM) pour codifier les actes médicaux à l'hôpital ou en ville.

Les thésaurus sont quant à eux des listes organisées de termes contrôlés et normalisés représentant les concepts d'un domaine de la connaissance. Si l'objectif pre-

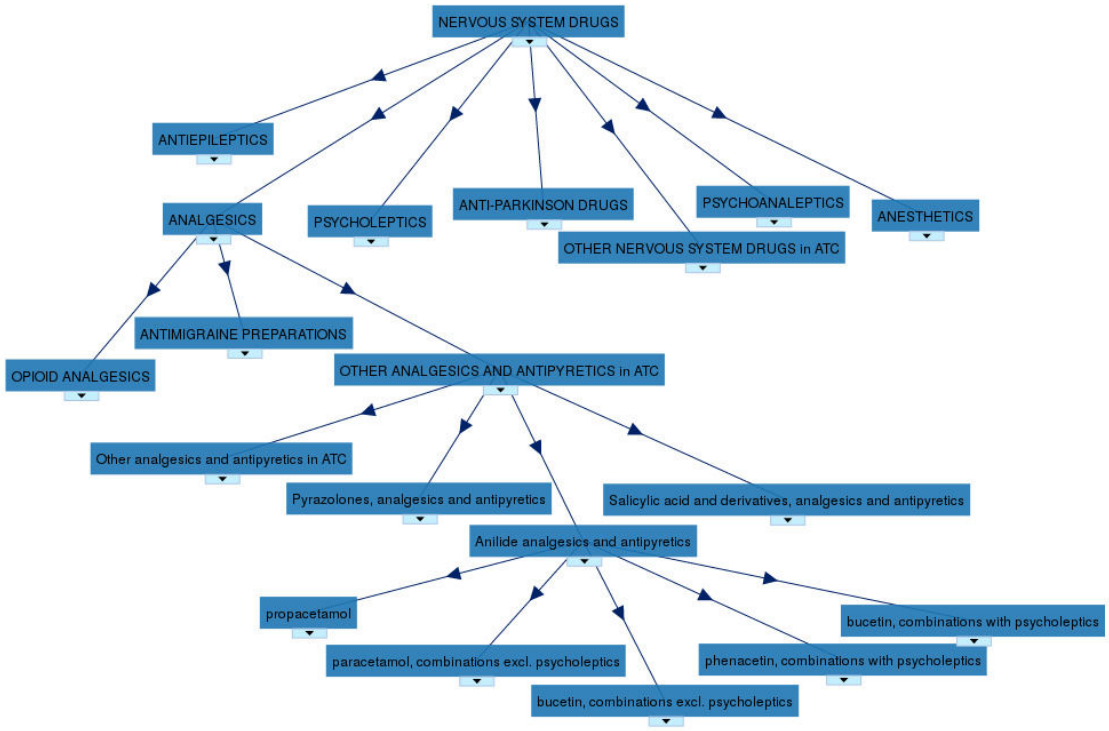

FIGURE 2.2 – Un extrait de l'ontologie de l'ATC, hébergée sur BioPortal. Les nœuds bleus correspondent aux classes de l'ontologie. Les arcs allant d'un rectangle à un autre représentent une relation de subsomption entre classes de médicaments. Par transitivité, le paracétamol est un médicament du système nerveux, dans l'ATC.

mier d'un thésaurus est l'indexation d'un document (par exemple un résumé de séjour à l'hôpital ou encore une fiche de prescription), de nombreuses ontologies se construisent aujourd'hui autour de thésaurus, notamment dans le domaine médical. On peut ici citer le thésaurus des interactions médicamenteuses de l'ANSM, qui répertorie les interactions médicamenteuses connues de l'ANSM, normalisées et codifiées selon l'ATC. Par abus de langage tous ces systèmes d'organisation de la connaissance sont généralement appelés ontologies, quand bien même certains ne décrivent pas en soi un domaine de la connaissance, mais plutôt l'organisent selon les normes d'un vocabulaire contrôlé, typiquement une nomenclature médicale dans notre cas. Cette nuance se ressent d'ailleurs de moins en moins à mesure que les ontologies se construisent en lien avec des thésaurus et taxonomies, et que les éditeurs de ces thésaurus et taxonomies les enrichissent de définitions, descriptions et relations entre concepts des domaines qu'ils couvrent.

Les formalisations de la connaissance ont à leur origine été décrites dans des formats papier. En France, le dictionnaire Vidal [11] rassemble par exemple depuis 1914 des résumés des caractéristiques de médicaments. Avec l'avènement de l'infor-

11. Le site de Vidal : https://www.vidal.fr/

matique, la formalisation des connaissances a évolué pour être exploitable par des machines.

2.2.2 Les technologies du Web Sémantique

À travers des standards et technologies. le Web Sémantique fournit un cadre technique pour supporter l'interopérabilité sémantique des connaissances et l'interopérabilité technologique de leurs formalismes.

2.2.2.1 Représentation de données et connaissances

RDF [12], pour *Ressource Description Framework*, est un formalisme de représentation de données. Ce formalisme, sous la forme de graphe, ou de réseau, décrit des données en tant que triplets, constitués d'un sujet, d'un prédicat et d'un objet, souvent écrits de la façon suivante :

$$\{sujet, prédicat, objet\}$$

Le sujet représente la ressource ou la données à décrire, la prédicat la propriété de cette donnée que le triplet va décrire, et enfin l'objet la valeur que prend le sujet pour la propriété en question. Un objet peut également être une autre donnée, c'est-à-dire un sujet dans un autre triplet.
Une hospitalisation provenant du PMSI peut alors être représentée sous la forme d'un ensemble de triplets :

$$\{:0003GZD5C27VYG8FF,\ :has\_hosp\_stay,\ :Hospitalisation\_36\}$$

$$\{:Hospitalisation\_36,\ :has\_date\_sortie,\ "2013-06"\text{^^}xsd{:}gYearMonth\}$$

$$\{:Hospitalisation\_36,\ icd10{:}has\_dp,\ icd10:I70.1\}$$

$$\{:Hospitalisation\_36,\ icd10{:}has\_das,\ icd10{:}\ I10\}$$

$$\{:Hospitalisation\_36,\ ccam{:}has\_ccam,\ ccam{:}EDAF001\}$$

$$\{:Hospitalisation\_36,\ :has\_duration,\ P2DT\text{^^}xsd{:}duration\}$$

Le patient fictif $0003GZD5C27VYG8FF$ a eu le séjour d'hospitalisation n°36, qui a duré deux jours et a pris fin durant le mois de juin en 2013. À cette hospitalisation sont associés des codes de diagnostics : le diagnostic principal et les diagnostics associés, codés selon la CIM-10. Le diagnostic principal de l'hospitalisation est ainsi codé en $I70.1$ selon la CIM-10, et représente une athérosclérose de l'artère rénale. Enfin, durant cette hospitalisation, le patient a reçu une pose d'endoprothèse, par voie artérielle transcutanée, pour une dilatation intraluminale sélective ou hypersélective de l'artère rénale. Cet acte médical est codé par $EDAF001$ selon la Classification Commune des Actes Médicaux (CCAM).
Un document RDF, donc composé de triplets, constitue alors un graphe orienté et étiqueté (figure 2.3). Les sujets et objets sont les nœuds du graphes, les prédicats sont les arcs qui relient entre eux les nœuds. Les arcs sont étiquetés par la propriété qui relie les sujets aux objets.

12. RDF sur le site du W3C : https://www.w3.org/RDF/

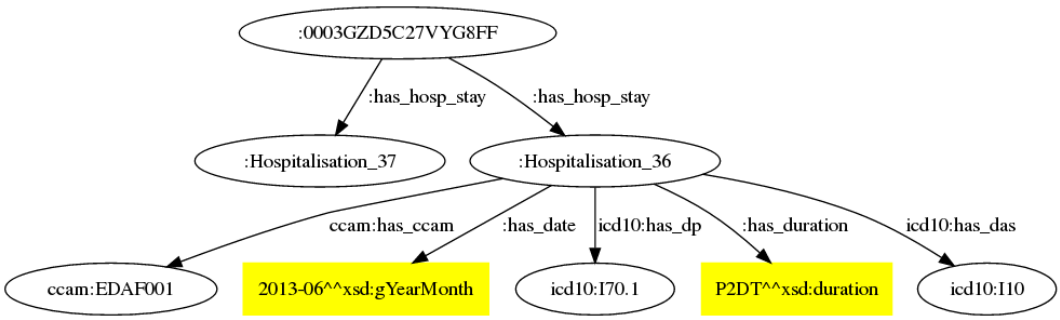

FIGURE 2.3 – Graphe RDF du patient fictif `0003GZD5C27VYG8FF` et de l'`Hospitalisation_36`.

Les connaissances d'une ontologie peuvent être formalisées selon plusieurs standards, surcouches de RDF. RDFS [13], pour RDF *schema*, est ainsi recommandé par le *World Wide Web Consortium* (W3C). Il ajoute à RDF des propriétés de base pour la représentation de classes, par exemple la propriété de subsomption entre classes avec $subClassOf$, très utilisée dans les ontologies taxonomiques (voir figure 2.2, où chaque flèche représente une relation de subsomption entre des classes de médicaments de l'ATC). OWL [14], pour *Web Ontology Language*, une extension de RDFS, introduit d'autres propriétés issues des logiques descriptives (Horrocks, 2005). Ces langages de représentation permettent ainsi d'intégrer facilement les données et les connaissances, et fournissent les primitives permettant de formaliser les connaissances afin de les réutiliser lors de raisonnements automatiques.

2.2.2.2 Interrogation de graphe RDF : SPARQL

SPARQL [15] pour *SPARQL Protocol And Query Language* est un langage de requête et un protocole. Cette technologie du Web Sémantique permet à la fois d'interroger un graphe de données RDF, d'ajouter, de supprimer ou de modifier les données d'un graphe RDF. De nombreux *triplestores*, bases de données stockant des données RDF, intègrent différentes implémentations de SPARQL [16] et permettent ainsi la récupération et la modification de triplets RDF.

Si on reprend l'exemple d'un graphe RDF qui décrit des patients et leurs hospitalisations, avec l'ontologie de la CIM-10, le code source 1 correspond à la requête SPARQL qui doit retourner les patients et caractéristiques de leurs hospitalisations (diagnostic principal, diagnostics associés, actes médicaux, date et durée) dont le diagnostic principal est issu du chapitre I70 de la CIM-10, autrement dit une athérosclérose. Ainsi, logiquement le patient 0003GZD5C27VYG8FF et l'hospitalisation

13. RDFS sur le site du W3C : https://www.w3.org/TR/rdf-schema/
14. OWL sur le site du W3C : https://www.w3.org/2001/sw/wiki/OWL
15. SPARQL sur le site du W3C : https://www.w3.org/TR/2013/REC-sparql11-overview-20130321/
16. Liste non exhaustive d'implémentations SPARQL sur le site du W3C : https://www.w3.org/wiki/SparqlImplementations

n°36 doivent se retrouver dans les résultats de cette requête.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX : <http://www.ansm.fr/>
PREFIX icd10: <http://www.semanticweb.org/owl/owlapi/turtle#>
PREFIX ccam: <http://www.ansm.fr/ccam/>
SELECT *
WHERE {
   ?patient :has_hosp_stay ?hospitalization .
   ?hospitalization icd10:has_dp ?dp .
   ?hospitalization icd10:has_das ?das .
   ?hospitalization ccam:has_ccam ?ccam .
   ?hospitalization :has_date ?date .
   ?hospitalizaiton :has_duration ?duration
   ?dp rdfs:subClassOf* icd10:I70 .
}
```

Code source 1: Requête SPARQL pour rechercher dans un graphe RDF de trajectoires de soins les patients et caractéristiques de leurs hospitalisations dont le diagnostic principal est une sous-classe de l'athérosclérose, codé en I70 dans la CIM-10.

L'apparition de ces standards et technologies, ainsi que l'informatisation des systèmes de soins, ont largement contribué au développement et à la publication des formalisations de la connaissance issue des domaines médicaux et pharmacologiques.

2.2.3 Les Données liées

Basé sur ces standards et technologies, le Web des données, ou Données Liées (*Linked Data* en anglais), est une initiative du W3C qui promouvoit la publication et le partage de données structurées et liées sur le Web. Si les données médico-administratives, personnelles et hautement confidentielles, n'ont bien sûr aucunement vocation à être partagées sur le Web –bien au contraire–, la connaissance pouvant être utile à leur réutilisation, c'est à dire la connaissance médicale et pharmacologique, formalisée sous la forme d'ontologies, est elle de plus en plus publiée et utilisée.

Le *Linked Open Data*[17] (LOD) rend ainsi libre des bases de données et des bases de connaissances sur le Web (figure 2.4). Et en plus de les rendre publics, les auteurs des données et connaissances doivent s'assurer de pouvoir les relier à d'autres bases. Comme souligné par Tim Berners-lee[18], l'idéal poursuivi par le Web des données, et qui fait sa force, est la capacité des données et connaissances à être liées entre elles. En plus de recommander les standards du Web Sémantique pour la publication libre de données et connaissances, il suggère et encourage leur interconnexion (figure 2.5).

17. *LOD cloud* : https://lod-cloud.net/
18. Données 5 étoiles : https://www.w3.org/2011/gld/wiki/5_Star_Linked_Data

Ce sont ces recommandations de bonnes pratiques, qui en partie doivent favoriser l'interopérabilité des données et connaissances.

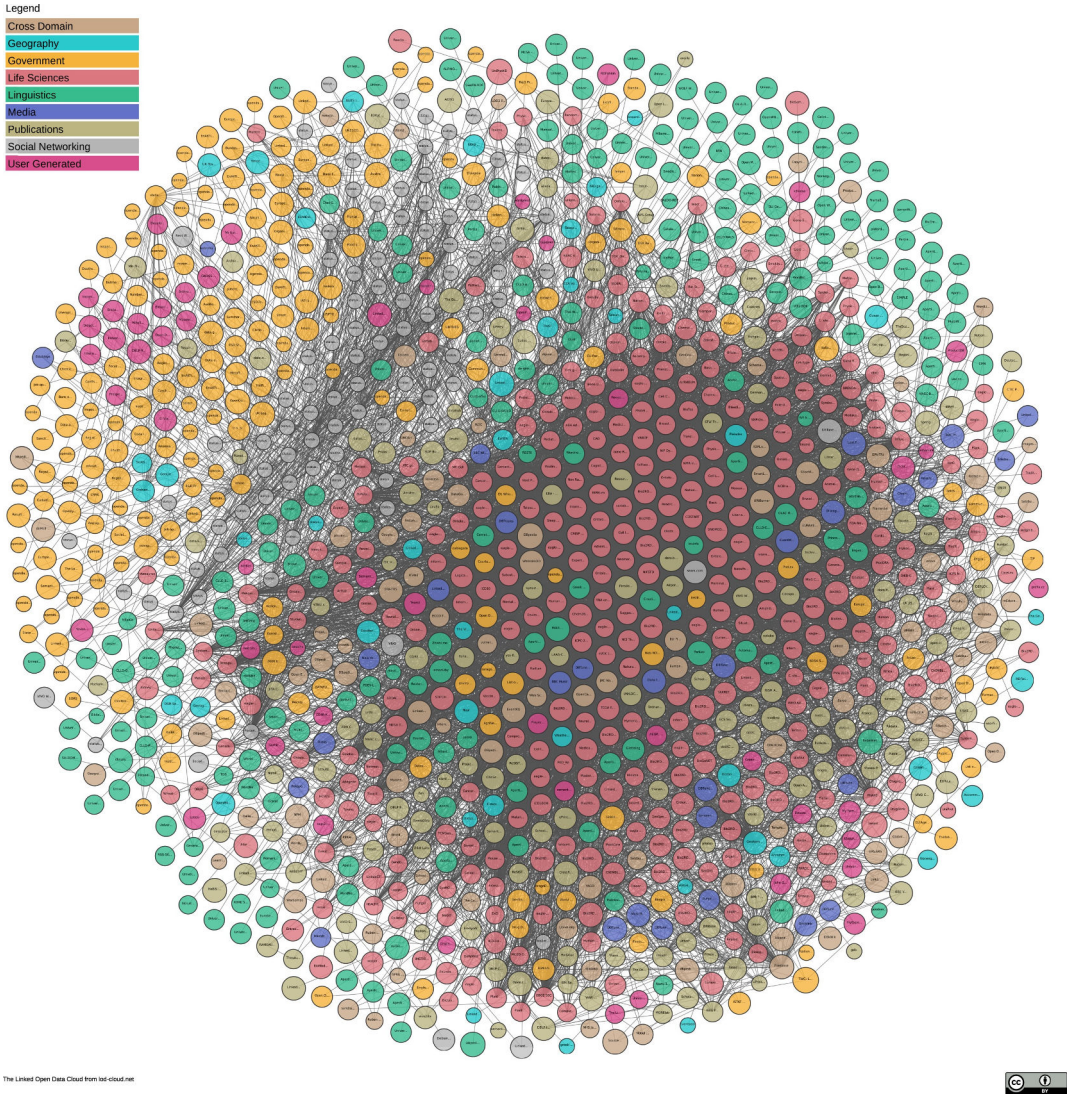

FIGURE 2.4 – *Linked Open Data Cloud* : données et connaissances aux standards du Web Sémantique publiées librement sur le Web. Les ronds représentent les publications de données liées sur le Web, et les traits l'interconnexion entre ces bases de données. Les sciences de la vie sont un des domaines apportant le plus de publication de données liées, avec ses rond rouges, que l'on retrouve principalement au centre du graphe. Ce domaine se caractérise notamment par une interconnexion des données très forte, comme peuvent le montrer les nombreux traits partant des ronds rouges. En juin 2018, le LOD contenait 1224 jeux de données et 16113 liens.

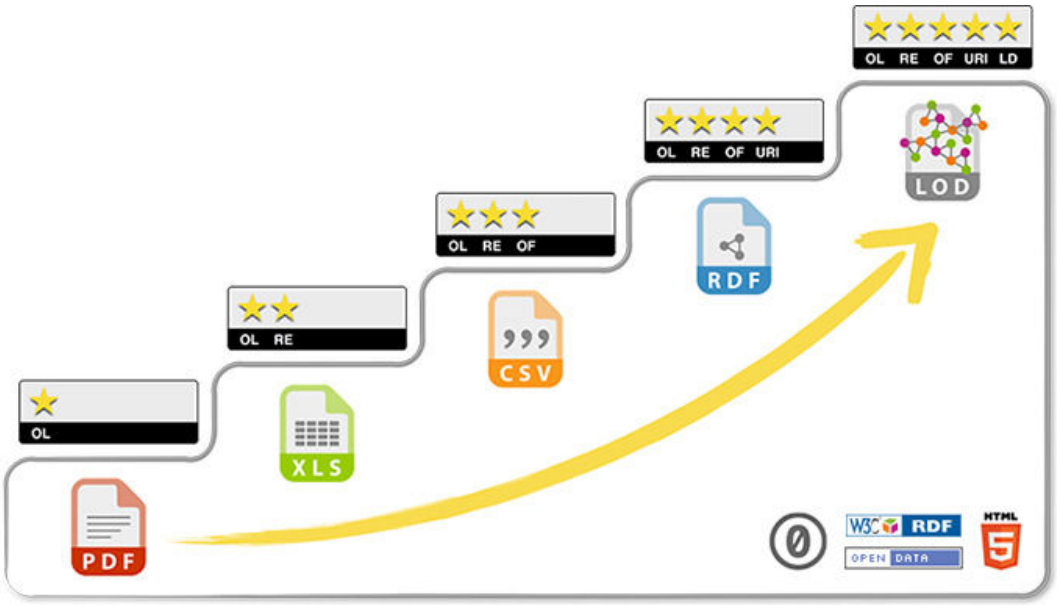

FIGURE 2.5 – *5 stars data* : vers la publication libre et l'interconnexion de données et connaissances aux standards du Web Sémantique. Les données 5 étoiles sont publiées sur le Web sous une licence libre, structurées pour être compréhensibles par l'humain comme par une machine, dans un format non propriétaire, sont notées par des identifiants uniformes de ressource(*Uniform Ressource Identifier* en anglais - URI), et sont liées à d'autres données du domaine.

La publication de données RDF et d'ontologies a alors pu prendre plusieurs formes. La multiplication des publications d'ontologies ont rapidement mené à l'arrivée de plateformes pour leur regroupement. Ainsi, dans le domaine biomédical, on peut par exemple citer BioPortal (Whetzel et al., 2011), The Open Biological and Biomedical Ontology (OBO) Foundry (Smith et al., 2007) ou encore SIFR BioPortal (Jonquet et al., 2016), un équivalent francophone de BioPortal. Ces dépôts d'ontologies ont pour but de rassembler et ainsi de publier des ontologies biomédicales (figure 2.2). En plus de pouvoir y déposer et y télécharger des ontologies, ces plateformes de dépôt mettent parfois à disposition un SPARQL *endpoint*. Ces serveurs distants hébergent les ontologies publiées sur la plateforme, et les rendent accessible par des requêtes SPARQL (voir figures 2.6 et 2.7). Sont présentés ci-après certaines des principales plateformes de dépôt et partage d'ontologies biomédicales, leurs SPARQL *endpoints*, certaines bases de connaissances biomédicales, ainsi que des nomenclatures pouvant servir à l'intégration et l'exploration de données et connaissances médicales pour la recherche en santé publique, particulièrement les données issues des bases médico-administratives françaises :

NCBO BioPortal Dirigé par le National Center for Biomedical Ontology (NCBO), BioPortal (Whetzel et al., 2011) est une plateforme de dépôt d'ontologies biomédicales. Il contient plus de 300 ontologies développées dans les formats

du Web Sémantique ou dans d'autres formats (OBO par exemple), ainsi que de nombreuses terminologies issues de la National Library of Medecine (NLM) ou de l'Organisation Mondiale de Santé (OMS) (par exemple l'ATC, voir figure 2.2). Il fournit un accès à une version RDF de ces ontologies via un SPARQL endpoint [19], en version bêta néanmoins (voir figure 2.6). BioPortal met également à disposition des outils pour travailler avec les ontologies de la plateforme, via des API REST. L'API *search* [20] et *annotator* [21] permettent par exemple de récupérer de l'information sur les ontologies de BioPortal et respectivement d'annoter du texte par des classes de ces ontologies (Salvadores et al., 2012).

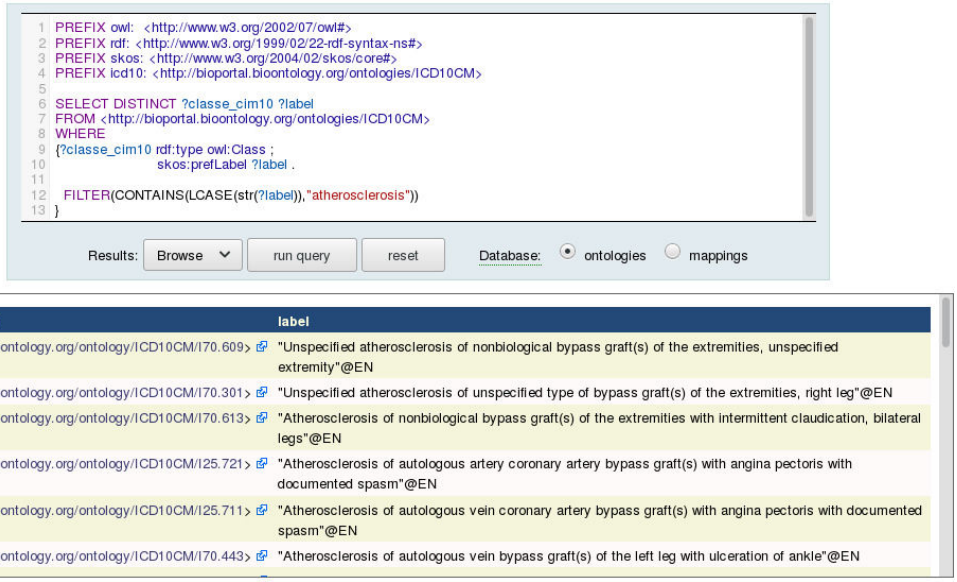

FIGURE 2.6 – Requête SPARQL sur le serveur de BioPortal : recherche des classes de la CIM-10 dont le label contient "atherosclerosis".

SIFR BioPortal En collaboration avec le NCBO, le projet SIFR [22] (*Semantic Indexing of French Biomedical Data Resources*, Indexation Sémantique de Ressources Biomédicales Francophones, en français) vise à développer une plateforme similaire à celle de BioPortal, pour les ontologies biomédicales francophones (Jonquet et al., 2009). Elle met ainsi à disposition des ontologies et nomenclatures biomédicales traduites ou même initialement conçues en français. Certaines ontologies sont ainsi extraites et traduites de l'*Unified Medical Language System* (UMLS - paragraphe 2.2.3), récupérées du NCBO BioPortal, déposées par les utilisateurs ou également fournies par le Catalogue de Index des Sites Médicaux de langue Française [23] (CISMEF) depuis la plateforme HeTOP (Grosjean et al., 2011). La plateforme compte

19. SPARQL endpoint de BioPortal : `http://sparql.bioontology.org/`
20. API search de BioPortal : `https://bioportal.bioontology.org/search`
21. API annotator de BioPortal : `https://bioportal.bioontology.org/annotator`
22. Projet SIFR : `http://sifr.lirmm.fr/`
23. Le site du CISMEF : `http://www.chu-rouen.fr/cismef/`

près de 30 ontologies francophones. Tout comme le NCBO Bioportal, SIFR BioPortal fournit un SPARQL endpoint [24], et des API REST [25, 26].

OBO Foundry OBO Foundry [27], pour *The Open Biological and Biomedical Ontology* est un collectif de développeurs d'ontologies biomédicales (Smith et al., 2007). Le collectif a pour objectif le développement d'ontologies biomédicales de référence, libres et interopérables.

Ontobee Ontobee est un SPARQL endpoint rendant accessibles des ontologies biomédicales aux standards du Web Sémantique, requêtable en SPARQL (Ong et al., 2017). Son serveur [28] héberge la plupart des ontologies issues d'OBO Foundry ainsi que d'autres ontologies biomédicales, rendant leurs connaissance accessible via des requêtes SPARQL (voir figure 2.7). Il offre tout comme BioPortal un outil pour l'annotation automatique [29] de texte brut par les classes de ses ontologies.

Bio2RDF Bio2RDF est un projet *open-source* utilisant les technologies du Web Sémantique pour la production d'un large réseaux de données liées et libres dans le domaine des sciences de la vie (Callahan et al., 2013). Plus de 11 milliards de triplets RDF, issues de 35 bases de données, sont ainsi requêtables sur un SPARQL endpoint [30]. Bio2RDF contient par exemple une partie de la base de connaissance DrugBank (Wishart et al., 2008), relative aux médicaments et à leurs effets, sous la forme de triplets RDF.

DBpedia DBpedia est un projet dont l'objectif est l'extraction automatique de données liées et structurées aux formats du Web Sémantique à partir de contenus de l'encyclopédie universelle et multilingue Wikipédia (Lehmann et al., 2015). Sa portée pluridisciplinaire couvre entre autre les domaines médicaux et pharmacologiques. Son SPARQL endpoint [31] permet ainsi de récupérer grâce à des requêtes SPARQL des connaissances relatives à des médicaments, actes médicaux et maladies.

UMLS Maintenu par la NLM, l'*Unified Medical Language System* (UMLS) est un condensé de plusieurs vocabulaires contrôlés, nomenclatures, du domaine médical (Bodenreider, 2004). Il fournit une correspondance entre les termes de ces différentes nomenclatures, via les *Concept Unique Identifier* (CUI). De par toutes les connaissances décrivant les termes de ces nomenclatures et de ces correspondances, il est souvent vu comme un réseau d'ontologies biomédicales.

24. SPARQL endpoint du SIFR BioPortal : `http://sparql.bioportal.lirmm.fr/test/`
25. API search de SIFR BioPortal : `http://bioportal.lirmm.fr/search`
26. API Annotator de SIFR BioPortal : `http://bioportal.lirmm.fr/annotator`
27. OBO Foundry : `http://www.obofoundry.org/`
28. SPARQL endpoint d'Ontobee : `http://www.ontobee.org/sparql`
29. API Annotator d'Ontobee : `http://www.ontobee.org/annotate`
30. SPARQL endpoint de Bio2RDF : `http://bio2rdf.org/sparql`
31. SPARQL endpoint de DBpedia : `https://dbpedia.org/sparql`

NDF-RT Produite par le Département des Anciens combattants des États Unis d'Amérique (*Department of Veterans Affairs*) et incluse dans l'UMLS, la *National Drug File - Reference Terminology* est une base de connaissances relatives aux médicaments de la liste *Veterans Health Administration National Drug File*. Elle décrit ainsi les effets physiologiques, les indications (pouvant traiter ou prévenir une maladie ou un état de santé), les compositions chimiques, les interactions ou encore les contre-indications des médicaments présents dans cette liste.

-- Prefixes -- | -- Template -- | -- Statement Help -- | Example 1, Ex 2, Ex 3, Ex 4, Ex 5, Ex 6, Ex 7, Ex 8

```
prefix ndf: <http://evs.nci.nih.gov/ftp1/NDF-RT/NDF-RT.owl#>
prefix umls: <http://bioportal.bioontology.org/ontologies/umls/>
 SELECT DISTINCT ?ndf_drug ?cui_drug ?label_drug ?ndf_diag ?cui_diag ?label_diag
 FROM <http://evs.nci.nih.gov/ftp1/NDF-RT/NDF-RT.owl>
 WHERE {
 {?ndf_drug ndf:UMLS_CUI ?cui_drug .} UNION {?ndf_drug_low rdfs:subClassOf ?ndf_drug .
                                             ?ndf_drug_low ndf:UMLS_CUI ?cui_drug .}
 ?ndf_drug rdfs:subClassOf ?CI .
 ?CI owl:onProperty ndf:CI_with ;
     owl:someValuesFrom ?ndf_diag .
 ?ndf_diag ndf:UMLS_CUI ?cui_diag .
 OPTIONAL{?ndf_diag rdfs:label ?label_diag .}
 OPTIONAL{?ndf_drug rdfs:label ?label_drug .}}
```

Output format: Table | Max Rows: 10 | Run Query | Reset

Result | Raw Request/Permalinks | Raw Response

ndf_drug	cui_drug	label_drug	ndf_diag	cui_diag	label_diag
http://evs.nci.nih.gov/ftp1/NDF-RT/NDF-RT.owl#N0000020091	"C0014704"	"ERGONOVINE"	http://evs.nci.nih.gov/ftp1/NDF-RT/NDF-RT.owl#N0000000273	"C0000821"	"Abortion, Threatened [Disease/Finding]"
http://evs.nci.nih.gov/ftp1/NDF-RT/NDF-RT.owl#N0000145814	"C0059514"	"ERGONOVINE MALEATE"	http://evs.nci.nih.gov/ftp1/NDF-RT/NDF-RT.owl#N0000000273	"C0000821"	"Abortion, Threatened [Disease/Finding]"
http://evs.nci.nih.gov/ftp1/NDF-RT/NDF-RT.owl#N0000023156	"C1572765"	"WARFARIN SODIUM ISOPROPANOL COMPLEX"	http://evs.nci.nih.gov/ftp1/NDF-RT/NDF-RT.owl#N0000000273	"C0000821"	"Abortion, Threatened [Disease/Finding]"
http://evs.nci.nih.gov/ftp1/NDF-RT/NDF-RT.owl#N0000022035	"C0244656"	"FOSPHENYTOIN"	http://evs.nci.nih.gov/ftp1/NDF-RT/NDF-RT.owl#N0000000304	"C0001396"	"Adams-Stokes Syndrome [Disease/Finding]"
http://evs.nci.nih.gov/ftp1/NDF-RT/NDF-RT.owl#N0000022099	"C0733758"	"FOLLITROPIN"	http://evs.nci.nih.gov/ftp1/NDF-RT/NDF-RT.owl#N0000000324	"C0001621"	"Adrenal Gland Diseases [Disease/Finding]"
http://evs.nci.nih.gov/ftp1/NDF-RT/NDF-RT.owl#N0000145817	"C0012258"	"DIGITOXIN"	http://evs.nci.nih.gov/ftp1/NDF-RT/NDF-RT.owl#N0000000375	"C0002726"	"Amyloidosis [Disease/Finding]"

FIGURE 2.7 – Requête SPARQL sur le serveur d'Ontobee : recherche des contre-indications entre médicaments et diagnostics dans l'ontologie NDF-RT. Le résultat est une liste de couples (médicament, diagnostic) où le médicament est contre-indiqué avec le diagnostic. Par exemple, l'ergonovine est contre-indiquée avec un avortement ou une menace d'avortement, selon NDF-RT. La requête retourne les codes NDF-RT des médicaments et diagnostics, leurs labels, ainsi que leurs correspondances CUI.

DIKB Le projet *Drug Interaction Knowledge Base* est un programme de recherche visant à produire des outils d'aide à la décision clinique basés sur des connaissances relatives aux interactions médicamenteuses. Il propose ainsi un portail [32] pour chercher des interactions entre médicaments, à partir d'un regroupement de plus d'une dizaine de sources, dont DrugBank, NDF-RT ou encore le thésaurus des interactions médicamenteuses de l'ANSM. Pour des applications de recherche, le projet rend également disponible la partie libre des toutes ces sources. Les médicaments y sont codés par les codes de médicament de DrugBank.

32. Le portail merged-PDDI : `https://www.dikb.org/Merged-PDDI/`

DID DID (Sharp, 2017), pour *Drug Indication Database*, est une base de données regroupant douze sources de connaissances sur les indications de médicaments. Les médicaments de ces sources sont codés en codes CUI, grâce à des correspondances entre les nomenclatures utilisées par chacune des sources et les CUI de l'UMLS.

CIM-10 La 10^{e} révision de la Classification Internationale des Maladies, maintenue par l'OMS, est une classification codant les maladies, signes, symptômes, causes externes de maladies ou de blessures. Elle est notamment utilisée dans le PMSI pour coder les diagnostics réalisés à l'hôpital. Des versions en RDF sont disponibles en anglais sur NCBO BioPortal et en français sur SIFR BioPortal.

ATC La classification Anatomique Thérapeutique et Chimique est une classification détenue par l'OMS pour classer les médicaments. Sa hiérarchie en cinq niveaux définit (i) le groupe anatomique d'un médicament, (ii) son sous-groupe pharmacologique ou thérapeutique principal, (iii et iv) ses sous-groupes chimiques, pharmacologiques ou thérapeutiques et (iv) sa substance chimique. Des versions en RDF sont disponibles en anglais sur NCBO BioPortal (voir figure 2.2) et en français sur SIFR BioPortal.

Le Linked Data met ainsi à disposition de nombreux outils et ontologies relatifs au domaine pharmacologique et médical. Utiliser ces outils et ontologies constitue un réel défi au regard de leur variété, de leur dispersion sur le Web et de leur spécificité. Ils peuvent cependant s'adosser aux bases de données médico-administratives et ainsi enrichir leur traitement, de l'exploration à l'analyse des données.

2.2.4 Le Web Sémantique et les Données liées pour la recherche en santé publique

La recherche en santé publique dépend en grande partie des données médicales disponibles, et donc des enjeux propres à leur traitement (Piro et al., 2016). L'intégration, l'analyse et l'exploration de ces données représentent particulièrement des obstacles techniques (Girardi et al., 2015). Si certaines complexités des données médicales sont souvent pointées du doigt, par exemple leur dispersion et leur hétérogénéité, de plus en plus d'articles et projets de recherche en santé publique soulignent la pertinence de l'utilisation des technologies du Web Sémantique pour y remédier (Piro et al., 2016; Ferreira et al., 2013). Parallèlement, la réalisation de tâches complexes -analyse ou exploration- sur de telles données de plus en plus massives et complexes, nécessite la gestion et l'utilisation de connaissances médicales externes aux données. Un nombre croissant de projets de recherche trouvent aussi dans les technologies du Web Sémantique et les ontologies médicales des Données liées une solution à ces besoins (Ivanović and Budimac, 2014). La représentation des données de santé selon les technologies du Web Sémantique s'est donc de plus en plus répandue ces dernières années. Pathak et al. (2012a); Piro et al. (2016) ont ainsi représenté des données de santé de patients en RDF. L'utilisation conjointe

des technologies du Web Sémantique pour la représentation des données et d'ontologies du Linked Data permet alors de réaliser des tâches complexes, nécessitant des connaissances expertes. Pathak et al. (2013) ont par exemple pu rechercher des potentielles interactions médicamenteuses dans des données de santé contenant des prescriptions, en utilisant la bases de connaissance DrugBank. Ils ont également basé leur étape de construction de cohortes sur l'utilisation du Web Sémantique et de plusieurs ontologies médicales et pharmacologiques (Pathak et al., 2012b).

Si les technologies du Web Sémantique et les ontologies médicales sont de plus en plus utilisées en médecine, elles restent relativement peu utilisées en santé publique, et particulièrement en épidémiologie (Ferreira et al., 2013). Bien que pouvant apporter des solutions aux complexités et enjeux liés au traitement de ces données, leur diffusion est ralentie par certains enjeux techniques. Comme mentionné par Jain et al. (2010), et plus spécifiquement dans le domaine médical par Jonquet et al. (2009), le nombre croissant des ontologies, l'hétérogénéité de leur schéma de représentation des ontologies, leur diversité et un manque de description rend leur utilisation compliquée. C'est particulièrement le cas dans l'élaboration de requêtes SPARQL, qui reposent sur ces schémas.

En France, la représentation de données médico-administratives selon les technologies du Web Sémantique, couplé à l'utilisation d'ontologies médicales n'a encore que rarement été menée.

2.3 Analyser les trajectoires de soins issues des bases de données médico-administratives françaises

Comme nous l'avons vu dans la section 2.1.5, les trajectoires de soins de patients peuvent être reconstituées à partir de données issues des bases médico-administratives françaises. Également, leur traitement permet la compréhension du contexte de soins, et donc aide à l'amélioration du système de santé et de la qualité des soins. Ce traitement peut prendre plusieurs formes. La comparaison de trajectoires de soins entre elles peut par exemple permettre de mesurer la ressemblance entre deux patients, du point de vue de leurs consommations et recours aux soins. De la même façon, la comparaison d'une trajectoire de soins à un parcours théorique, permet de situer un patient par rapport à des bonnes pratiques et recommandations associées à son état de santé (Ainsworth and Buchan, 2012; Li et al., 2015). La constitution de groupes homogènes de patients selon leur trajectoire de soins permet de discerner des tendances, comme des séquences de consommation et de recours aux soins partagées au sein d'une population. La découverte d'associations fréquentes entre des événements qui constituent les trajectoires permet également de comprendre le contexte de consommation et de recours aux soins d'une population. Les méthodes employées pour poursuivre ces objectifs sont assez diverses, souvent empruntées à l'informatique avec notamment de la fouille de données et de la comparaison de chaînes de caractère, ainsi qu'à la statistique avec des méthodes de classification.

2.3.1 Comparaison de trajectoires de soins

Dans leur formalisation, les trajectoires de soins sont généralement représentées comme des séquences d'événements médicaux. Des méthodes de comparaison de chaînes de caractères (*string metrics* en anglais) peuvent alors s'adapter à la comparaison de trajectoires de soins, où les événements médicaux remplacerait les caractères. Des mesures de similarités basées sur ce type de comparaison ont alors été utilisées pour mesurer la ressemblance entre trajectoires de soins au sein d'une population (Williams et al., 2014). Ces mesures de similarité, autant que les méthodes de comparaison de chaînes, sont d'ailleurs très nombreuses et variées (Studer and Ritschard, 2016). Les distances d'édition font partie des méthodes de comparaison de chaînes de caractères les plus populaires pour la comparaison de trajectoires de soins (Le Meur et al., 2015; Roux et al., 2018b). Elles peuvent se définir de la sorte :

Définition 1 *distance d'édition*
Une distance d'édition entre deux séquences est proportionnelle au nombre minimum d'éditions –c'est à dire d'un changement d'un caractère, dans notre cas d'un événement– nécessaires pour passer d'une séquence à une autre.

De cette définition très généraliste, dépendent plusieurs variantes, basés sur différents types de changements de caractères, ou encore sur différentes pondérations

associés à ces différents changements (Navarro, 2001). Par exemple, la première distance de ce genre, la distance de Levenshtein (1966), comprenait la substitution, l'insertion et la délétion. La distance de Damerau-Leveinshtein introduit ensuite la transposition entre éléments (Damerau, 1964). Celle de Jaro (1989) n'utilise justement que la transposition, celle de Hamming (1950) uniquement la substitution. Le principe de la plus grande sous séquence commune (*Longest Common Subsequence* - LCS, en anglais) considère uniquement l'insertion et la délétion (Hirschberg, 1975). La LCS permettant d'introduire la notion de sous-séquence, sa définition est détaillée ci-après :

Définition 2 *Sous-séquence*
Soient deux séquences $X = (x_1, x_2, ..., x_m)$ *et* $Y = (y_1, y_2, ..., y_n)$*, avec* $m \leq n$*.* X *est une sous-séquence de* Y *s'il existe les indices* $1 \leq j_1 \leq j_2 \leq ... \leq j_m \leq n$ *tels que* $x_i = y_{j_m}$ *pour tout* $i = 1, 2, ..., m$*.*

Définition 3 *Plus grande sous-séquence commune*
Étant données deux séquences $X = (x_1, x_2, ..., x_m)$ *et* $Y = (y_1, y_2, ..., y_n)$*,* Z *est une plus grande sous séquence commune de* X *et* Y *si* Z *est une sous-séquence de* X *et de* Y*, et que* $|Z| \geq |Z'|$*, pour toute autre sous-séquence commune* Z' *de* X *et* Y*.*

Si on considère deux trajectoires de soins, écrites en séquences, par exemple les trajectoires suivantes composées de codes de médicaments de l'ATC et d'un code de diagnostic CIM-10, $T_1 = (N02BE01, A02BC05, M17.1)$ et $T_2 = (B01AC04, A02BC05, M17.1)$, la plus grande sous séquence commune à T_1 et T_2 est $(A02BC05, M17.1)$.
Pour obtenir une mesure de similarité, il est d'usage de rapporter une telle distance à une valeur maximale que l'on pourrait obtenir (Zhao et al., 2002).

Définition 4 *Mesure de similarité basée sur la LCS*
La mesure de similarité entre deux séquences X *et* Y *se définit alors de la façon suivante :*

$$sim(X, Y) = \frac{LCS(X, Y)}{min(|X|, |Y|)}$$

Cette mesure de similarité ne prenant pas en compte la potentielle différence de taille entre les deux séquences, une variante est parfois préférée, rapportant la distance à la taille maximale des deux séquences.

Définition 5 *Mesure de similarité basée sur la LCS - variante*
La mesure de similarité entre deux séquences X *et* Y *se définit alors de la façon suivante :*

$$sim(X, Y) = \frac{LCS(X, Y)}{max(|X|, |Y|)}$$

Sur la base de telles mesures de similarité, des méthodes de classification statistique ont souvent été appliquées dans le but de créer des groupes homogènes de

patients, selon leurs consommations de soins et états de santé, parmi un ensemble plus général de patients. (Le Meur et al., 2015) ont par exemple utilisé une telle méthodologie pour l'étude du suivi médical des femmes enceintes en France, afin de classer les patientes au sein de trois groupes relatifs au degré de leur consommation de soins. De telles études mènent ensuite à des analyses statistiques pour la caractérisation des groupes, aussi bien pour leurs description et compréhension que pour constater si les effets de groupes, et donc les différences de trajectoires, sont déterminants d'un état de santé ou d'un résultat clinique. Ces mesures de similarités peuvent également permettre de comparer des trajectoires de soins à des parcours de recommandations, comme le montrent (Williams et al., 2014).

2.3.2 Motifs à partir de trajectoires de soins

D'autres méthodes de traitement et d'exploration de données ont été utilisées pour l'analyse des trajectoires de soins, avec notamment des méthodes issues de la fouille de données (Egho et al., 2013a). Par la vision des trajectoires de soins en séquences composées d'événements de santé, de telles méthodes permettent de rechercher des motifs fréquents dans un ensemble de trajectoires, pour identifier des tendances de consommation et de recours aux soins.

2.3.2.1 Motifs fréquents

Dans la recherche de motifs fréquents, les bases de données constituées de trajectoires de soins sont appelées des bases de transactions.

Id transaction	Séquence
T_1	(atc:A02BC05, atc:N02BE01, icd10:M16.1)
T_2	(atc:N02BE01, atc:A02BC05, icd10:M17.1)
T_3	(atc:B01AC04, atc:A02BC01, icd10:M17.1)
T_4	(atc:B01AB05, atc:N02BE01, icd10:M16.1)
T_5	(atc:B01AC04, atc:A02BC05, icd10:M17.1)
T_6	(atc:A02BC05, icd10:M17.1)

TABLE 2.1 – Exemple fictif d'une base de données de transactions.

Un motif est de la même forme qu'une trajectoire. C'est donc une séquence d'éléments. $(atc : A02BC05, atc : N02BE01)$ est par exemple un motif. Chaque motif possède un support, définit de la façon suivante :

Définition 6 *Support*
Étant donné une bases de transactions $B = t_1, t_2, ..., t_n$, le support d'un motif m dans B, noté $support_B(m)$ est la proportion de transactions dont le motif est une sous-séquence. On dit aussi que c'est la proportion de transactions qui vérifient le motif.

Le motif $(atc : A02BC05, icd10 : M17.1)$ a ainsi un support de 1/2 dans la base de transaction 2.1. Un motif est fréquent selon un seuil minimal de support. Les motifs fréquents se définissent alors comme suit :

Définition 7 *Motif fréquent*
Étant donné une base de transactions B et un seuil minimal de support s, un motif M est dit fréquent si $support_B(M) \geq s$.

Si on considérait un seuil de support minimal inférieur ou égale à 1/2, le motif $(atc : A02BC05, icd10 : M17.1)$ serait alors fréquent dans la base de transaction 2.1.

L'extraction de motifs fréquents à partir de trajectoires de soins a par exemple été utilisée par Pinaire et al. (2017a) pour l'identification de motifs les plus pronostics de décès hospitalier, dans le cadre de la prise en charge de l'infarctus du myocarde, en utilisant des données issues du PMSI. Dans ce travail, les auteurs utilisent également des mesures de similarité basées sur des distances d'édition pour mesurer l'éloignement de trajectoires de soins à ces motifs les plus pronostics de décès hospitalier.

2.3.2.2 Règles d'association

L'extraction de règles d'association est une méthode de fouille de données très proche de celle de l'extraction de motifs fréquents. Elle peut en effet être vue comme un cas particulier de cette dernière, où les motifs extraits sont des séquences de deux ensembles. L'utilisation des règles d'associations, et d'algorithmes pour leur extraction, s'est d'ailleurs fortement popularisée grâce à un article d'Agrawal et al. (1993), dont certains des auteurs ont ensuite contribué aux méthodes d'extraction de motifs fréquents (Agrawal and Srikant, 1995). Dans cet article, les auteurs définissent les règles d'association de la façon suivante :

Définition 8 *Règle d'association*
Soit $I = i_1, i_2, ..., i_m$ un ensemble d'objets. Soit B un ensemble de transactions où chaque transaction T est un ensemble d'objets provenant de I tels que $T \subseteq I$. Une règle d'association est une implication de la forme $X \longrightarrow Y$, où $X \subset I, Y \subset I$, et $X \cap Y = \emptyset$. X est alors généralement appelé l'antécédent (ou côté gauche, left hand side en anglais), quand Y est appelé le conséquent (ou côté droit, right hand side en anglais).

Pour une règle, donc pour une implication entre un antécédent et un conséquent, on peut calculer tout un ensemble de mesures de qualité. Le support et la confiance ont été les premières mesures à caractériser des règles d'association (Agrawal et al., 1993). De la même façon que pour les motifs, le support d'une règle est la part de transactions $T \in D$ telle que T satisfait la règle. Il peut se définir en termes de probabilités, de la façon suivante :

Définition 9 *Support d'une règle d'association*

Si on considère la règle $R : X \longrightarrow Y$ *et* B *une base de transactions :*

$$support(R) = P(XY)$$

c'est à dire, la probabilité qu'une transaction de B *contienne à la fois* X *et* Y*, qu'elle satisfasse la règle autrement dit.*

Avec l'exemple de transactions (table 2.1), on peut ainsi mesurer le support de la règle R_1 : $atc{:}A02BC05 \longrightarrow icd10{:}M17.1$:

$$support(R_1) = P(atc{:}A02BC05, icd10{:}M17.1) = 3/6 = 0,5$$

Trois transactions sur six satisfont la règle R_1.

La confiance d'une règle représente la part de transactions à satisfaire le conséquent parmi celle qui satisfont déjà l'antécédent. Elle reflète le caractère prédictif de la règle, la valeur prédictive de l'antécédent sur le conséquent, de X sur Y. En gardant cette écriture probabiliste, elle peut être définie de la sorte :

Définition 10 *Confiance d'une règle d'association*

$$confiance(R) = P(Y|X)$$

Et si on reprend à nouveau l'exemple :

$$confiance(R_1) = P(Y|X) = P(icd10{:}M17.1|atc{:}A02BC05) = 3/4 = 0,75$$

Parmi les quatre transactions qui contiennent $icd10{:}M17.1$, trois contiennent également $atc{:}A02BC05$.

Après l'article d'Agrawal et al. (1993), et avec l'engouement de la méthode, de nombreuses autres mesures de qualité des règles (table 2.2) ont été publiées et utilisées (Guillet and Hamilton, 2007). Notamment, les mesures de qualité dites objectives, c'est à dire ne nécessitant pour leur calcul uniquement les transactions ayant servi à l'extraction des règles, sont très largement étudiées.

Dans l'utilisation de données médicales, Concaro et al. (2009) ont par exemple utilisé des méthodes d'extraction de règles d'associations pour caractériser des groupes de patients, en soulignant des associations fréquentes liées à ces patients, entre des diagnostics et des prescriptions de médicaments.

Mesure (en anglais)	Formule
$Support$	$P(XY)$
$Confidence$	$P(Y\|X)$
$Coverage$	$P(X)$
$Prevalence$	$P(B)$
$Recal$	$P(X\|Y)$
$Specificity$	$P(\neg Y\|\neg X)$
$Accuracy$	$P(XY) + P(\neg X\neg Y)$
$Lift$ ou $Interest$	$\frac{P(Y\|X)}{P(Y)}$
$Leverage$	$P(Y\|X) - P(X)P(Y)$
$AddedValue/ChangeofSupport$	$P(Y\|X) - P(Y)$
$RelativeRisk$	$\frac{P(Y\|X)}{P(Y\|\neg X)}$

TABLE 2.2 – Mesures objectives de qualité de règles, tirées de (Geng and Hamilton, 2007).

2.3.2.3 Chroniques

L'extraction de chroniques fréquentes peut être vue comme un cas particulier de l'extraction de motifs fréquents. Les chroniques ajoutent en effet au motif des contraintes temporelles entre les éléments du motif. Les base de transactions pour l'extraction de chroniques fréquentes doit donc être pourvue de cet aspect temporel (voir figure 2.3).

Id transaction	Séquence
T_1	(atc:A02BC05, 1), (atc:N02BE01, 3), (icd10:M16.1, 6)
T_2	(atc:N02BE01, 2), (atc:A02BC05, 4), (icd10:M17.1, 7)
T_3	(atc:B01AC04, 1), (atc:A02BC01, 5), (icd10:M17.1, 9)
T_4	(atc:B01AB05, 3), (atc:N02BE01, 6), (icd10:M16.1, 8))
T_5	(atc:B01AC04, 2), (atc:A02BC05, 5), (icd10:M17.1, 10)
T_6	(atc:A02BC05, 4), (icd10:M17.1, 5)

TABLE 2.3 – Exemple d'une base de données fictive de transactions temporelles.

Pour définir les chroniques, il faut donc définir tout d'abord les contraintes temporelles entre éléments.

Définition 11 *Contraintes temporelles*
Une contrainte temporelle entre deux événements e_1 et e_2 est définie par l'ensemble (e_1, e_2, t^-, t^+). Elle peut aussi être notée $e_1[t^-, t^+]e_2$, où $t^- \leq t^+$. Un couple d'événements $((e,t),(e',t'))$ satisfait la contrainte temporelle (e_1, e_2, t^-, t^+) si et seulement si $e = e_1,\ e' = e_2,\ et\ (t'-t) \in [t^-, t^+]$.

Définition 12 *Chroniques*
Une chronique est un couple $(\mathcal{E}, \mathcal{T})$ *où* $\mathcal{E} = \{e_1, ..., e_n\}$ *est un ensemble fini d'événements partiellement ordonnés, et* $\mathcal{T} = \{\tau_{ij}\}_{i \leq i < j \leq n}$ *est un ensemble de contraintes temporelles sur* $\mathcal{E}$*.*

Comme pour l'extraction de motifs fréquents, une chronique est dite fréquente selon un seuil minimal de support. Le support d'une chronique est également la proportion de transactions qui vérifient la chronique parmi une base.
Par exemple, la chronique $atc{:}A02BC05[1,3]icd10{:}M17.1$ a un support de $2/6$. Si on considérait un support minimal inférieur ou égal à $1/3$, elle serait alors une chronique fréquente.

Avec une application aux bases de données médico-administratives, Dauxais et al. (2017) ont par exemple extrait des chroniques de prescriptions de médicaments, discriminantes pour des événements correspondant à des crises d'épilepsie. Une telle étude peut ensuite permettre de surveiller les patients dont la trajectoire de soin se rapproche ou bien même vérifie ces chroniques.

2.3.3 Limites

Si des méthodes d'analyse et de fouille de données sont de plus en plus utilisées dans le cadre de la réutilisation des bases de données médico-administratives françaises, leur efficacité est parfois limitée.

Extraction de motifs Les contributions dans la recherche de motifs fréquents, de règles d'associations ou de chroniques, s'accordent largement sur le fait que l'intérêt de ces méthodes est limité par la génération de résultats trop volumineux. Ces résultats sont de fait compliqués à analyser. Différentes méthodes pour filtrer les résultats existent, et celles basées sur des mesures d'intérêt des motifs sont sans doute les plus populaires. Ces mesures statistiques mènent généralement à filtrer les motifs rares, qui peuvent toutefois être intéressants.

Comparaison de trajectoires Les méthodes utilisées sont plus adaptées à des séquences courtes. Elles ont été utilisées en premier lieu pour la comparaison de chaînes de caractères, et le nombre d'événements médicaux dépasse bien souvent la taille de l'alphabet. Des pré-traitements pour réduire la taille des séquences et l'alphabet des événements étudiés sont alors souvent nécessaires.

D'autres limites à ces méthodes sont directement liées à certaines caractéristiques des données médico-administratives.

Variabilité des données et profondeur des nomenclatures médicales Dans les distances d'édition et autres méthodes de comparaison de chaînes de caractères, la variabilité des trajectoires couplée à la profondeur des nomenclatures médicales utilisées mène souvent à des mesures de similarité faibles, proches de 0, et à peu de fortes mesures, proches de 1. Ceci est du à la comparaison stricte des éléments de

deux trajectoires. Dans la plupart de ces méthodes, si deux événements médicaux de deux trajectoires comparées, ne sont pas strictement les mêmes, au moins une édition sera nécessaire. Le nombre d'édition peut ainsi se retrouver très grand, amenant à une similarité très faible. Pour la même raison, dans les méthodes de fouille de données, une grande variabilité des éléments constituant les transactions va mener à l'extraction de nombreux motifs, mais qui ne seront que peu fréquents. Dans les deux cas, des pré-traitements sont là aussi d'usage pour contourner ces limites, avec par exemple la considération d'un niveau de granularité des données plus général.

Ce genre de méthodes peut aussi être impacté par l'aspect massif des données.

Volumétrie des données Une comparaison de trajectoires de soin dans un ensemble de patients implique une comparaison deux à deux de leurs trajectoires. Réaliser ce genre de calcul gourmand est parfois impossible dans un temps convenable sur une grande population, par exemple pour un million de patients. Il en va de même pour les méthodes de fouille de données. Face à cette limite, il est aussi d'usage de réduire la taille des trajectoires et transactions par la sélection d'un nombre restreint d'événements de santé d'intérêt.

Les trajectoires, plus courtes, sont alors plus rapidement comparées, et on peut utiliser ces méthodes sur des plus grands ensembles de patients. La réduction de la taille des transactions réduit le nombre de motifs obtenus, et facilite leur analyse. Cependant, la réduction des trajectoires et transactions demande d'établir des *apriori* sur les motifs recherchés, ou sur les événements menant à la discrétisation des trajectoires.

La réduction de la variabilité par le fait de résumer les événements de soins par leurs classes ancêtres, va rapprocher certaines trajectoires, qui après cette modification partageront plus d'événements en commun. Il en va de même pour les transactions, ce qui mènera à la découverte de motifs plus fréquents. Cependant, cette modification diminue la précision des méthodes. Pouvoir conserver l'information détaillée des événements, tout en ajoutant celle de leurs classes ancêtres pourrait être une solution.

2.3.4 Les connaissances médicales et pharmacologiques pour l'analyse des trajectoires de soins

Parallèlement à l'idée d'enrichir l'intégration, la représentation et l'exploration de données médicales par l'apport de connaissances médicales externes, certains projets de recherche s'intéressent à l'apport des connaissances pour l'analyse de ces données, car il pourrait représenter des solutions aux limites des méthodes utilisées.

Connaissances médicales et comparaison de trajectoires Les structures hiérarchiques des ontologies et nomenclatures médicales ont ainsi été utilisées pour le calcul de similarités sémantiques entre événements de santé (Girardi et al., 2016). Ces mesures de similarités sémantiques, une fois introduites dans des méthodes de comparaison de trajectoires de soins, permettent de mieux prendre en compte la

similarité entre deux événements ou états de santé proches mais pourtant différents, et rendent ainsi la comparaison de trajectoires de soins plus robuste à la variabilité des données.

Connaissances médicales et recherche de motifs Dans le cadre de la fouille de données, la structure hiérarchique des nomenclatures médicales a également pu être prise en compte dans la découverte de motifs fréquents (Kost et al., 2012; Egho et al., 2013b). Les méthodes peuvent alors être qualifiées de fouille de données généralisée, multidimensionnelles ou encore multi-niveaux. Contrairement à une recherche de motifs qui se fixerait un niveau de granularité dans une hiérarchie, de telles méthodes multi-niveaux recherchent des motifs dont les éléments peuvent appartenir à plusieurs niveaux de granularité dans les hiérarchies. Ces méthodes profitent ainsi des liens hiérarchiques pour obtenir des règles plus fortes –plus fréquentes, ou plus prédictives–, sans pour autant être concernées par la perte de précision résultante du choix d'un seul niveau de granularité. Certains articles définissent enfin des mesures de qualités des motifs, basées sur l'apport de connaissances, grâce à l'utilisation d'ontologies du domaine médical et à l'introduction de similarités sémantiques. Ces mesures de qualité peuvent alors êtres dites de subjectives, à l'inverse des mesures classiques dites objectives. Paul et al. (2014) ont par exemple définit de telles mesures subjectives pour obtenir des règles d'association entre phénotypes et diagnostics liés au syndrome de dysplasie squelettique. Enfin, d'autres contributions ont montré que les méthodes de fouille de données pouvaient être enrichies de connaissances externes après traitement, lors de l'analyse des résultats. Chen et al. (2008) ont par exemple, sur la bases de connaissances médicales de l'UMLS, filtré parmi des règles d'association, celles qui étaient déjà connues.

2.4 Synthèse

La réutilisation des bases de données médico-administratives françaises, fortes de leurs couverture nationale, représente de réelles opportunités pour la recherche en santé publique, avec notamment des applications en épidémiologie, pharmaco-épidémiologie et pharmaco-vigilance. Bien qu'encouragée par le législateur, certaines particularités et complexités des données rendent leur réutilisation compliquée et limitée.

Ces complications et limitations se ressentent notamment lors des étapes d'exploration et d'analyse des trajectoires de soins, traces pour chaque patient des événements médicaux enregistrés. Des approches proposant l'intégration de connaissances dans ces étapes peuvent s'avérer prometteuses comme solutions à ces limites. De plus, le lien entre connaissances médicales et pharmacologiques externes et données médico-administratives n'est quant-à lui que peu limité. En effet, la codification systématique des données par des nomenclatures médicales rend possible ce lien. Les technologies du Web Sémantique et les ontologies médicales et pharmacologiques du *Linked Data* pourraient permettre de réaliser cette intégration de connaissances, dans l'exploration ainsi que dans l'analyse des trajectoires de soins.

Cette voie n'est pourtant que rarement mise en œuvre dans le cadre de la réutilisation des données médico-administratives française. En effet, enrichir l'exploration et l'analyse des trajectoires de soins, par des connaissances complexes et éparses, reste une tâche difficile. Des outils la facilitant sont alors nécessaires pour la dissémination de ces approches basées sur l'apport de connaissances externes aux données.

Chapitre 3

Objectifs

Sommaire

3.1 Étudier la faisabilité et l'intérêt des technologies du Web Sémantique et des ontologies médicales du *Linked Data* pour l'exploration de trajectoires de soins

L'utilisation des technologies du Web Sémantique permet une exploration de données qui intègre les connaissances d'ontologies. Les ontologies médicales et pharmacologiques du *Linked Data* formalisent des connaissances complexes, qui pourraient justement s'avérer très enrichissantes pour l'exploration de trajectoires de soins.

Pour étudier la faisabilité et l'intérêt de cette approche, plusieurs objectifs sont poursuivis durant cette thèse :

— Rassembler les ontologies distribuées pouvant être reliées aux événements médicaux des trajectoires de soins issues des bases de données médico-administratives ;

— Représenter en RDFS ou OWL les nomenclatures et thésaurus typiquement français qui manqueraient au *Linked Data* et dont l'intégration aux données des bases médico-administratives serait utile et pertinente ;

— Représenter les trajectoires de soins en RDF de façon à pouvoir intégrer les ontologies utiles en épidémiologie, pharmaco-épidémiologie et pharmacovigilance ;

— Étudier la faisabilité et l'expressivité de l'exploration de trajectoires de soins enrichie de l'intégration de ces ontologies par des requêtes SPARQL.

3.2 Utiliser des connaissances médicales et pharmacologiques du *Linked Data* pour enrichir des méthodes d'analyse de trajectoires de soins

Les connaissances d'ontologies qui décrivent les événements des trajectoires de soins pourraient permettre d'améliorer la réutilisation de certaines méthodes d'analyse et de fouille de données pour le traitement des trajectoires de soins. Afin d'utiliser les connaissances médicales et pharmacologiques du *Linked Data* pour enrichir ces méthodes, cette thèse poursuit les objectifs suivants :

— Proposer des enrichissements pour des méthodes d'analyse et de fouille de données sur la bases de connaissances médicales et pharmacologiques du *Linked Data* ;

— Proposer une exploration des résultats de ces méthodes enrichie d'ontologies médicales et pharmacologiques.

3.3 Faciliter l'accès aux connaissances médicales et pharmacologiques du *Linked Data*

Après avoir fondé les principaux objectifs de cette thèse sur l'apport de connaissances médicales et pharmacologiques, il devenait assez naturel de vouloir rendre ces approches plus généralisables et réutilisables par la communauté des chercheurs en santé publique, spécifiquement en pharmaco-épidémiologie et pharmacovigilance. Notamment, la diversité des technologies pour accéder sur le Web aux connaissances du Linked Data, celles du Web Sémantique, ou encore les REST API, ainsi que la diversité des accès avec les nombreux SPARQL endpoints, rendent la récupération des connaissances laborieuse. En plus de la récupération, l'utilisation de ces connaissances est elle aussi difficile, du fait de schémas de représentation complexes et variés, et parfois d'un manque de leur documentation. Enfin, les nombreuses ontologies médicales du *Linked Data* n'utilisent pas toutes les nomenclatures utilisées dans les bases de données médico-administratives pour codifier leurs concepts. L'utilisation de correspondances reliant plusieurs nomenclatures médicales entre elles est nécessaire pour augmenter le nombre de sources potentielles d'apport de connaissances.

L'objectif est donc de faciliter aussi bien l'accès de différentes sources de connaissances médicales et pharmacologiques, que leur liaison entre elles et avec les bases de données médico-administratives. Cet objectif s'est concentré sur la réalisation d'un paquet R. Le langage de programmation R (R Core Team, 2017), adapté à l'exploration et l'analyse statistique de données est de plus en plus utilisé par la communauté des chercheurs en santé publique, et offre la liberté de programmation pour répondre aux objectifs d'un tel paquet :

— Rendre les ontologies du *Linked Data* accessibles depuis R via l'utilisation des technologies du Web Sémantique, notamment SPARQL, ou d'autres technologies comme les REST API ;

— Faciliter le requêtage SPARQL depuis R pour la récupération de connaissances du *Linked Data*, pour des utilisateurs non-experts ;

— Faciliter l'utilisation des REST API depuis R pour la récupération de connaissances du *Linked Data*, pour des utilisateurs non-experts ;

— Faciliter l'utilisation de correspondances entre nomenclatures médicales, notamment par l'utilisation des CUI de l'UMLS, depuis R, pour relier données médico-administratives et connaissances du *Linked Data*.

Après ces objectifs pour favoriser la récupération et la liaison de connaissances du *Linked Data* aux données médico-administratives, un dernier objectif d'un tel paquet et de pouvoir proposer aux chercheurs un moyen simple d'explorer de façon efficace des tables de données R, avec des critères portant sur les connaissances récupérées.

Le chapitre 4 : Lier connaissances médicales et pharmacologiques aux bases de données médico-administratives, présente les aspects techniques liés aux objectifs des sections 3.1 et 3.3.

Le chapitre 5 : Enrichissement de l'analyse de trajectoires de soins par connaissances, porte lui sur les aspect méthodologiques plus fondamentaux liés aux objectifs énumérés en section 3.2.

Chapitre 4

Lier connaissances médicales et pharmacologiques aux bases de données médico-administratives

Sommaire

Ce chapitre traite de l'utilisation des technologies du Web Sémantique pour relier les bases de données médico-administratives à des ontologies médicales et pharmacologiques du *Linked Data*.

La section 4.1 présente les expériences menées dans le cadre de cette thèse pour baser une exploration des bases de données médico-administratives et notamment des trajectoires de soins qu'y en sont issues sur ces liens entre données et connaissances. Dans cette section on s'intéresse donc particulièrement à la faisabilité d'une telle approche, aux ontologies et aux bases de connaissances pouvant être utilisées. L'expressivité des technologies du Web Sémantique pour l'exploration de trajectoires de soins est également étudiée.

La section 4.2 présente le package R *queryMed* (Rivault et al., 2018c). Il vise à rendre plus accessible à la communauté des chercheurs en santé publique l'utilisation des technologies du Web Sémantique et les ressources médicales et pharmacologiques du *Linked Data*.

4.1 Améliorer l'intégration, la représentation et l'exploration des données médico-administratives françaises en utilisant les technologies du Web Sémantique

4.1.1 Introduction et objectifs

La réutilisation des bases de données médico-administratives pour la recherche en santé publique repose sur leur exploration. Lors de chaque étude épidémiologique, une première recherche permet de constituer une cohorte de patients à étudier. Cette exploration peut reposer sur des critères socio-démographiques, le sexe, l'âge, la commune de résidence, comme sur des critères médicaux, avec par exemple les médicaments prescrits et délivrés, les actes médicaux réalisés pour un patients, ainsi que ses diagnostics. L'exploration, sous cette forme assez simple, est le plus souvent réalisée grâce à des outils dédiés à la gestion et à l'exploration de données.

Les chercheurs en santé publique explorent de plus en plus les données médico-administratives selon des critères riches. En pharmaco-épidémiologie, l'étude des conséquences de la consommation d'un médicament sur une population requiert des connaissances propres au médicament étudié, par exemple ses indications ou ses contre-indications à un état de santé. En pharmacovigilance, on s'intéresse à rechercher des effets indésirables de médicaments, par exemple les effets secondaires, les interactions ou les contre-indications médicamenteuses.

Les données médico-administratives, par leurs contenus, permettent justement de répondre à ce type de questions :

— Quels patients présentent une contre-indication médicamenteuse ?

— Quels patients présentent une interaction médicamenteuse ?

— Quelles sont les prescriptions indiquées dans le cadre d'un état de santé étudié ?

— Quelles sont les prescriptions contre-indiquées dans le cadre d'un état de santé étudié ?

Bien que l'exploration de données médico-administratives soit souvent fondée sur des critères médicaux, ceux-ci représentent la plupart du temps des codes de médicaments, d'actes médicaux, ou de diagnostics, sélectionnés par des experts. Des explorations systématiques répondant à ce type de questions font appel à une quantité telle de connaissances qu'il peut être difficile pour les experts de les traduire par un ensemble de codes de médicaments, diagnostics et actes médicaux à investiguer. L'intégration d'ontologies dans l'exploration permet de rendre systématique cette étape.

Les technologies du Web Sémantique et les ontologies biomédicales du *Linked Data* ont montré qu'elles supportaient l'interopérabilité technologique et sémantique entre connaissances biomédicales et bases de données médicales. Nous proposons une approche utilisant ces technologies pour enrichir l'exploration de données médico-administratives, notamment dans le but de répondre à des questions nécessitant l'apport de connaissances médicales et pharmacologiques supplémentaires.

Dans le cadre de cette thèse, les objectifs devant mener à améliorer l'intégration, la représentation et l'exploration des trajectoires de soins ont tout d'abord été des objectifs visant à étudier la faisabilité d'une exploration de données médico-administratives françaises par l'utilisation des technologies du Web Sémantique. Cela a consisté à :

1. Réaliser une exploration d'un graphe RDF de données médico-administratives grâce à SPARQL ;
2. Comparer cette approche aux autres méthodes d'exploration, notamment dans la recherche de trajectoires de soins ;

Ces premiers objectifs étaient ensuite accompagnés d'objectifs relevant de l'amélioration de l'exploration par l'intégration de connaissances externes :

3. Intégrer dans l'exploration de trajectoires de soins des ontologies médicales et pharmacologiques décrivant les éléments médicaux des trajectoires (médicaments, actes médicaux, diagnostics), ainsi que la structure hiérarchique des nomenclature les codant ;
4. Évaluer l'expressivité de l'approche par l'exploration de trajectoires de soins sur des critères nécessitant l'utilisation de connaissances.

4.1.2 RDF pour la représentation de données médico-administratives

Pour étudier la faisabilité d'une exploration des données médico-administratives grâce aux technologies du Web Sémantique, nos études ont porté sur plusieurs jeux de données issues de bases médico-administratives françaises, que nous avons transformés en graphes RDF, sérialisés en *turtle*[1] (code source 2 et figure 4.1). Nous détaillons ci-après les jeux de données qui ont permis de poursuivre les objectifs 4.1.1.

Liste MSAP Avec l'introduction de la liste de gestes MSAP[2], la loi de financement de la Sécurité Sociale de 2008[3] vise à encourager le développement des opérations chirurgicales courantes en ambulatoire. Un acte sur cette liste, pour être

1. *turtle* sur le site du W3C : https://www.w3.org/TR/turtle/
2. La liste des gestes MSAP sur le site de l'Assurance Maladie : https://www.ameli.fr/medecin/exercice-liberal/presciption-prise-charge/accord-prealable/accord-prealable-chirurgie-ambulatoire
3. La loi de financement de la Sécurité Sociale de 2008 : https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000017726554#LEGIARTI000017730757

accompagné d'un hébergement dans un établissement MCO, doit ainsi bénéficier d'un accord préalable explicite de l'Assurance Maladie. En 2012, la liste s'étendait à 43 gestes, soit 203 codes CCAM. Dans le cadre de l'étude des réhospitalisations et des complications post-chirurgie pour des patients ayant bénéficié d'une opération chirurgicale issue de la liste MSAP, les données hospitalières des patients ayant reçu au moins une de ces 43 opérations en 2012 et 2013 ont été extraites du PMSI (approbation IDS #95, 30 septembre 2014). Ce sont ainsi des données de 2 700 256 séjours hospitaliers, pour un total de 15 597 374 actes médicaux, qui ont été transformées en un graphe RDF de plus de 140 millions de triplets. Le graphe RDF contenait les données socio-démographiques disponibles dans le PMSI pour les patients anonymisés associés à ces hospitalisations, leur sexe, leur âge et leur commune de résidence. Le graphe contenait également des données relatives aux séjours hospitaliers, les modes d'entrée et de sortie, les diagnostics principaux, reliés et associés, les actes médicaux, la date de sortie, le GHM, ainsi que d'autres variables hospitalières.

```
#Un patient :
:patient_X rdf:type foaf:person ;
        ansm:has_by 1975 ;
        ansm:residence ansm:18220;
        ansm:has_hospital_stay :patient_X_hospital_stay_0 .
#Et son hospitalisation :
:patient_X_hospital_stay_0 ansm:date_sortie "2012-6"^^xsd:gYearMonth ;
        ansm:GHM ansm:28Z04J ;
        ansm:entry mode:8 ;
        ansm:exit mode:8 ;
        ansm:duration 0 ;
        icd10:has_dr icd10:N189 ;
        icd10:has_dp icd10:Z490 ;
        icd10:has_das icd10:I509 ;
        icd10:has_das icd10:H903 ;
        ccam:has_hospital_procedure ccam:EZMA001 ;
        ccam:has_hospital_procedure ccam:YYYY467 .
```

Code source 2: Extrait d'un graphe RDF de données médico-administratives en sérialisation *turtle*, où sont représentés le patient fictif `patient_X` et une de ses hospitalisations, `patient_X_hospital_stay_0`.

Arthroplastie Dans le cadre de l'étude des trajectoires de soins des patients opérés pour une arthroplastie (de la hanche ou du genou) en 2012, des données de soins de ville et d'hôpital issues de l'EGB, ont été extraites pour 1700 patients (approbation IDS #202, 29 juin 2016). Ces données couvrent la période de 2012 à 2013 afin de pouvoir étudier les consommations après hospitalisation pour tous les patients. Les données de ces 1700 patients contenaient principalement des GHM, des actes médicaux et diagnostics (principaux, reliés et associés) réalisés en établissement MCO,

ainsi que des délivrances en pharmacie de médicaments qui leur étaient prescrits et remboursés. Ces données ont ensuite été transformées en un graphe RDF de plus de 1 140 000 triplets.

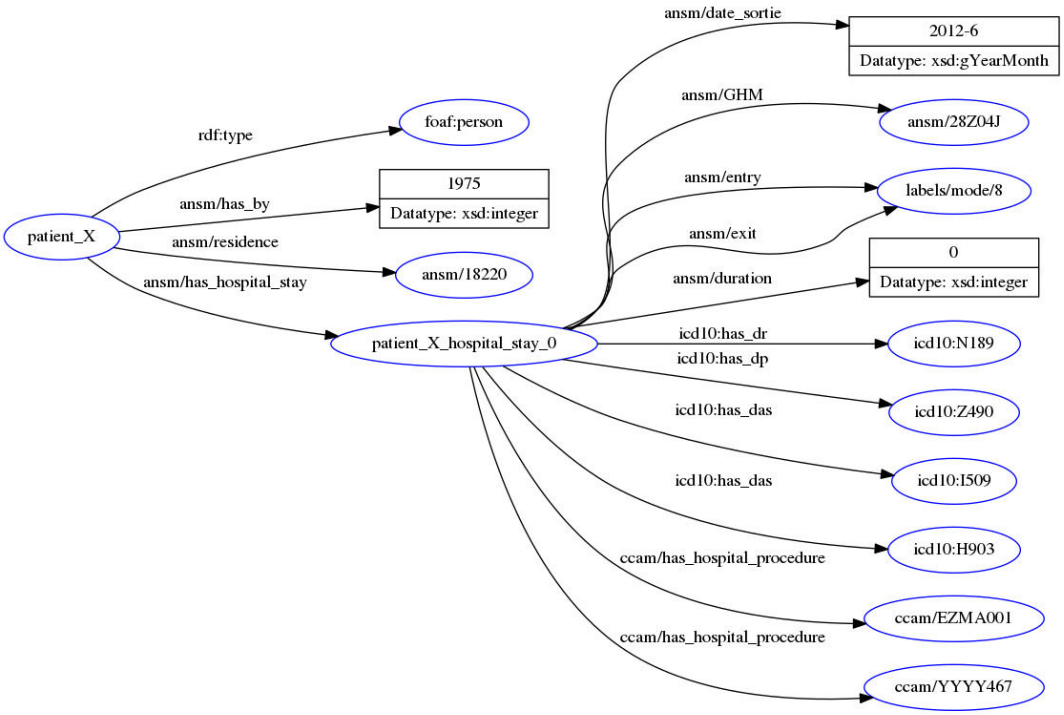

FIGURE 4.1 – Graphe RDF du patient fictif `patient_X` et d'une de ses hospitalisations, `patient_X_hospital_stay_0` (code source 2).

4.1.3 SPARQL pour l'exploration de données médico-administratives

4.1.3.1 Data management avec SPARQL

Les premières expérimentations concernant l'exploration de données médico-administratives grâce aux technologies du Web Sémantique ont tout d'abord porté sur la reproduction des explorations pouvant être réalisées dans des études d'épidémiologie ou de pharmaco-épidémiologie. Ces explorations consistent en premier lieu à l'identification et au dénombrement de patients ou d'événements répondant à des critères de sélection relevant des variables hospitalières et des variables propres aux patients. Le graphe de triplets des patients ayant bénéficié d'une opération parmi la liste des actes MSAP a été chargé dans le *triplestore* FUSEKI[4]. Une fois le graphe de triplets chargé, des requêtes SPARQL permettent d'effectuer l'identification ou le dénombrement de patients et d'événements. Dans le cadre de cette étude, la première exploration a consisté à distinguer et à dénombrer les hospitalisations longues

4. FUSEKI sur le site d'Apache Jena : https://jena.apache.org/documentation/serving_data/

des hospitalisations en ambulatoire. Une seconde requête a permis de dénombrer les hospitalisations qui étaient suivies par une réhospitalisation, jusqu'à trois mois après l'opération. Enfin, une dernière requête permettait d'identifier, parmi ces réhospitalisations, celles qui étaient liées à une complication post-chirurgie, pouvant s'apparenter à une infection nosocomiale, d'après la liste de codes CIM-9 fournie par Owens et al. (2014), une fois adaptée à la CIM-10.

Les critères suivant traduisent une hospitalisation en ambulatoire, une hospitalisation longue, une réhospitalisation ou une complication :

— Une hospitalisation était dite ambulatoire lorsque son GHM finissait par J ou par T, indiquant une hospitalisation en ambulatoire ou de très courte durée, et que sa durée de séjour (mesurée en jours) était nulle. Les patient de ses hospitalisations devaient en plus arriver de leur domicile (mode d'entrée=8) pour repartir également vers leur domicile (mode de sortie=8) ;

— Une réhospitalisation était une nouvelle hospitalisation survenant après une hospitalisation contenant un geste de la liste MSAP, dans un délai de 90 jours ;

— Une réhospitalisation était dite avec complication, si son diagnostics principal ou un de ses diagnostics associés était présent dans la liste de codes CIM-10 de complications [5].

Ils sont traduits par autant de requêtes SPARQL, donc les résultats sont présentés par la figure 4.2.

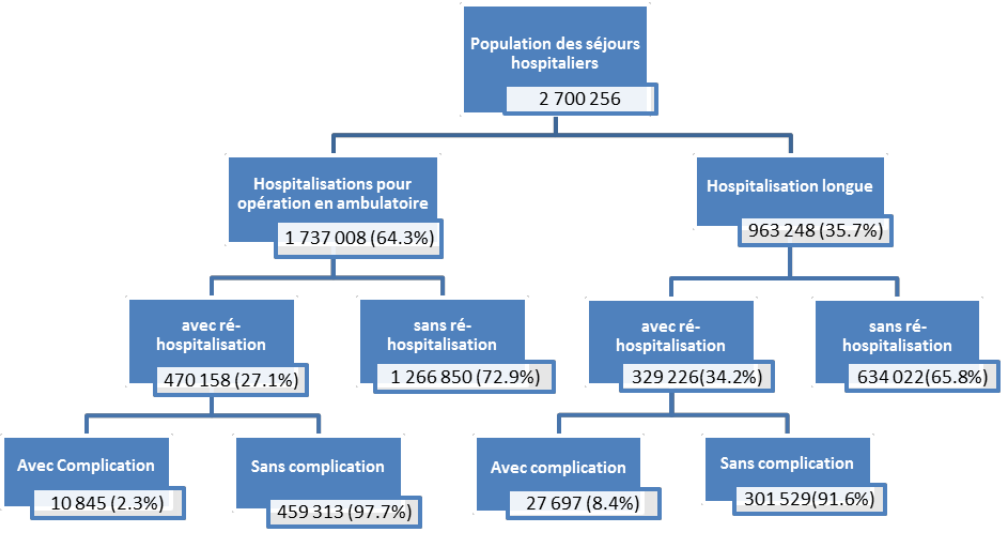

FIGURE 4.2 – Les technologies du Web Sémantique appliquées à l'exploration de données médico-administratives. Chaque catégorie et son dénombrement correspondent au résultat d'une requête SPARQL.

5. Liste fournie par Owens et al. (2014) : https://jamanetwork.com/data/Journals/JAMA/929772/JOI140003supp1_prod.pdf

Ces résultats portaient sur la faisabilité d'une exploration de données médico-administratives utilisant les technologies du Web Sémantique plus que sur les résultats épidémiologiques de ces dénombrements. Une exploration similaire a été réalisée en amont à une étude statistique plus détaillée sur le risque de complication post-chirurgie ambulatoire pour une angioplastie (Rivault et al., 2016b).

Cette première expérimentation a pu montrer que les technologies du Web Sémantique et le *triplestore* FUSEKI sont adaptés à l'exploration des données médico-administratives, et ce notamment pour un volume de donnée important (plus de 140 millions de triplets pour plus de 2,7 millions de séjours hospitaliers).

4.1.3.2 Comparaison de SPARQL, SQL et R pour l'exploration des bases de données médico-administratives et de leurs trajectoires de soins

Bien qu'adaptée, nous voulions pouvoir comparer cette approche aux approches plus répandues dans la réutilisation des données médico-administratives pour la recherche en santé publique. Le langage de requête structurée SQL est sans doute la méthode la plus répandue pour la gestion et l'exploration de données médico-administratives volumineuses (Collen, 2012). Cependant, il s'avère peu adapté à la recherche de séquences d'événements, telles que les trajectoires de soins (Wongsuphasawat et al., 2012). Au contraire, en parcourant un graphe, SPARQL offre un moyen adapté à la recherche de séquences d'événements (Rinne et al., 2012). Le langage de programmation R, dédié à la statistique plutôt qu'à l'exploration de données, est lui de plus en plus utilisé pour cette étape de gestion et d'exploration de données, en amont à leur analyse. Il offre notamment une grande liberté de programmation, qui a favorisé le développement de packages pour de la gestion et de l'exploration de données tout comme pour l'utilisation des technologies du Web Sémantique à travers R. Le package *rrdf* (Willighagen, 2014) permet par exemple de charger un graphe de trajectoires de soins, et également de l'explorer au moyen de requêtes SPARQL. Le package *sqldf* (Grothendieck, 2017) permet lui de requêter les tables d'une base SQLite et des jeux de données R au moyen de requêtes SQL.

Dans le but de comparer une exploration de graphe RDF avec SPARQL à une exploration de données dans leur format initial en table, avec SQL ou R, le logiciel et langage de programmation R a donc été choisi. L'objectif n'était pas de comparer SPARQL, SQL et R, trois technologies et outils bien différents, mais plutôt les usages qui leur sont fait dans l'exploration de données médico-administratives.

La principale limite à utiliser R pour de l'exploration de données, qu'elles soient sous la forme de tables de données ou de graphes RDF, est sa capacité à gérer des données volumineuses. R n'étant pas en soi un outil dédié pour la gestion et l'exploration de données, charger un graphe RDF très volumineux (tout comme une table de données très volumineuse) est bien plus limité que dans un système de gestion de données, par exemple un triplestore. Le chargement des tables de données ou du graphe RDF des patients opérés d'un acte dans la liste MSAP était

alors impossible avec les fonctions de base de R comme avec le package *rrdf*, du fait de leur grande volumétrie. Pour cette raison, un sous-échantillon des données liées à la liste MSAP a été sélectionné, en se concentrant sur les patients ayant été opérés en 2012 d'un accès vasculaire arterioveineux d'un membre avec pose d'endoprothèse par voie artérielle transcutanée (EZAF002 dans la CCAM). Les trajectoires de soins des 435 patients de cet échantillon ont été chargées dans R au format RDF grâce au package *rrdf*. Les tables de données ayant servi à la création du graphe RDF ont également été chargées dans R. L'étude a consisté à reproduire les recherches d'hospitalisations ambulatoires ou non, menant ou non à une réhospitalisation, avec ou sans diagnostic de complication (section 4.1.3.1). Plusieurs méthodes ont été utilisées et comparées, en terme d'efficacité (i.e. en temps de calcul) et en terme d'expressivité, pour l'identification de ces sous-ensembles d'hospitalisations :

— Des algorithmes écrits en R *base* ;

— Des requêtes SPARQL grâce au package *rrdf* ;

— Des requêtes SQL grâce au package *sqldf*.

Les résultats des trois méthodes ont mené aux mêmes identifications de sous-ensembles de séjours hospitaliers. Les temps d'exécution sont présentés en table 4.1.

	Algorithmes R	SPARQL dans R	SQL dans R
Détection des hospitalisations ambulatoires ou non	≈ 5 sec	$\approx 0,6$ sec	$\approx 0,3$ sec
Détection des réhospitalisations pour complication	10 à 20 sec	$\approx$ 6.3 sec	$\approx$ 0,6 sec

TABLE 4.1 – Temps d'exécution approximatifs pour l'identification des hospitalisations ambulatoires ou non, et des réhospitalisations pour complication.

Les temps de calculs plus importants pour les algorithmes écrits en R étaient attendus. Les jointures entre tables et les processus itératifs d'exploration rendent la recherche longue et compliquée. Les sous-ensembles de patients ont été identifié plus rapidement avec SQL qu'avec SPARQL. Suite à ces résultats, nous avons étudié l'expressivité de SPARQL et de SQL pour cette application : dédié à la recherche dans des graphes, SPARQL est-il aussi adapté que SQL, voire plus, dans l'écriture de requêtes complexes pour l'exploration de données médico-administratives ? Pour tenter de répondre à cette question, les codes source 3 et 4 présentent les requêtes SPARQL et SQL permettant l'identification des hospitalisations ambulatoires.

```
prefix [...]
SELECT DISTINCT ?hospitalisation
  WHERE {
    ?patient ansm:has_hospital_stay ?hospitalisation .
    ?hospitalisation ansm:GHM ?ghm .
    ?hospitalisation ansm:duration ?duree .
    ?hospitalisation ansm:entry ?entree .
    ?hospitalisation ansm:exit ?sortie .
    ?hospitalisation ansm:date_sortie ?date_sortie .
    ?hospitalisation ccam:has_hospital_procedure ?ccam .
    #accès vasculaire arterioveineux d'un membre
    #avec pose d'endoprothèse par voie artérielle transcutanée
    FILTER (?ccam = ccam:EZAF002)
    #L'hospitalisation se termine en 2012
    FILTER (?date_sortie <= '2012-12-31'^^xsd:gYearMonth)
    FILTER (?date_sortie >= '2012-01-01'^^xsd:gYearMonth)
    #GHM terminant par J ou T
    FILTER(strEnds(str(?ghm), 'T') || strEnds(str(?ghm), 'J')
    #Durée de séjour nulle
    FILTER (?duree=0)
    #Sortie et entrée vers domicile
    FILTER (?entree=mode:8 && ?sortie=mode:8)
}
```

Code source 3: Requête SPARQL pour la détection des hospitalisations ambulatoires pour accès vasculaire arterioveineux d'un membre avec pose d'endoprothèse par voie artérielle transcutanée (EZAF002 dans la CIM-10). La transformation des tables en un graphe RDF rend les clefs de jointures invisibles pour l'utilisateur.

Les deux requêtes SPARQL et SQL sont assez similaires. Leur différence essentielle réside dans une recherche qui parcourt un graphe pour l'une, et des tables pour l'autre. La résultante à cette différence est la nécessité d'une clef de jointure entre la table des séjours et la table des actes médicaux pour les requêtes SQL. Sur une exploration relativement simple comme celle-ci, la complexité à utiliser une clef de jointure ne se fait pas forcément ressentir. Néanmoins, lorsque l'on réalise une exploration plus complexe, les jointures de multiples tables peuvent rendre l'écriture des requêtes SQL moins intuitive que celle des requêtes équivalentes SPARQL.

```
SELECT DISTINCT sejour.num_Anonyme, sejour.num_de_sejour,
                sejour.FINESS_PMSI, sejour.RSA,
                sejour.Duree_Sejour, sejour.Date_Sortie,
       --Jointure des tables sejour et actes
       FROM sejour INNER JOIN actes
       --Clefs de jointure
       ON sejour.FINESS_PMSI=actes.FINESS_PMSI AND
          sejour.RSA=actes.RSA AND
          sejour.Date_Sortie = actes.Date_Sortie AND
          --accès vasculaire arterioveineux d'un membre
          --avec pose d'endoprothèse par voie artérielle transcutanée
          actes.CCAM='EZAF002'
          WHERE
          --L'hospitalisation se termine en 2012
                sejour.Date_Sortie BETWEEN '2012-01-01'
                AND '2012-12-31' AND
          --Sortie et entrée vers domicile
                sejour.Mode_Sortie=8 AND
                sejour.Mode_Entree=8 AND
          --Durée de séjour nulle
                sejour.Duree_Sejour=0 AND
          --GHM terminant par J ou T
                (sejour.GHM_Obtenu LIKE '%T' OR sejour.GHM_Obtenu LIKE '%J')
```

Code source 4: Requête SQL pour la détection des hospitalisations ambulatoires pour accès vasculaire arterioveineux d'un membre avec pose d'endoprothèse par voie artérielle transcutanée (EZAF002 dans la CIM-10).

Nous souhaitions également comparer l'expressivité de requêtes SPARQL et SQL dans la recherche de suites d'événements, c'est à dire de trajectoires de soins. La recherche de ré-hospitalisation est justement une suite d'événements : on recherche une première hospitalisation selon certains critères, puis une seconde qui survient après la première. Une telle recherche avec SQL peut se faire avec une requête imbriquée jointe aux tables initiales, quand SPARQL offre la possibilité de parcourir un graphe RDF selon divers chemins avec une seule requête. Les codes sources 5 et 6 des requêtes SPARQL et SQL pour l'identification des ré-hospitalisations pour complication illustrent bien cette différence notable pour la recherche de séquences d'événements, ou trajectoires de soins. Elle se ressentirait d'autant plus avec la recherche de trajectoires de soins à plus de deux événements.

```
SELECT DISTINCT ?hospitalisation
  WHERE{
    #Le patient a une hospitalisation_1
    ?patient ansm:has_hospital_stay ?hospitalisation_1 .
    ?hospitalisation_1 ansm:date_sortie ?date_hosp_1 .
    ?hospitalisation_1 ccam:has_hospital_procedure ?ccam .
    #Survenue en 2012
    FILTER (?date_hosp_1 <= '2012-12-31'^^xsd:gYearMonth)
    FILTER (?date_hosp_1 >= '2012-01-01'^^xsd:gYearMonth)
    #Pour un accès vasculaire arterioveineux d'un membre
    #avec pose d'endoprothèse par voie artérielle transcutanée
    FILTER (?ccam = ccam:EZAF002)

    #Le patient a une hospitalisation_2
    ?patient ansm:has_hospital_stay ?hospitalisation_2 .
    ?hospitalisation_2 ansm:date_sortie ?date_hosp_2 .
    ?hospitalisation_2 icd10:has_dp ?diag_principal .
    OPTIONAL {?hospitalisation_2 icd10:has_das ?diag_associe .}
    #Cette hospitalisation est une complication
    FILTER (?das IN (icd10:T813,icd10:T818, ...) ||
    ?dp IN (icd10:T813,icd10:T818, ...))}

    #hospitalisation_2 est une réhospitalisation de hospitalisation_1
    FILTER (?date_hosp_2 > ?date_hosp_1)
    #Qui survient dans un délai de 90 jours
    FILTER (?date_hosp_2-?date_hosp_1 <= 'P90D'^^xsd:duration)
```

Code source 5: Requête SPARQL pour la détection des hospitalisations pour accès vasculaire arterioveineux d'un membre avec pose d'endoprothèse par voie artérielle transcutanée (EZAF002 dans la CIM-10) menant à une complication.

Le parcours multiple du graphe RDF ne complexifie pas tant l'écriture de la requête SPARQL. Chaque paragraphe détaille une recherche dans le graphe, et chaque recherche peut communiquer avec les autres, par exemple avec des clauses *FILTER*. Avec SQL, la sous-requête pour identifier les premières hospitalisations pour un accès vasculaire arterioveineux est jointe aux tables des séjours et des actes. Cette sous-requête rend l'écriture de la requête finale relativement compliquée. L'augmentation du nombre d'événements de la trajectoire à rechercher ne fait que complexifier d'avantage l'écriture de la requête SQL.

```
SELECT DISTINCT hospitalisation_num_Anonyme, hospitalisation_num_de_sejour,
                hospitalisation_FINESS_PMSI, hospitalisation_RSA,
                hospitalisation_Date_Sortie
FROM sejour INNER JOIN --Jointure avec la table des hospitalisations premières
--Sous-requête pour la recherche des hospitalisations premières
(SELECT DISTINCT sejour.num_Anonyme as hospitalisation_num_Anonyme,
                 sejour.num_de_sejour as hospitalisation_num_de_sejour,
                 sejour.FINESS_PMSI as hospitalisation_FINESS_PMSI,
                 sejour.RSA as hospitalisation_RSA,
                 sejour.Date_Sortie as hospitalisation_Date_Sortie
 FROM sejour INNER JOIN actes
 ON sejour.FINESS_PMSI=actes.FINESS_PMSI AND sejour.RSA=actes.RSA AND
    sejour.Date_Sortie = actes.Date_Sortie AND  --Clef de jointure
    ccam.CCAM='EZAF002' --accès vasculaire arterioveineux [...]
    WHERE sejour.Date_Sortie<='2012-12-31' --L'hospitalisation se termine en 2012
) AS subquery
ON sejour.num_Anonyme=subquery.hospitalisation_num_Anonyme
LEFT JOIN das --Jointure avec la table des diagnostics associés
ON sejour.FINESS_PMSI=das.FINESS_PMSI AND
   sejour.RSA=das.RSA AND
   sejour.Date_Sortie = das.Date_Sortie
WHERE (DAS IN (T813,T818, ...) OR DP IN (T813,T818, ...)) AND --complications
 --Les complications qui surviennent après une hospitalisation première
 julianday(sejour.Date_Sortie)>julianday(hospitalisation_Date_Sortie) AND
 --Dans un délai de 90 jours
 (julianday(sejour.Date_Sortie)-julianday(hospitalisation_Date_Sortie))<=90
```

Code source 6: Requête SQL pour la détection des hospitalisations pour accès vasculaire arterioveineux d'un membre avec pose d'endoprothèse par voie artérielle transcutanée (EZAF002 dans la CIM-10) menant à une complication.

Pour une exploration simple des données médico-administratives, il est donc assez difficile de comparer l'expressivité de requêtes SPARQL à celle de requêtes SQL, qui peuvent alors être jugées équivalentes. Lorsque l'exploration traite de la recherche de trajectoires de soins, c'est-à-dire de séquences d'événements médicaux, le parcours de graphe qu'offre SPARQL rend l'écriture des requêtes plus simple que celle de requêtes équivalentes SQL. Également, lorsque l'exploration nécessite l'agrégation de données éparses et hétérogènes, l'utilisation de nombreuses tables et clefs de jointure peut rendre l'exploration d'un seul graphe plus intuitive. Néanmoins, ce gain d'expressivité de SPARQL se fait généralement au détriment de son efficacité, en terme de temps d'exécution. Les résultats en termes de temps d'exécution sont tout de même à considérer avec certaines précautions : les différentes approches ont été réalisées dans R. Les temps de calcul pour les requêtes SQL et SPARQL lancées dans des systèmes de gestions de données adaptés seraient sûrement plus faibles.

Adapté à l'exploration de graphes de données et de connaissances, c'est surtout dans des explorations nécessitant l'apport de connaissances externes aux données, telles que les connaissances du Web Sémantique, que SPARQL doit pouvoir révéler tous ses avantages en terme d'efficacité et d'expressivité. Dans ce sens, Tim Berners-Lee expliquait en 2008, lors de la publication de SPARQL[6] : « Essayer d'utiliser le Web Sémantique sans SPARQL revient à essayer d'utiliser une base de données relationnelle sans SQL ».

4.1.4 Intégration d'ontologies biomédicales

Dans le but d'évaluer l'expressivité et les apports d'une approche basée sur les technologies du Web Sémantique pour l'exploration de données médico-administratives, nous nous sommes intéressés à les relier à des ontologies biomédicales du *Linked Data*.

4.1.4.1 Le *Linked Data* pour l'exploration des bases de données médico-administratives

De nombreuses ontologies du *Linked Data* décrivant des concepts médicaux et pharmacologiques peuvent être reliées aux données médicales présentes dans les bases médico-administratives. L'intégration d'ontologies décrivant les nomenclatures médicales (par exemple l'ATC ou la CIM-10) dans l'exploration des trajectoires de soins permet une exploration tenant compte des relations hiérarchiques entre médicaments et classes de médicaments, maladies et chapitres de maladies. D'autres ontologies médicales, en plus d'apporter des hiérarchies de nomenclatures, décrivent des relations entre médicaments, actes médicaux et maladies. NDF-RT contient par exemple des relations d'indication et contre-indication entre médicaments et maladies. La suite de cette section détaille l'utilisation et l'intégration des ontologies et bases de connaissances biomédicales qui ont été intégrées à des graphes RDF de trajectoires de soins dans le cadre de cette thèse.

Ontologie de la CIM-10 Disponible sur BioPortal[7] en anglais et au format RDFS, ou sur SIFR BioPortal[8] en français, elle permet d'adosser à un graphe de trajectoires de soins la structure hiérarchique de la CIM-10. Notamment, les requêtes SPARQL peuvent se baser sur cette structure pour une recherche transitive à partir des chapitres de diagnostics et états de santé de la CIM-10 (code source 7). De plus, une correspondance des classes CIM-10 avec les CUI de l'UMLS est mise à disposition dans cette ontologie. Cette correspondance permet ainsi de relier facilement des codes CIM-10 à d'autres concepts et classes issues de l'UMLS (figure 4.3).

6. Le W3C ouvre le Web des données avec SPARQL : https://www.w3.org/2007/12/sparql-pressrelease

7. La CIM-10 en RDFS sur BioPortal : https://bioportal.bioontology.org/ontologies/ICD10CM

8. La CIM-10 sur SIFR BioPortal : http://bioportal.lirmm.fr/ontologies/CIM-10

```
PREFIX [...]

SELECT DISTINCT *
WHERE {
  ?CIM10 rdfs:subClassOf* icd10:A00-A09 .
  ?CIM10 skos:prefLabel ?CIM10_label .
  OPTIONAL{?CIM10 umls:cui ?umls .}
  OPTIONAL{?NDFRT  ndf:UMLS_CUI ?umls .
           ?NDFRT rdfs:label ?NDFRT_label .}
  }
```

Code source 7: Requête SPARQL recherchant labels et correspondances CUI des sous-classes de la classe A00-A09 de la CIM-10. Les classes de l'ontologie NDF-RT associés à ces codes CUI, sont ensuite recherchés, ainsi que leur label. L'« * » après `rdfs:subClassOf` signifie que toutes les sous-classes de A00-A09 ainsi que la classe A00-A09 elle-même sont parcourues par la requête. Cette opération, appellée *property path*, permet ainsi une recherche transitive de graphe RDF.

	CIM10	CIM10_label	umls	NDFRT	NDFRT_label
1	icd10:A00-A09	"Intestinal infectious diseases (A00-A09)"@eng	"C0178238"		
2	icd10:A00	"Cholera"@eng	"C0008354"	ndf:N0000000761	"Cholera [Disease/Finding]"
3	icd10:A00.0	"Cholera due to Vibrio cholerae 01, biovar cholerae"@eng	"C0008354"	ndf:N0000000761	"Cholera [Disease/Finding]"
4	icd10:A00.0	"Cholera due to Vibrio cholerae 01, biovar cholerae"@eng	"C0494021"		
5	icd10:A00.1	"Cholera due to Vibrio cholerae 01, biovar eltor"@eng	"C0343372"		
6	icd10:A00.9	"Cholera, unspecified"@eng	"C0008354"	ndf:N0000000761	"Cholera [Disease/Finding]"
7	icd10:A01	"Typhoid and paratyphoid fevers"@eng	"C0275976"		
8	icd10:A01.0	"Typhoid fever"@eng	"C0041466"	ndf:N0000003059	"Typhoid Fever [Disease/Finding]"
9	icd10:A01.0	"Typhoid fever"@eng	"C2880084"		
10	icd10:A01.00	"Typhoid fever, unspecified"@eng	"C0041466"	ndf:N0000003059	"Typhoid Fever [Disease/Finding]"

FIGURE 4.3 – Extrait des résultats de la requête 7 exécutée sur le *triplestore* FUSEKI. Si une grande majorité des concepts de la CIM-10 conduisent au moins à un code CUI, ces codes ne mènent néanmoins pas tous à un concept de l'ontologie NDF-RT.

Ontologie de l'ATC Également disponible sur Bioportal [9] en anglais et en RDFS, ainsi qu'en français sur SIFR BioPortal [10], cette ontologie permet d'adosser aux triplets RDF la structure hiérarchique en arbre de l'ATC. Cette nomenclature internationale peut être reliée aux codes CIP, qui sont utilisés pour coder les médicaments dans les bases de données médico-administratives françaises. De la même façon, une correspondance avec les CUI de l'UMLS est fournie dans l'ontologie.

National Drug File-Reference Terminology L'ontologie de NDF-RT, disponible sur BioPortal [11], propose des hiérarchies pour différents types de concepts médicaux, associés à la pharmacologie. La hiérarchie centrale relative aux médicaments, s'accompagne ainsi d'autres hiérarchies concernant les maladies et états de santé, les ingrédients composant les médicaments, leurs mécanismes d'action ou encore leurs effets physiologiques. Les concepts de ces hiérarchies peuvent être reliés par des relations de rôle. Ces relations décrivent ainsi des connaissances propres aux médicaments et/ou états de santé. Dans le cadre de la réutilisation des bases médico-administrative pour une application de pharmaco-épidémiologie, les relations comme "has_PE" pour *has a physiologic effect*, "may_treat", "may_prevent", "induces", "CI_with" pour *contraindicated with*, ou encore "effect_may_be_inhibited_by" sont particulièrement intéressantes à relier aux nomenclatures utilisée dans ces bases. Faisant partie de l'UMLS, chaque concept est associé à au moins un CUI. Un lien entre les concepts de NDF-RT et ceux de l'ATC et la CIM-10, et donc de données médico-administratives française, est donc très souvent réalisable (figure 4.3 et code source 7).

DrugBank DrugBank [12] est une base de connaissances libre, accessible en ligne, qui décrit près de 12 000 médicaments, par leurs caractéristiques chimiques, pharmacologiques et pharmaceutiques. Un sous-ensemble de toutes ces descriptions est disponible aux formats du Web Sémantique sur le SPARQL endpoint de Bio2RDF [13]. Nous nous intéressons particulièrement aux relations d'interaction entre médicaments présentes dans DrugBank. Une correspondance avec la classification ATC est souvent disponible pour les médicaments de DrugBank, facilitant ainsi le lien entre DrugBank et données médico-administratives françaises.

DID et DIKB Malgré une représentation en table de données, on peut tout de même classer ces deux bases de connaissances parmi le *Linked Data*, dans le sens où elles sont toutes les deux construites en partie par l'agrégation de connaissances issues d'ontologies du *Linked Data*. Leur liaison à d'autres ontologies et nomenclatures est en plus assurée grâce à une codification des médicaments avec DrugBank pour DIKB et des médicaments et états de santé en CUI pour DID. De plus, leur

9. L'ATC sur BioPortal : https://bioportal.bioontology.org/ontologies/ATC
10. L'ATC sur SIFR BioPortal : http://bioportal.lirmm.fr/ontologies/ATCFRE
11. NDF-RT sur BioPortal : https://bioportal.bioontology.org/ontologies/NDFRT
12. Le site internet de DrugBank : https://www.drugbank.ca/
13. Le SPARQL endpoint de Bio2RDF : http://sparql.bioontology.org/

transformation en graphes RDFS est relativement simple. Reliées à des trajectoires de soins de bases de données médico-administratives, elles permettent de baser une recherche sur des interactions entre médicaments ainsi que sur des indications entre médicaments et états de santé.

4.1.4.2 Représentation en RDFS de nomenclatures et thésaurus manquant au *Linked Data*

Certaines nomenclatures ou sources de connaissances médicales et pharmacologiques ne font cependant pas encore partie du *Linked Data*, et ne sont pas disponibles aux formats du Web Sémantique. Leur utilisation pourrait néanmoins représenter un réel apport dans l'exploration et l'analyse des données médico-administratives françaises. En particulier, au commencement de cette thèse, à notre connaissance aucune version de la CCAM, nomenclature française utilisée pour coder les actes médicaux dans les bases médico-administratives, n'avait alors été publiée au format du Web Sémantique. Également, aucune version du thésaurus des interactions médicamenteuses de l'ANSM n'avait été publiée dans un format exploitable.

Code	Texte	Activité	Phase	Tarif Secteur 1 / adhérent OPTAM/OPTAM-CO (en euro)	Tarif Hors secteur 1 / hors adhérent OPTAM/OPTAM-CO (en euro)
1	SYSTÈME NERVEUX CENTRAL, PÉRIPHÉRIQUE ET AUTONOME *À l'exclusion de : analgésie postopératoire* *Par intrathécal, on entend : dans l'espace subarachnoïdien.* *Par infiltration anesthésique d'un nerf, on entend : injection d'un agent pharmacologique au contact d'un nerf, par voie transcutanée.* *Par bloc anesthésique continu d'un nerf, on entend : injection d'un agent pharmacologique au contact d'un nerf avec pose d'un cathéter, par voie transcutanée.*				
01.01	ACTES DIAGNOSTIQUES SUR LE SYSTÈME NERVEUX *À l'exclusion de : actes diagnostiques au niveau* *- des muscles oculomoteurs ou de la paupière (cf chapitre 02)* *- du larynx (cf chapitre 06)* *- du périnée (cf chapitre 08)* *- des muscles ptérygoïdiens (cf chapitre 11)* *- du diaphragme (cf chapitre 12)*				
01.01.01	Explorations électrophysiologiques du système nerveux				
01.01.01.01	Électromyographie [EMG] *Facturation : les examens électromyographiques doivent être pratiqués avec un appareil comportant un système d'enregistrement permettant en, différé, une étude qualitative et quantitative*				
AHQP001	Électromyographie par électrode de surface, sans enregistrement vidéo	1	0	Non pris en charge	Non pris en charge
AHQP002	Électromyographie par électrode de surface, avec enregistrement vidéo	1	0	Non pris en charge	Non pris en charge

FIGURE 4.4 – Extrait du fichier Excel®de la CCAM, disponible sur le site de l'Assurance Maladie.

CCAM La classification commune des actes médicaux est une nomenclature française pour la codification des gestes pratiqués par les médecins. Elle est ainsi utilisée dans le SNIIRAM et dans le PMSI pour codifier les actes médicaux. Produite par l'Assurance Maladie [14], elle est délivrée au format propriétaire de tableau

14. CCAM sur le site de l'Assurance Maladie : https://www.ameli.fr/accueil-de-la-ccam/telechargement/index.php

Excel®(figure 4.4) et n'obtient ainsi que deux étoiles dans le classement des données liées par Tim Berners-lee (figure 2.5). Cependant, une version en RDFS est depuis peu fournie par le CISMEF, et disponible sur SIFR BioPortal[15], obtenant elle cinq étoiles. Dans le cadre de cette thèse et dans l'exploration des bases de données médico-administratives au moyen des technologies du Web Sémantique, nous avons transformé ce fichier de la CCAM en un graphe RDFS (code source 8, reprenant l'extrait de la figure 4.4).

```
ccam:01 rdf:type owl:Class ;
        skos:prefLabel "SYSTÈME NERVEUX CENTRAL, PÉRIPHÉRIQUE ET AUTONOME"@fr ;
        skos:notation "01"^^xsd:string ;
        rdfs:subClassOf owl:Thing .
ccam:01.01 rdf:type owl:Class ;
        skos:prefLabel "ACTES DIAGNOSTIQUES SUR LE SYSTÈME NERVEUX"@fr ;
        skos:notation "01.01"^^xsd:string ;
        rdfs:subClassOf ccam:01 .
ccam:01.01.01 rdf:type owl:Class ;
        skos:prefLabel "Explorations électrophysiologiques du système nerveux"@fr ;
        skos:notation "01.01.01"^^xsd:string ;
        rdfs:subClassOf ccam:01.01 .
ccam:01.01.01.01 rdf:type owl:Class ;
        skos:prefLabel "Électromyographie [EMG]"@fr ;
        skos:notation "01.01.01.01"^^xsd:string ;
        rdfs:subClassOf ccam:01.01.01 .
ccam:AHQP001 rdf:type owl:Class ;
        skos:prefLabel "Électromyographie par électrode de surface [...]"@fr ;
        skos:notation "AHQP001"^^xsd:string ;
        rdfs:subClassOf ccam:01.01.01.01 .
```

Code source 8: Extrait de la CCAM en RDFS, en sérialisation *turtle*. Les prédicats $rdfs{:}subClassOf$ décrivent les liens de sous-classe entre concepts de la CCAM, et ainsi la structure hiérarchique de cette nomenclature. On peut retrouver ces prédicats dans les requêtes SPARQL pour la recherche de trajectoires de soins possédant un actes appartenant à un chapitre d'actes de la CCAM. Le script Python pour transformer un fichier Excel®de la CCAM de l'Assurance Maladie au format RDFS est disponible sur https://github.com/yannrivault/CCAM_ontology.

Thésaurus des interactions médicamenteuses de l'ANSM Ce thésaurus[16] recense l'ensemble des interactions médicamenteuses identifiées par l'ANSM. Il doit fournir aux professionnels de santé un guide pharmaco-thérapeutique pour l'aide à

15. CCAM sur SIFR BioPortal : http://bioportal.lirmm.fr/ontologies/CCAM
16. Thésaurus des interactions médicamenteuses de l'ANSM : https://ansm.sante.fr/var/ansm_site/storage/original/application/de444ea9eb4bc084905c917c902a805f.pdf

la prescription. Chaque interaction est associée à un niveau de gravité, allant de la prise en compte, de la précaution d'emploi, d'une association déconseillé jusqu'à la contre-indication. Délivré au format PDF, ces données obtiennent une seule étoile selon le classement des données liées. Il est depuis peu présent dans la base de connaissance DIKB. Dans le but de pouvoir baser une exploration de données médico-administratives sur des critères complexes, tels que les interactions médicamenteuses, nous avons également transformé ce fichier en un graphe RDFS.

Les ontologies du *Linked Data* et bases de connaissances énumérées en section 4.1.4.1, la CCAM et le thésaurus des interactions médicamenteuses, forment alors un graphe de relations entre concepts pharmacologiques et médicaux (médicaments, actes et diagnostics), dont une partie est représentée par la figure 4.5.

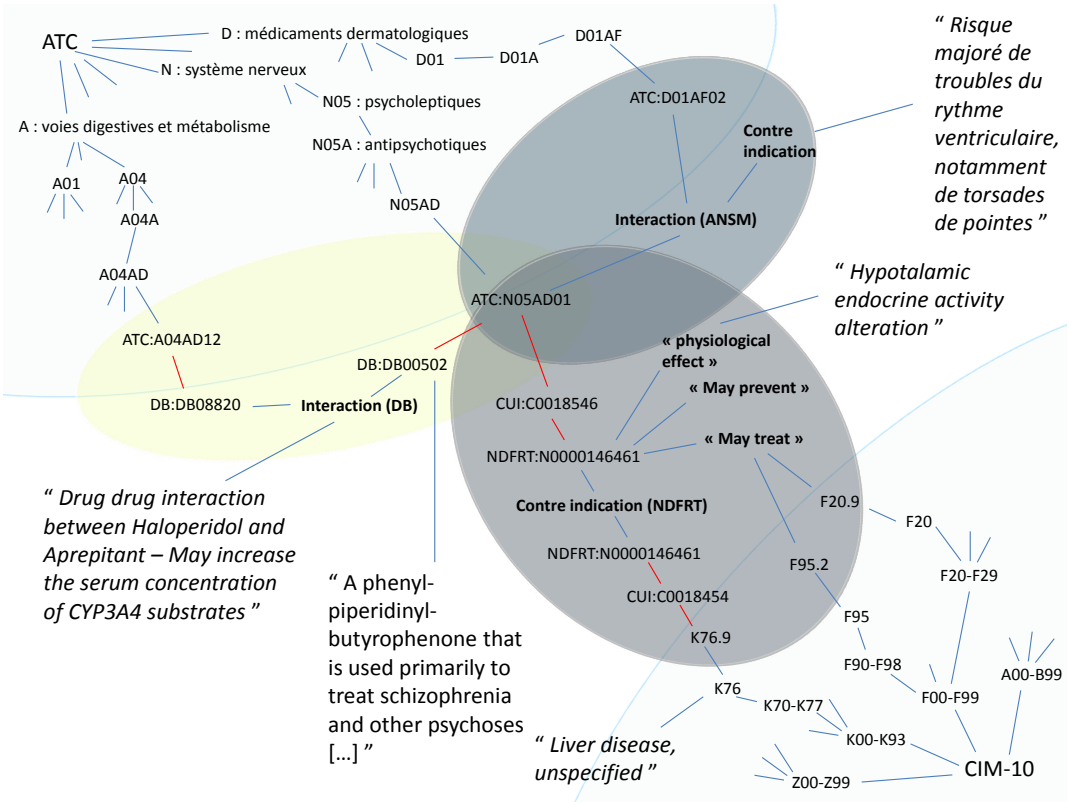

FIGURE 4.5 – Extrait du graphe de connaissance résultant d'une agrégation de bases de connaissances et d'ontologies du *Linked Data*, pour une exploration des trajectoires de soins basée sur des connaissances externes. Les traits rouges représentent des relations de correspondance –*mapping*– entre concepts similaires de différentes nomenclatures.

Ces relations décrivent une partie des connaissances du domaine pharmacologique et médical par les relations entre leurs concepts. Elles peuvent être des relations hiérarchiques, ou bien des relations décrivant un lien médical entre deux

concepts, tels que les interaction médicamenteuses, les indications de médicaments, les contre-indications médicamenteuses, ou encore les contre-indications entre médicaments et diagnostics. Les médicaments et diagnostics peuvent ainsi se trouver reliés entre eux. Cependant, ce graphe ne contient pas encore de liaison entre des actes et des concepts autres, tels que les diagnostics ou les médicaments. À notre connaissance, il n'existe pas d'ontologie décrivant les actes médicaux de la CCAM par des relations avec d'autres concepts tels que des diagnostics ou médicaments. Étant une nomenclature française, il n'existe pas non plus de correspondance entre la CCAM et une nomenclature internationale qui pourrait elle être reliée à une telle ontologie.

4.1.5 SPARQL pour une exploration des trajectoires de soins basée sur l'apport de connaissances d'ontologies biomédicales

Jusqu'ici, l'exploration de données médico-administratives utilise les technologies du Web Sémantique mais ne nécessite pas encore l'intégration d'ontologies et bases de connaissances du *Linked Data*. Cette sous-section présente certaines explorations rendues possibles grâce à l'utilisation de ces technologies et à l'intégration de graphes RDF de trajectoires de soins couplée à des graphes de connaissances (en RDFS et OWL) eux même reconstitués à partir de plusieurs ontologies et bases de connaissances du *Linked Data* comme présenté précédemment. Notamment, l'intégration d'ontologies telles que NDF-RT, le thésaurus de l'ANSM, DrugBank, DID ou encore DIKB, et les technologies du Web Sémantique, permettent de rendre systématique la recherche de certains événements dans les trajectoires de soins, des interactions médicamenteuses, des contre-indications, ou encore des indications de médicaments. L'intégration de ces connaissances peut ainsi permettre de répondre aux questions évoquées en section 4.1.1.

4.1.5.1 Interactions médicamenteuses

La base de connaissances DrugBank et le thésaurus des interactions médicamenteuses de l'ANSM (en RDFS), ont permis de rechercher des interactions médicamenteuses dans les trajectoires de soins des 1700 patients opérés pour une arthroplastie en 2012. Ces recherches se sont traduites par une requête SPARQL pour chaque source, une pour le thésaurus et une pour DrugBank. Une recherche équivalente pourrait être menée avec DIKB, qui contient des sources supplémentaires à DrugBank et au thésaurus de l'ANSM. Les résultats de la requête utilisant DrugBank pour la recherche d'interactions médicamenteuses dans un intervalle d'un mois sont représentés par la figure 4.6.

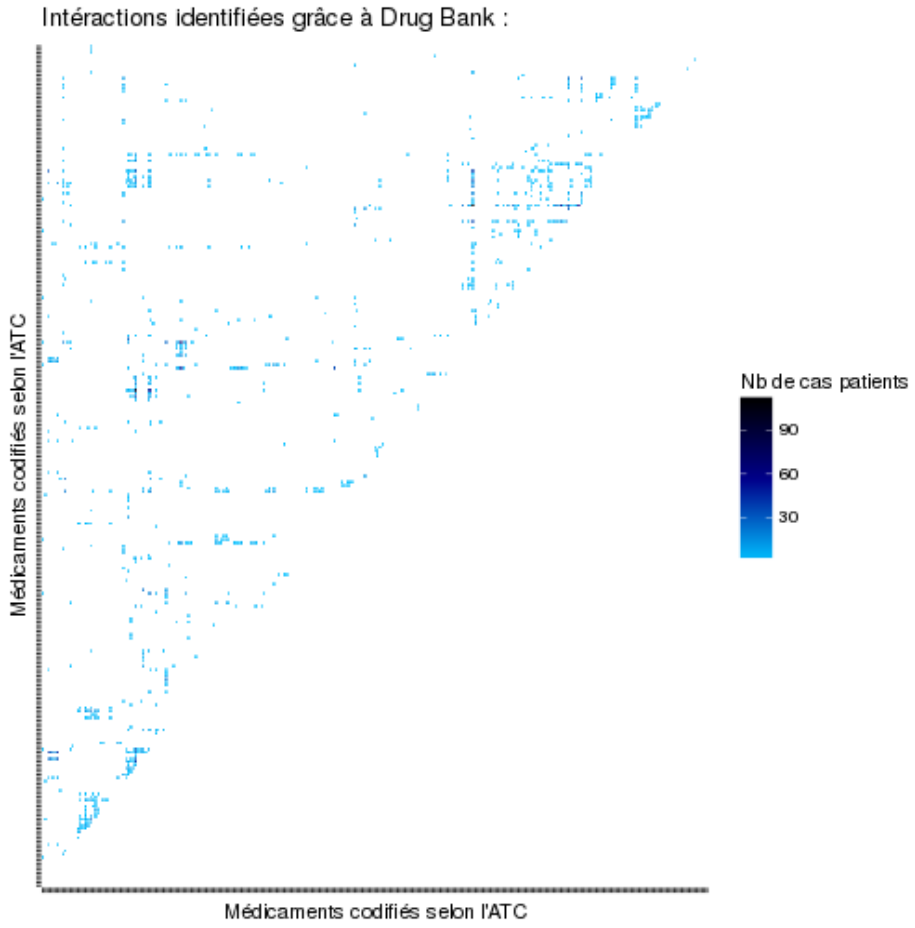

FIGURE 4.6 – Interactions médicamenteuses identifiées grâce à DrugBank. Le nombre important d'interactions détectées ne permet pas d'afficher les labels des médicaments. Ces derniers sont triés par ordre alphabétique de leur codage ATC, en abscisse et en ordonnée.

D'après les spécialistes cliniciens, ces très nombreuses interactions coïncident avec l'importante comorbidité des patients atteints d'arthroplastie. Relativement âgés, ces patients sont souvent polypathologiques. De fait, de nombreux médicaments leurs sont administrés pour le traitement de ces différentes maladies. Les risques d'interaction peuvent alors être importants.

4.1.5.2 Contre-indications médicament-médicament

La quatrième niveau de sévérité d'une interaction médicamenteuse dans le thésaurus de l'ANSM correspond à une contre-indication. Une requête SPARQL peut alors permettre de rechercher les interactions médicamenteuses contre-indiquées. Les résultats de cette requête pour la recherche de contre-indications dans un intervalle d'un mois, pour les 1700 patients opérés en 2012 pour une arthroplastie, sont synthétisés par la figure 4.7.

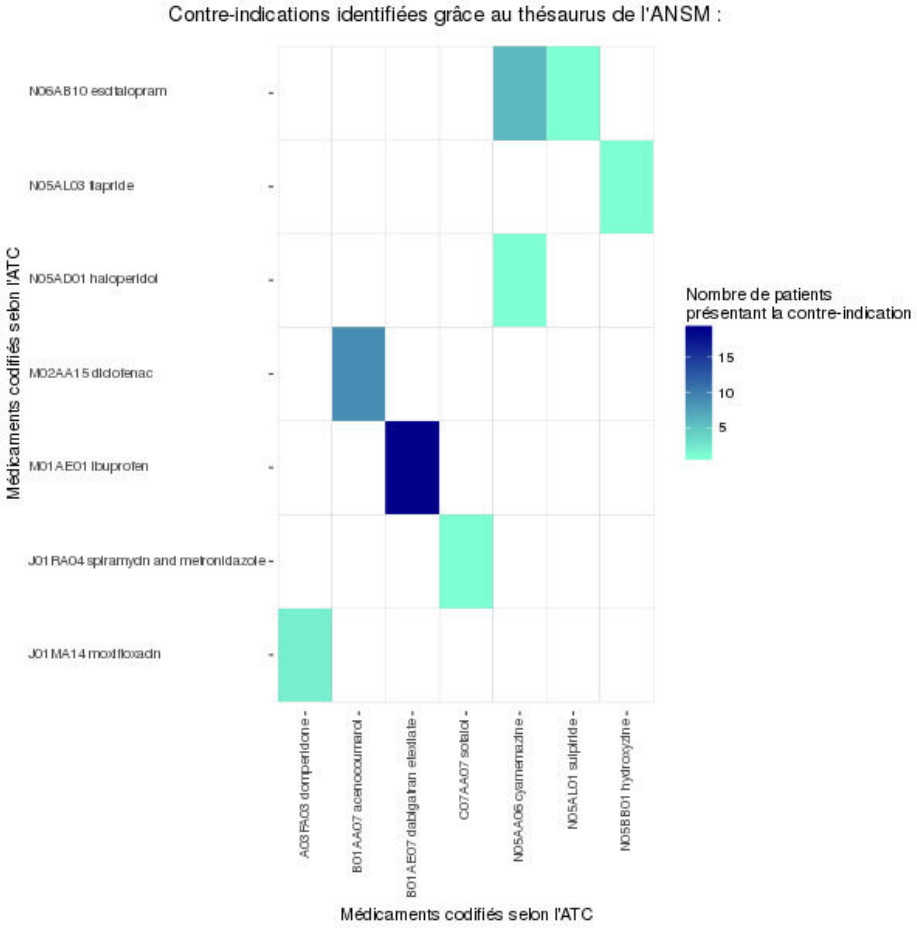

FIGURE 4.7 – Interactions médicamenteuses contre-indiquées par l'ANSM, détectées dans les données de l'EGB à partir d'un échantillon de 1700 patients opérés pour une arthroplastie. À l'intersection de deux médicaments contre-indiqués est représenté le nombre de patients présentant l'interaction contre-indiquée dans un intervalle d'un mois.

Bien que les interactions soient fréquentes dans cette population, elles ne sont pas toutes contre-indiquées, d'après l'ANSM. De plus, l'intervalle assez large d'un mois ne permet pas réellement d'assurer que les patients consomment les deux médicaments simultanément.

4.1.5.3 Contre-indications médicament-diagnostic

L'ontologie NDF-RT, dans la description de médicaments et maladies, définit des relations de contre-indication entre ces concepts. Les médicaments peuvent être associés à des codes ATC et les maladies à des codes de la CIM-10. Une requête SPARQL adaptée (code source 9) permet alors de rechercher de façon systématique les contre-indications éventuelles entre un diagnostic réalisé et une délivrance d'un médicament prescrit et remboursé.

```
PREFIX [...]
SELECT DISTINCT *
WHERE {#Contre-indications dans NDF-RT :
  ?ndf_med rdfs:subClassOf ?CI . #Médicament NDF
  ?CI owl:onProperty ndf:CI_with .
  ?CI owl:someValuesFrom ?ndf_diag . #Diagnostic ou état de santé NDF
  #En plus de ?ndf_diag et ?ndf_med, toutes leurs sous-classes supportent la CI :
  ?subclass_ndf_diag rdfs:subClassOf* ?ndf_diag .
  ?subclass_ndf_med rdfs:subClassOf* ?ndf_med .
  #Correspondance à l'ATC et à la CIM-10 via les CUI :
  ?subclass_ndf_diag ndf:UMLS_CUI ?cui_diag . #CUI
  ?ATC umls:cui ?cui_med . #ATC
  ?subclass_ndf_med ndf:UMLS_CUI ?cui_med . #CUI
  ?CIM10 umls:cui ?cui_diag . #CIM-10
  #En plus de ?ATC et ?CIM10, toutes leurs sous-classes supportent la CI :
  ?atc_med rdfs:subClassOf* ?ATC .
  ?cim10_diag rdfs:subClassOf* ?CIM10 .
  #On peut maintenant chercher les patients qui ont ?atc_med et ?cim10_diag :
  #Prestations qui mènent à la délivrance d'un médicament :
  ?patient :has_presta ?presta .
  ?presta :has_date ?date_presta .
  ?presta :has_cip_13 ?cip13 . #Délivrance d'un médicament, codé en CIP13
  ?cip13 :cip2atc ?atc_med . #Correspondance ATC
  #Hospitalisations et leurs diagnostics principaux, reliés ou associés
  ?patient :has_hosp_stay ?hospit .
  ?hospit :has_date ?date_hospit .
  {?hospit icd10:has_dp ?cim10_diag .} UNION {?hospit icd10:has_dr ?cim10_diag .}
  UNION {?hospit icd10:has_das ?cim10_diag .}
  FILTER(?date_hospit-?date_presta <= 'P30D'^^xsd:duration) #Contraintes temporelles
  FILTER(?date_presta-?date_hospit <= 'P30D'^^xsd:duration)}
```

Code source 9: Requête SPARQL pour la recherche de contre-indications (CI) selon NDF-RT dans un graphe RDF de données médico-administratives. Les correspondances avec des CUI permettent de relier des codes CIM-10 et ATC aux nomenclatures de NDF-RT, mais compliquent la requête du fait de plusieurs hiérarchies à prendre en considération. Des précautions doivent notamment être prises pour gérer l'extension par transitivité des contre-indications aux sous-classes de médicaments et diagnostics et n'en rater aucune, et ce pour chaque nomenclature utilisée.

Cette exploration, bien plus complexe que celles en section 4.1.3, illustre bien la nécessité d'intégrer des connaissances externes aux données, ainsi que de technologies adaptées à leur représentation et exploration, en l'occurrence RDF, RDFS, OWL et SPARQL. Les résultats de cette requêtes, pour les 1700 patients opérés d'une arthroplastie en 2012, sont représentés par la figure 4.8.

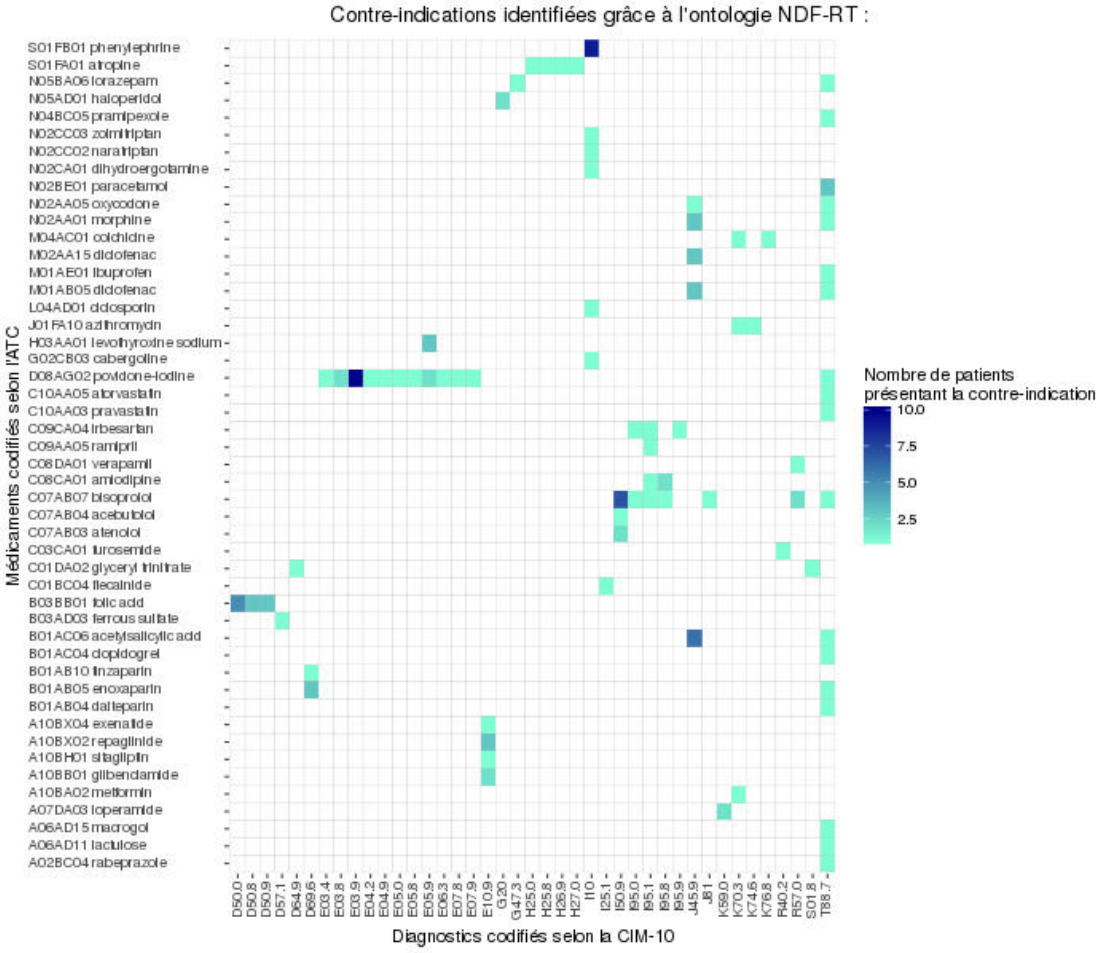

FIGURE 4.8 – Contre-indications médicament-diagnostic détectées via une requête SPARQL, grâce à l'apport des connaissances issues de l'ontologie NDF-RT. À l'intersection d'un diagnostic et d'un médicaments est représentée le nombre de patients présentant la contre-indication dans un intervalle d'un mois.

4.1.5.4 Indications des médicaments

Moins particuliers, car beaucoup plus attendues, des relations d'indications entre médicaments et états de santé peuvent être recherchées dans les trajectoires de soins grâce à l'intégration de NDF-RT (relations *may_treat* et *may_prevent*) ou de DID (NDF-RT étant présent dans DID). Il devient alors également possible de rechercher des absences d'indications pour un état de santé étudié, ou encore les prescriptions de médicaments qui ne sont pas indiquées dans le cadre d'une maladie. La recherche d'absence de relation est cependant plus délicate. Nous présentons un exemple dans la section 4.2.

4.1.6 Synthèse

Pour une exploration standard des données médico-administratives, les technologies du Web Sémantique ne sont pas plus adaptées que d'autres méthodes souvent plus populaires, telles que SQL ou que de l'exploration de données avec R. En revanche, ces dernières technologies ne sont pas adaptées pour traiter des requêtes plus complexes, comme l'exploration de trajectoires de soins, qui sont rendues nécessaires par l'évolution de la pharmaco-épidémiologie. Le travail présenté dans cette section montre que les technologies du Web Sémantique ont l'expressivité requise tout en permettant le passage à l'échelle. Adaptées à l'utilisation d'ontologies, c'est d'autant plus dans l'exploration de données nécessitant l'utilisation de connaissances externes que les technologies du Web Sémantique montrent tout leur intérêt par rapport aux méthodes classiques d'exploration et de gestion de données. Elles permettent notamment de répondre à des questions pertinentes en pharmaco-épidémiologie et pharmacovigilance.

Cependant, les nomenclatures médicales et représentations de la connaissance médicale et pharmacologique sont de plus en plus nombreuses, complexes, et difficiles à utiliser. L'exploration de trajectoires de soins des bases médico-administratives, grâce à SPARQL et à l'utilisation d'ontologies du *Linked Data*, reste une tâche longue et compliquée pour les pharmaco-épidémiologistes. C'est sans doute pour cette raison que l'exploration de données médico-administrative est aujourd'hui encore majoritairement envisagée par le biais de requêtes SQL, rendant de fait l'exploration basée sur l'apport de connaissance peu répandue.

Certaines avancées et développements de packages R ont considérablement amélioré l'expressivité et l'efficacité d'une exploration avec R. Les packages *data.table* (Dowle and Srinivasan, 2018) et *dplyr* (Wickham et al., 2018) permettent par exemple de gérer des données plus volumineuses, de les explorer plus rapidement et plus facilement, qu'ils en deviennent de vraies alternatives à SQL. La liberté de programmation qu'offre R, ces dernières avancées, et le fait que cet outil soit de plus en plus utilisé par la communauté des chercheurs en santé publique, nous a amené à proposer un package R pour faciliter l'utilisation de connaissances médicales et pharmacologiques du *Linked Data* dans l'exploration des bases de données médico-administratives française à des fins de recherche en santé publique.

4.2 Faciliter la réutilisation des connaissances médicales pour l'exploration et l'analyse de données médico-administratives avec R

4.2.1 Introduction

Si la pertinence de lier des systèmes d'organisation de la connaissance à des bases de données a pu être démontrée, notamment pour baser leur exploration et leur analyse sur les connaissances d'un domaine, dans la pratique on constate que cette approche est rarement mise en œuvre en santé publique, en épidémiologie ou en pharmaco-épidémiologie. Malgré le cadre technique qu'offre les recommandations du W3C, peu de projets de recherche en santé publique basent l'exploration de leurs données médicales sur les technologies du Web Sémantique et les ontologies médicales du *Linked Data*. Selon Jain et al. (2010), c'est le nombre croissant d'ontologies, l'hétérogénéité de leurs schémas de représentation, et un manque de leurs descriptions qui rendent leur utilisation compliquée, particulièrement l'écriture de requêtes SPARQL qui dépendent des schémas des ontologies. Enfin, la multiplicité des technologies permettant de récupérer la connaissance médicales et pharmacologique sur le Web ajoute une difficulté. En effet, si une grande partie des ontologies médicales sont disponibles via des SPARQL endpoint, d'autres méthodes et protocoles permettent de récupérer des connaissances d'ontologies sur le Web. BioPortal propose par exemple plusieurs outils sous la forme d'API REST, alors que son sparql endpoint, en version beta, est déconseillé dans le développement d'outil ou de réalisation d'études.

Néanmoins, plusieurs travaux ont contribué à réduire les difficultés propres aux ontologies médicales du *Linked Data*. Le *Concept Unique Identifier* (CUI) du *Unified Medical Language System* (UMLS) a par exemple largement été utilisé dans les ontologies médicales pour établir des correspondances entre les nomenclatures les plus utilisées. Enfin, d'autres travaux ont rassemblé des sources de connaissances médicales et pharmacologiques similaires. La *Drug Indication Database* (DID) (Sharp, 2017) a par exemple rassemblé plus d'une dizaine de sources de connaissances du Web des données concernant les indications de médicaments. De la même façon, la *Drug Interaction Knowledge Base* (DIKB) (Ayvaz et al., 2015) a rassemblé elle aussi plus d'une dizaine de sources de connaissances relatives aux interactions médicamenteuses. Malgré ces travaux pour rendre les connaissances médicales et pharmacologiques plus facilement réutilisables, leur intégration dans une base de données médicales reste une tâche laborieuse pour un non-expert. Dans ce sens, Ferreira et al. (2013) ont souligné l'importance pour les épidémiologistes de disposer d'outils facilitant cette tâche, pour que les méthodes basée sur les connaissances médicales et pharmacologiques se répandent dans les champs de la santé publique, notamment de la pharmaco-épidémiologie et de l'épidémiologie. Plusieurs projets ont ainsi pour objectif de rendre l'utilisation des technologies du Web Sémantique plus simple (Yamaguchi et al., 2014; Bettembourg et al., 2015), par exemple en proposant des outils intéractifs pour la création de requêtes SPARQL pour des bases de données RDF

par des non experts de ces technologies. Puisque l'exploration de données est fortement liée à l'analyse des données, de nombreux outils (Kurbatova and others, 2011; Hage and others, 2013; Willighagen, 2014) pour rendre les technologies du Web Sémantique disponibles à travers le logiciel et langage de programmation R (R Core Team, 2017) ont également été développés. Cependant, ces outils ne traitent pas de la difficulté à utiliser des ontologies ou les technologiques du Web Sémantique, mais les rendent seulement accessibles dans un environnement dédié à la programmation à des fins statistiques, R. Si l'envoi de requêtes SPARQL sur des serveurs distants ou sur des données RDF importées depuis R est ainsi possible, leur réalisation reste encore difficile pour un non-expert des ontologies et de leurs schémas de représentation.

4.2.2 Objectifs

L'objectif du package R *queryMed* (Rivault et al., 2018c) est de fournir un outil visant à faciliter l'utilisation de connaissances médicales et pharmacologiques pour de l'exploration et de l'analyses de données de santé, notamment des données médico-administratives.

4.2.3 Méthodes

Pour que cet outil soit accessible au plus grand nombre de statisticiens et épidémiologistes, notre choix s'est tourné vers un outil pouvant s'intégrer dans l'environnement de statistique R, de plus en plus utilisé par la communauté des épidémiologistes. Une des premières fonctions du package, `sparql(requête SPARQL, url du serveur)`, permet d'envoyer des requêtes SPARQL sur les serveur distants hébergeant des données aux standards du Web Sémantique et autorisant leur requêtage, les SPARQL endpoints. Cette première fonction requiert une expertise dans les ontologies et données RDF disponibles, ainsi que la connaissances des SPARQL endpoints. queryMed propose d'autres fonctions qui encapsulent des requêtes prédéfinies adaptées aux principaux SPARQL endpoints du champs médical. Le package offre ainsi des requêtes pour les serveurs de Bio2RDF (Callahan et al., 2013), DBpedia (Lehmann et al., 2015), et Ontobee (Ong et al., 2017). Ces requêtes permettent d'obtenir des annotations sur des codes de diagnostics et médicaments, telles que des définitions, des résumés, des labels, des synonymes, des traductions ou encore des correspondances entre termes de différentes nomenclatures. De telles fonctions permettent également de récupérer des annotations plus complexes, reliant plusieurs codes médicaux entre eux, tels que des contre-indications entre médicaments et diagnostics.

De façon similaire, une fonction de base du package permet d'envoyer des requêtes sur deux API REST de BioPortal. `annotator(texte)` permet d'annoter du texte par des classes d'ontologies présentes sur BioPortal. `search(termes, ontologies)` permet de récupérer une partie des connaissances relatives à des termes ou à des codes médicaux, dans les ontologies de BioPortal.

La fonction `search()` permet notamment d'associer un CUI à des termes de nomenclatures médicales, quand ces derniers existent et sont disponibles. La fonction `mapping_cui(codes, source, cible)`, sur la base de `search()`, permet alors de faciliter la récupération de correspondances entre code de nomenclatures différentes, en utilisant la notion de CUI. De plus, le package intègre les bases de données DID et DIKB, ajoutant des connaissances sur les indications de médicaments et leurs potentiels interactions.

Si le package permet de récupérer des annotations simples, comme des labels, elle permet également de récupérer des relations plus complexes incluant deux codes ou deux termes. Les indications d'un médicaments en sont un exemple. Elles relient un médicaments et les états de santé pour lesquels le médicament est indiqué. La fonction `find_relations(...)` permet alors de rechercher de telles relations impliquant plusieurs concepts médicaux, dans une base de données médicales de patients, une fois les relations récupérées et traduites dans les bonnes nomenclatures.

4.2.4 Résultats et applications

Dans le cadre d'une première application du package, nous nous sommes intéressés à l'annotation de trajectoires de soins d'un ensemble de 1003 patients hospitalisés pour une artériopathie oblitérante des membres inférieurs (AOMI) et opérés par angioplastie, durant l'année 2015. Les données du SNIIRAM, prescriptions de médicaments, diagnostics principaux, reliés et associés, ont permis de constituer les trajectoires de soins des patients. L'application s'est concentré sur la recherche de médicaments indiqués et contre-indiqués dans le cadre de l'AOMI. Elle s'est divisé en trois étapes (figure 4.9) :

1. Récupération des contre-indications médicaments-diagnostics présents dans NDF-RT, sur Ontobee, et des indications dans DID, en tenant compte des hiérarchies ;
2. Correspondance des médicaments et diagnostics vers les CUI et NDF-RT ;
3. Utilisation de `find_relations()` pour rechercher les relations de contre-indication et d'indication entre médicaments et diagnostics, pour chaque patient.

queryMed a ainsi permis de détecter 72 patients, qui selon DID, n'ont pas de prescription de médicament indiqué dans le cadre d'une AOMI. S'il est aisé d'interpréter une présence, une absence l'est beaucoup moins : ces patients peuvent en effet manquer d'une prescription de médicament, tout comme DID peut être loin de l'exhaustivité en ce qui concerne les médicaments indiqués dans le cadre de l'AOMI. Ils peuvent également avoir eu une délivrance de médicament à l'hôpital, qu'une étude utilisant des données du SNIIRAM ne pourrait détecter.

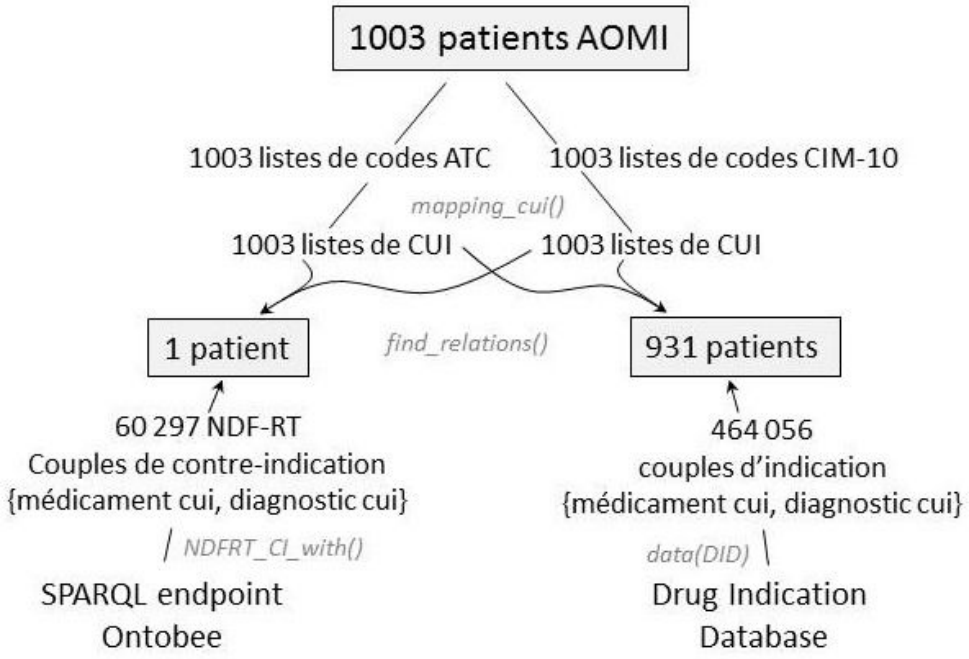

FIGURE 4.9 – Application de querymed sur 1003 trajectoires de soins de patients opérés par angioplastie pour une AOMI. Les appels de fonction sont en gris.

Selon NDF-RT, 91 médicaments sont contre-indiqués dans le cadre de la maladie. Parmi les 1003 patients, queryMed a permis de détecter un patient avec une contre-indication médicament-diagnostic. En effet, le patient a reçu une délivrance d'un vasoconstricteur, un médicament qui agît à réduire la lumière des vaisseaux sanguins, de ce fait contre-indiqué dans le cadre de l'AOMI, maladie qui se caractérise déjà par la diminution du diamètre des artères irriguant les membres inférieurs.

4.2.5 Codes sources et vignette

Les codes sources du package sont disponibles sur `https://github.com/yannrivault/queryMed`. De plus, pour que les utilisateurs du package puissent tester une application similaire à celle présentée en sous-section 4.2.4, le package contient une vignette R (annexe 6). Ce document décrit les étapes à réaliser, sur un jeu de données simulées, pour rechercher les contre-indications entre des médicaments et diagnostics, les interactions médicamenteuses ainsi que les indications de médicaments. Il couvre la majeure partie des fonctions disponibles dans le package, de l'annotation simple de données médicales jusqu'à la correspondances entre codes, et à la recherche de relations sémantiques complexes dans un jeu de données de trajectoires de soins.

4.3 Synthèse

Ce premier aspect de la thèse a pu démontrer l'intérêt de baser une exploration des données médico-administratives sur un apport de connaissances issues du *Linked Data*. Les technologies du Web Sémantique ont montré une expressivité suffisante pour la représentation des données ainsi que pour leur exploration. Couplées à l'utilisation d'ontologies médicales et pharmacologiques, elles permettent en effet d'intégrer des connaissances aux données, et d'explorer efficacement des trajectoires de soins sur des critères complexes. L'approche permet ainsi d'intégrer dans l'exploration les dimensions hiérarchiques des nomenclatures utilisées dans les bases médico-administratives. Elle a également permis d'identifier des trajectoires comportant des événements ou relations particulières entre leurs éléments, tels que des interactions médicamenteuses, des contre-indications entre médicaments ou entre médicaments et maladies, ou encore des relations d'indication entre médicaments et maladies. Durant cette thèse, cette approche a été présentée à plusieurs reprises sous la forme d'articles et de communication dans des congrès (Rivault et al., 2015, 2016a, 2018b,a).

Face à une certaine réticence à mettre en œuvre une telle approche pour l'exploration des données médico-administratives, nous avons proposé l'outil *queryMed* (Rivault et al., 2018c). Ce package R facilite l'accès aux connaissances médicales et pharmacologiques du *Linked Data*, leur liaison avec des données de santé, et la recherche de relations entre concepts, comme des indications, des contre-indications et des interactions, dans de telles données. Il rend ainsi plus accessible l'approche d'exploration de données médico-administratives utilisant des ontologies du *Linked Data*. Il a été présenté par des articles courts (Rivault et al., 2018b,a) associés à des communications dans des congrès francophones. Une note d'application a également été soumise au comité de lecture de la revue Bioinformatics, et est disponible en annexe 6

Chapitre 5

Enrichissement de l'analyse de trajectoires de soins par connaissances

Sommaire

Dans l'étude de trajectoires de soins, un traitement de données plus poussé qu'une exploration est très souvent réalisé. On parle alors d'analyse de trajectoires de soins. Bien que l'analyse des données médico-administratives soit aujourd'hui systématique dans la réutilisation scientifique de ces données, leur vision holistique en trajectoires de soins ne l'est pas autant. En effet, les méthodes d'analyse de données médico-administratives empruntées à la statistiques, ne permettent que rarement de considérer les données patient dans leur globalité, par exemple sous la forme de séquences. Les méthodes de fouille de données et de comparaison (ou alignement) de séquences sont des méthodes qui gagnent en popularité dans l'analyse des données médico-administratives, justement car elles permettent d'analyser des trajectoires de soins.

Il existe cependant certaines limites à ces méthodes qui se font d'autant plus ressentir dans la réutilisation des données médico-administratives. Considérer les trajectoires de soins comme la concaténation de médicaments, de diagnostics et d'actes médicaux mène généralement à des calculs longs et compliqués. Associés à cette limite, les résultats de ces méthodes, et notamment ceux de la fouille de données, sont souvent eux aussi complexes, surtout trop volumineux pour qu'ils soient compréhensibles. Constituer les trajectoires de soins de codes de médicaments, de diagnostics et actes médicaux, rend les trajectoires très variables, très différentes les unes des autres, notamment du fait de la profondeur des nomenclature utilisées. Si cette variabilité est importante à considérer, les méthodes classiques de fouille et d'analyse de trajectoires ne considèrent pas les similarités possibles entre deux éléments différents, par exemple deux médicaments différents du point de vue de leur représentation dans les données (i.e. leur codage) mais proches du point de vue de leurs actions.

Ces limites sont généralement contournées par des techniques de pré-traitement. Il est ainsi d'usage de regrouper certains éléments d'une trajectoire en un groupe d'éléments (en un événement de santé par exemple) ou de réduire les séquences à un nombre restreint d'éléments d'intérêt. Également, il est assez fréquent de représenter des médicaments, diagnostics ou actes médicaux par un de leurs ascendants dans les hiérarchies des nomenclatures les codant. Le degré de granularité des éléments des trajectoires est ainsi réduit pour contourner le problème de trajectoires trop variables. Si ces pré-traitements sont forcément nécessaires lors de l'analyse de trajectoires de soins, nous présentons dans ce chapitre des modifications de certaines méthodes d'analyse et de fouille de trajectoires de soins, afin de réduire leur besoin.

La section 5.1 présente un travail d'analyse de trajectoires de soins reposant sur

leur comparaison. Nous y présentons l'adaptation de la notion de la plus grande sous-séquence commune aux formalisme des séquences d'ensembles, ainsi qu'à l'intégration de la prise en compte de hiérarchies de nomenclatures médicales à l'aide de similarités sémantiques.

Dans la section 5.2, nous présentons une modification d'une méthode de fouille de données, l'extraction de règles d'associations. Appliquée aux trajectoires de soins, nous proposons d'étendre l'extraction de règles d'associations aux règles d'associations multi-niveaux, qui peuvent alors se constituer de médicaments, diagnostics et actes médicaux à plusieurs degrés hiérarchiques.

La section 5.3 présente une piste pour l'exploration de résultats d'une extraction de règles d'associations à partir de trajectoires de soins. S'appuyant sur l'approche présentée en chapitre 4, nous proposons d'explorer des règles d'associations grâce à aux technologies du Web Sémantique et à des connaissances du *Linked Data* afin de les filtrer ou de sélectionner des sous-ensembles de règles d'intérêt.

Enfin, dans la section 5.4, dans le but de rechercher dans une bases de données de trajectoires de soins des motifs rares qui seraient fournis par des experts, nous testons les technologies du Web Sémantique pour de la reconnaissance de chroniques.

5.1 Comparaison de trajectoires de soins

5.1.1 Introduction

Les méthodes de comparaison de séquences sont de plus en plus utilisées sur des trajectoires de soins pour mesurer des similarités ou dissimilarités entre trajectoires et ainsi les discrétiser en groupes homogènes de trajectoires. Cela permet ainsi de regrouper des patients selon leurs trajectoires de soins, c'est-à-dire selon leurs consommations de soins et états de santé. Cependant, considérer les trajectoires de soins comme des traces linéaires d'événements de santé, par exemple les diagnostics, les actes médicaux et les délivrances de médicaments, peut parfois être un peu simpliste. En effet, des actes médicaux et diagnostics peuvent être regroupés au sein d'une hospitalisation, et plusieurs médicaments délivrés peuvent être issus d'une seule et même prescription.

En outre, la profondeur des nomenclatures utilisées pour codifier les éléments des trajectoires de soins induit une importante variabilité des données et donc des trajectoires elles-même. Cette variabilité peut rendre les méthodes classiques de comparaison de trajectoires trop strictes. Deux médicaments différents, quand bien même ils pourraient être similaires de par leurs compositions chimiques, actions ou indications, seront souvent jugés comme étant différents si leur codage l'est. En ne prenant pas en compte les similarités entre événements lors de la comparaison des trajectoires, les méthodes classiques surestiment la variabilité des trajectoires.

5.1.2 Objectifs

Afin de rendre la comparaison de trajectoires de soins plus adaptée au contexte des données médico-administratives, c'est à dire au codage des données et aux trajectoires constituées d'événements de santé, nous proposons d'apporter des modifications à une méthode classique de comparaison de séquence : la notion de plus grande sous-séquence commune. Les objectifs de ces modifications étaient les suivantes :

— Proposer un formalisme de trajectoires de soins qui prenne en compte l'aspect multidimensionnel des éléments en événements de santé ;

— Adapter la notion de la plus grande sous-séquence commune à ce formalisme ;

— Intégrer dans la méthode la hiérarchie des nomenclatures utilisées pour coder les données médico-administratives afin de mieux estimer la variabilité des trajectoires de soins.

5.1.3 Formalisme des séquences d'ensembles

Le formalisme des séquences d'ensembles (*sequences of itemsets* dans la litérature anglophone) est un sujet de recherche assez récent dans l'analyse (Egho et al., 2015) et la fouille (Plantevit et al., 2010; Jay and d'Aquin, 2013) de séquences. Ainsi, les travaux comparant des trajectoires de soins voient souvent les trajectoires comme des séquences simples d'événements médicaux (figure 5.1). Les événements sont parfois des médicaments, des actes médicaux, ou encore des diagnostics.

FIGURE 5.1 – Trajectoire de soins vue comme une séquence simple composée d'événements médicaux.

C'est le nombre important de méthodes pour la comparaison de séquences simples (Navarro, 2001; Studer and Ritschard, 2016) qui a mené à considérer les trajectoires de soins en tant que séquences simples plutôt qu'en séquences d'ensembles. Pourtant, les habituels pré-traitements pour regrouper des éléments d'une trajectoires en un événement de santé illustrent bien le caractère multi-dimensionnel des trajectoires de soins. Dans ce travail, nous considérons les trajectoires de soins comme des séquences d'ensembles (figure 5.2), dans le but de mieux capturer la simultanéité de certaines éléments de soins et de s'affranchir de pré-traitements visant à regrouper des éléments en un événement.

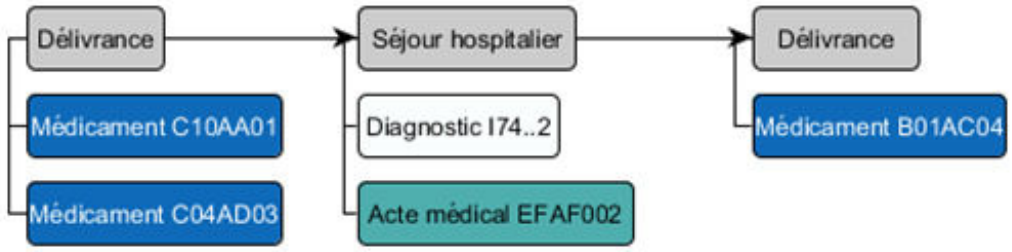

FIGURE 5.2 – Trajectoire de soins vue comme une séquence d'ensembles. Les ensembles constitués d'un ou de plusieurs codes représentent une hospitalisation ou une délivrance. Les données d'une hospitalisation ou d'une délivrance ne permettent pas d'ordonner dans le temps de façon systématique leurs éléments. Le formalisme n'intègre donc pas d'ordonnancement temporel entre les éléments d'un ensemble.

Dans le but de comparer des trajectoires deux-à-deux, nous étudions dans la sous-section suivante la notion de plus grande sous-séquence commune, que nous étendons à ce formalisme.

5.1.4 Généralisation de la notion de plus grande sous-séquence commune au formalisme des séquences d'ensembles

La notion de plus grande sous-séquence commune (Hirschberg, 1975), en découvrant les plus longs motifs partagés par deux séquences, peut permettre de quantifier la partie commune de deux trajectoires de soins. En particulier, des mesures de similarités entre trajectoires de soins peuvent se baser sur la taille de ces plus grands motifs partagés. Nous nous intéressons donc au calcul de la taille d'une plus grande sous-séquence commune à deux trajectoires de soins, lorsqu'elles sont considérées

comme des séquences d'ensembles. Les définitions et théorème ci-après permettent la définition d'un algorithme pour le calcul de cette mesure.

Définition 13 *Séquence d'ensembles*
Une séquence d'ensembles $X = (x_1, x_2, ..., x_m)$ *est une séquence dont les éléments* x_i *sont des ensembles non-vides.*

Définition 14 *Taille d'une séquence d'ensembles*
La taille d'une séquence d'ensemble $X = (x_1, x_2, ..., x_m)$ *est définie de la façon suivante :*

$$|X| = \sum_{i=1}^{m} |x_i|$$

Définition 15 *Sous-séquence d'une séquence d'ensembles*
Étant données deux séquences d'ensembles $X = (x_1, x_2, ..., x_m)$ *et* $Y = (y_1, y_2, ..., y_n)$, *avec* $m \leq n$, X *est une sous-séquence de* Y *s'il existe les indices* $1 \leq j_1 < j_2 < ... < j_m \leq n$ *tels que* $x_i \subseteq y_{j_i}$ *pour tout* $i = 1, 2, ..., m$.

Définition 16 *Plus grande sous-séquence commune de deux séquences d'ensembles*
Étant données deux séquences d'ensembles $X = (x_1, x_2, ..., x_m)$ *et* $Y = (y_1, y_2, ..., y_n)$, Z *est une séquence commune à* X *et à* Y *si elle est à la fois une sous-séquence de* X *et de* Y. Z *est une plus grande sous-séquence commune à* X *et à* Y *si* $|Z| \geq |Z'|$, *pour toute autre sous-séquence commune* Z' *à* X *et à* Y.

Par la suite, les notation suivantes sont adoptées :

$CS(X, Y)$	l'ensemble des sous-séquences communes aux séquences d'ensembles X et Y ;
$LCS(X, Y)$	l'ensemble des plus grandes sous-séquences communes aux séquences d'ensembles X et Y ;
X_i	le i-préfixe de X, $X_i = (x_1, x_2, ..., x_i)$ avec $1 \leq i \leq m$.

Lemme 1 *Soient* $X = (x_1, x_2, ...x_m)$ *et* $Y = (y_1, y_2, ..., y_n)$ *deux séquences d'ensembles, et* $Z = (z_1, z_2, ...z_k) \in LCS(X, Y)$ *de longueur* $|Z| = \sum_i^k |Z_i|$.
Alors :

1. *alors* $z_k = x_m \cap y_n$ *implique* $Z_{k-1} = (z_1, z_2, ..., z_{k-1}) \in LCS(X_{m-1}, Y_{n-1})$;
2. *si* $z_k \neq x_m \cap y_n$ *alors :*
   (a) $z_k \not\subseteq x_m$ *implique* $Z \in LCS(X_{m-1}, Y_n)$;
   (b) $z_k \not\subseteq y_n$ *implique* $Z \in LCS(X_m, Y_{n-1})$;

Preuve 1 *1 :* $Z_{k-1} \in CS(X_{m-1}, Y_{n-1})$. *Montrons que* Z_{k-1} *appartient également aux plus longues sous-séquences communes de* X_{m-1} *et de* Y_{n-1}, *autrement dit qu'il n'existe pas d'autres sous-séquence commune à* X_{m-1} *et* Y_{n-1} *dont la taille soit supérieure à celle de* Z_{k-1}.

Pour cela, supposons qu'il existe $W \in CS(X_{m-1}, Y_{n-1})$ *telle que* $|W| > |Z_{k-1}|$. *Si on créé* W' *par concaténation de* W *et de* $z_k = x_m \cap y_m$, *on a* $W' \in CS(X, Y)$. *De plus, on aurait* $|W'| = |W| + |x_m \cap y_n|$ *et donc* $|W'| > |Z_{k-1}| + |x_m \cap y_n|$ *soit* $|W'| > |Z|$ *(puisque* $z_k = x_m \cap y_n$*). Ceci contredirait le fait initial que* $Z \in LCS(X, Y)$. *De fait, il ne peut exister* $W \in CS(X_{m-1}, Y_{n-1})$ *telle que* $|W| > |Z_{k-1}|$. Z_{k-1} *appartient donc aux plus longues sous-séquence communes de* X_{m-1} *et de* Y_{n-1}.
2.(a) : si $z_k \notin x_m$ *alors* $Z \in CS(X_{m-1}, Y_n)$. *Or* $Z \in LCS(X_m, Y_n)$. *Il n'existe donc pas de sous-séquence commune à* X_m *et à* Y_n *dont la taille dépasse celle de* Z. *Il n'existe donc pas plus de sous-séquences commune à* X_{m-1} *et à* Y_n *dont la taille dépasse celle de* Z. Z *est donc une plus grande sous-séquence commune à* X_{m-1} *et à* Y_n.
2.(b) : symétrique à 2.(a).

Le lemme 1 permet de déduire un formule récursive pour le calcul de la taille de la plus grande sous-séquence commune, notée $N(X_i, Y_j) = |LCS(X_i, Y_j)|$.

Théorème 1 *Taille de la plus grande sous-séquence commune*

$$N(X_i, Y_j) = \begin{cases} 0 \text{ si } i = 0 \\ 0 \text{ si } j = 0 \\ max(N(X_i, Y_{j-1}), N(X_{i-1}, Y_j), N(X_{i-1}, Y_{j-1}) + |x_i \cap y_j|) \text{ sinon} \end{cases}$$

Ce théorème permet ensuite le calcul de la plus grande sous-séquence commune à deux séquences d'ensembles, par programmation dynamique (Bellman, 1954).

Algorithme 1 : Calcul de la plus grande sous-séquence commune à deux séquences d'ensembles par programmation dynamique

Données : Deux séquences d'ensembles $X = (x_1, x_2, ..., x_m)$ et $Y = (y_1, y_2, ..., y_n)$.
Résultat : La taille d'une plus grande sous-séquence commune à X et Y, soit $|LCS(X, Y)|$.

```
initialisation : N[(0, ..., m), (0, ..., n)] = 0;
pour i dans 1,2,...,m faire
    pour j dans 1,2,...,n faire
        N(i, j) = max(N(i, j − 1), N(i − 1, j), N(i − 1, j − 1) + |x_i ∩ y_j|)
```

Une mesure de similarités entre séquences d'ensembles peut alors reposer sur le calcul de la taille de la plus grande sous-séquence commune à deux séquences d'ensemble. Nous proposons d'utiliser la mesure de similarité suivante :

Définition 17 *Similarité entre trajectoires de soins*
Étant données deux séquences d'ensembles $X = (x_1, x_2, ..., x_m)$ *et* $Y = (y_1, y_2, ..., y_n)$, *nous définissons la mesure de similarité entre* X *et* Y *suivante :*

$$sim(X, Y) = \frac{|LCS(X, Y)|}{max(|X|, |Y|)}$$

Par exemple, nous considérons deux trajectoires de soins composées de codes de médicaments, de diagnostics et d'actes médicaux, et représentées en séquences d'ensembles :

$$S_1 = (\{C10AA01\}, \{EZAF001, I87.1, YYYY034\}, \{B01AC04\})$$

$$S_2 = (\{C10AA04\}, \{I87.1, EZAF001, YYYY200\}, \{B01AC04\})$$

La plus grande sous-séquence commune à S_1 et à S_2 est alors :

$$LCS(S_1, S_2) = (\{I87.1, EZAF001\}, \{B01AC04\})$$

Et nous pouvons calculer une mesure de similarité entre S_1 et S_2, basée sur la taille de leur plus grande sous-séquence commune :

$$sim(S_1, S_2) = \frac{LCS(S_1, S_2)}{max(S_1, S_2)} = \frac{3}{5} = 0,6$$

Adapter la notion de plus grande sous-séquence commune au formalisme de séquences d'ensembles fait suite à l'objectif de mieux considérer le caractère multi-dimensionnel des trajectoires de soins dans leur comparaison. En revanche, la méthode ne prend pas en compte les similarités ou proximités potentielles entre des éléments différents. Du fait de la grande profondeur des nomenclatures médicales, deux codes de médicaments, de diagnostics ou d'actes médicaux peuvent représenter des concepts proches quand bien même ces codes seraient différents. Dans notre exemple, les codes $C10AA01$ (simvastatine) et $C10AA04$ (fluvastatine) sont deux codes différents représentant des médicaments appartenant à la même famille des statines, aux effets et indications similaires. De la même façon, $YYYY034$ et $YYYY200$ sont deux actes de radiologie. On peut juger que la mesure obtenue égale à 0,6 sous-estime la similarité des deux trajectoires, en considérant de façon trop stricte la différence entre ces codes. Dans l'objectif de pallier à ce problème, nous proposons d'introduire dans la comparaison des trajectoires des similarités entre éléments, c'est à dire entre diagnostics, médicaments et actes médicaux.

5.1.5 Introduction des connaissances hiérarchiques grâce à l'introduction de similarités sémantiques

La notion de similarités sémantiques entre deux concepts peut avoir plusieurs définitions. Avec le développement des ontologies, elles ont été utilisées pour mesurer la ressemblance entre deux concepts. On distingue généralement les similarités sémantiques basées sur les arcs d'une hiérarchie commune à deux concepts, des similarités basées sur les nœuds des deux concepts. Les similarités basées sur les arcs mesurent ainsi la ressemblance entre concepts selon leur proximité géographique dans une hiérarchie commune, tandis que les similarités basées sur les nœuds l'a mesurent selon les annotations qu'ils peuvent avoir en commun. Le développement important d'ontologies biomédicales s'est accompagné de l'utilisation de nombreuses et différentes similarités sémantiques, basées sur les arcs ou les nœuds des concepts,

dans le but de mesurer des similarités entre gènes, médicaments ou encore maladies (Pesquita et al., 2009). Comme l'ont montré Girardi et al. (2016), l'utilisation de similarités sémantiques dans le calcul de similarités entre patients permet de rassembler des trajectoires de soins qui se ressemblent alors qu'une mesure classique les aurait jugé différentes.

Dans l'objectif de rendre une mesure de similarité basée sur la notion de plus grande sous-séquence d'ensembles plus robuste à la grande variabilité des événements des trajectoires de soins, nous nous sommes intéressés à introduire des similarités sémantiques dans cette notion. Elle n'est donc plus à proprement parler la notion de plus grande sous-séquence d'ensembles commune. On pourrait la qualifier de plus grande sous-séquence d'ensembles similaires.

Nous proposons d'utiliser la similarité de Wu and Palmer (1994), définie en considérant la distance entre deux concepts, leur plus proche ancêtre commun, et la profondeur de la nomenclature utilisée :

Définition 18 *Similarité sémantique de Wu et Palmer*
Si nous considérons $C1$ et $C2$ deux concepts appartenant à la classification de concepts suivants :

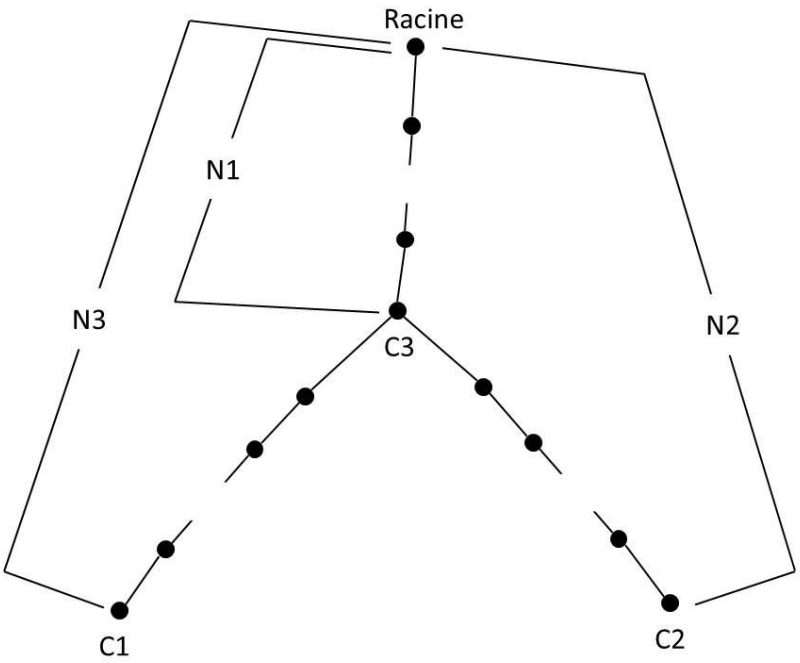

FIGURE 5.3 – Classification de concepts, adapté de Wu and Palmer (1994).

La similarité de Wu et Palmer est définie de la façon suivante :

$$sim_{wp}(C1, C2) = \frac{2.N3}{N1 + N2}$$

Où $C3$ est l'ancêtre commun le plus proche de $C1$ et $C2$, $N1$ est le nombre de nœuds entre $C2$ et la racine de la classification, $N2$ est le nombre de nœuds entre entre $C2$ et la racine, et $N3$ est le nombre de nœuds entre $C3$ et la racine.

Cette similarité est facilement intégrable à la notion de sous-séquence commune lorsque les trajectoires sont des séquences simples. Les comparaisons d'éléments des deux trajectoires, au lieu de valoir 0 lorsqu'ils sont différents ou 1 lorsqu'ils sont identiques, prennent une valeur entre 0 et 1 reflétant leur similarité. Lorsque les séquences sont des séquences d'ensembles, il est nécessaire de calculer des similarités sémantiques entre ensembles de concepts (i.e. codes de médicaments, de diagnostics et d'actes médicaux). Nous proposons d'utiliser la mesure suivante :

Définition 19 *Similarité sémantique entre deux ensembles de concepts*
Étant donné deux ensembles de concepts $X = (x_1, x_2, ..., x_m)$ et $Y = (y_1, y_2, ..., y_n)$ avec $n \leq m$, nous notons C l'ensemble des appariements possibles entre les éléments des deux concepts. La mesure de similarité entre X et Y est définie comme suit :

$$sim(X, Y) = \max_{c=((k_1,l_1),...,(k_n,l_n))\in C} \sum_{i=1}^{n} sim_{wp}(x_{k_i}, y_{l_i})$$

Si on reprend l'exemple en sous-section 5.1.4, nous pouvons calculer une similarité entre deux trajectoires de soins qui prenne en compte la similarité entre leurs éléments. Il est d'abord nécessaire de mesurer la similarité de Wu and Palmer (1994) entre chaque combinaison de concepts, c'est à dire des codes de diagnostics, médicaments et actes médicaux présents dans S_1 et S_2. La matrice de ces similarités est présentée en table 5.1.

	$C10AA01$	$I87.1$	$YYYY034$	$EZAF001$	$B01AC04$
$C10AA04$	5/6	0	0	0	1/6
$I87.1$	0	1	0	0	0
$EZAF001$	0	0	1/6	1	0
$YYYY200$	0	0	5/6	1/6	0
$B01AC04$	1/6	0	0	0	1

TABLE 5.1 – Matrice des similarités de Wu and Palmer (1994) entre les concepts de S_1 et S_2.

On peut ensuite calculer les similarités entre ensembles de concepts, présentés dans la table 5.2.

Le calcul d'une telle similarité entre ensembles de concepts peut lui aussi être réalisé grâce à une approche de programmation dynamique. Il est cependant bien plus complexe que le calcul de similarités sémantiques entre éléments atomiques dans le cas de séquences simples.

	$\{C10AA01\}$	$\{I87.1, YYYY034, EZAF001\}$	$B01AC04$
$\{C10AA04\}$	5/6	0	1/6
$\{I87.1, EZAF001, YYYY200\}$	0	17/6	0
$\{B01AC04\}$	1/6	0	1

TABLE 5.2 – Matrice des similarités entre ensembles de concepts issus de S_1 et S_2.

5.1.6 Applications et résultats

5.1.6.1 Données et méthodes

Afin d'évaluer notre mesure de similarité entre trajectoires de soins, nous avons réalisé une analyse rétrospective en utilisant des données de l'EGB. Des données d'hospitalisation ainsi que les prescriptions de médicaments ont été extraites de l'EGB pour les patients adultes ayant été opérés en ambulatoire d'un geste de la liste MSAP (section 4.1.2), durant l'année 2012. Cet échantillon de l'EGB représente un ensemble de 14 441 patients. Ce sont les codes d'actes médicaux, de médicaments prescrits, de diagnostics principaux, reliés et associés qui ont permis de construire autant de trajectoires de soins.

Tout d'abord, nous nous sommes intéressés à comparer une approche considérant les trajectoires comme des séquences simples à notre approche considérant les trajectoires comme des séquences d'ensembles. Nous dressons une comparaison de l'expressivité des deux formalismes pour une application aux trajectoires de soins, ainsi que de la complexité algorithmique et de temps d'exécution des deux approches dans la section suivante.

Cette première comparaison n'intègre pas encore les similarités sémantiques entres éléments des trajectoires. Dans le but d'évaluer la pertinence de cet enrichissement de notre approche, nous avons également comparé les similarités obtenues avec et sans tenir compte des similarités sémantiques entre éléments des trajectoires. Les similarités sémantiques ont été calculées en utilisant les ontologies de l'ATC pour les médicaments, de la CIM-10 pour les diagnostics et de la CCAM pour les actes médicaux. Trois groupes d'actes de chirurgie bien distincts ont été choisis dans la liste MSAP afin de comparer les similarités obtenues entre trajectoires dans les deux approches. Un échantillons de 287 patients opérés pour une angioplastie, une chirurgie du cristallin ou une chirurgie du sein a ainsi été constitué à partir de l'échantillon précédent. Le choix de trois groupes d'actes ne partageant *a priori* peu de comorbidités permet d'avoir une appartenance d'un patient à un groupe qui soit relativement sûre. Dans la comparaison de ces similarités, nous nous sommes particulièrement intéressés aux similarités intra-groupe et inter-groupe, pour rechercher si l'introduction de similarités sémantiques dans l'approche rapprochait plus les patients d'un même groupe entre eux qu'avec les patients des autres groupes.

Également, nous nous sommes assurés qu'une méthode de classification ascendante hiérarchique appliquée sur cet échantillon de 14 441 patients, sur la bases des similarités entre trajectoires de soins en séquences d'ensembles avec similarités sémantiques, menait bien à la constitution de trois groupes bien distincts. Cette

classification a été réalisée avec un lien de Ward, avec le logiciel R. Trois classes ont été définies en se basant sur le premier grand saut de l'inertie entre classes. La représentation des groupes (ou clusters) a également été réalisée avec R, grâce au package *qgraph* (Epskamp et al., 2012).

5.1.6.2 Plus grande sous-séquence commune sur séquences simples et sur séquences d'ensembles : comparaison

Expressivité du formalisme Le formalisme des séquences d'ensemble permet notamment de gérer la simultanéité de certains éléments des trajectoires de soins. Il est par exemple fréquent que des prescriptions englobent plusieurs médicaments. Également, dans les données médico-administratives, certains éléments ne peuvent être ordonnés dans le temps. Par exemple, les diagnostics associés à un séjour hospitalier peuvent être des diagnostics réalisés lors d'un séjour hospitalier antérieur. Les diagnostics associés représentent d'ailleurs plus un état de santé qu'un diagnostic réalisé lors de ce séjour hospitalier. Lorsque les trajectoires de soins sont représentées comme des séquences simples d'éléments atomiques, ces incertitudes ou simultanéités doivent cependant être ordonnées. Dans ces cas, on pourrait imaginer un ordonnancement aléatoire ou bien un ordonnancement alphanumérique. Intuitivement, l'ordonnancement alphanumérique semble plus pertinent que l'ordonnancement aléatoire, car imposerait les mêmes règles d'ordre pour toutes les trajectoires. L'ordonnancement aléatoire mène inexorablement à sous-estimer la similarité entre trajectoires. Prenons par exemple une trajectoire qu'il conviendrait de représenter en une séquence d'ensembles $T = (A, \{B, C\})$. Si toutefois on veut la représenter en une séquence simple, on pourrait la représenter comme $T_1 = (A, B, C)$ ou bien comme $T_2 = (A, C, B)$. Une similarité entre ces deux trajectoires, entre ces deux façons de représenter T, basée sur la taille de leur plus grande sous-séquence commune mènerait à une mesure inférieur à 1. Or, elles représentent toutes les deux la même trajectoire, initialement T. Si l'ordonnancement alphanumérique n'engendre pas ce problème, il peut néanmoins mener à une sur-estimation de la similarité entre deux trajectoires. Considérons un autre exemple, celui de deux trajectoires de soins $T_3 = (A, B, \{C, D\})$ et $T_4 = (A, \{B, C, D\})$. Si nous transformons ces trajectoires en séquences simples en utilisant un ordre alphanumériques sur les ensembles, on a alors les trajectoires $T'_3 = (A, B, C, D)$ et $T'_4 = (A, B, C, D)$, qui sont alors les mêmes. Leur similarité vaut de fait 1, alors que T_3 et T_4 étaient initialement deux trajectoires différentes, avec par conséquent une similarité inférieure à 1. Dans le cas où la simultanéité des éléments des trajectoire est importante, ces considérations démontrent l'intérêt de représenter les trajectoires en séquences d'ensembles.

Complexité algorithmique et temps d'exécution Pour deux séquences simples de taille n et m, le calcul de la taille de la plus grande sous-séquence commune aux deux séquences peut être effectué avec une complexité $\mathcal{O}(m \times n)$. En effet, l'approche par programmation dynamique calcule la taille de la plus grande sous-séquence commune de façon récursive, dans une matrice de taille $(n+1) \times (m+1)$.

Pour arriver à la dernière case de la matrice correspondant à la taille de la plus grande sous-séquence commune, il est nécessaire de réaliser autant d'opérations que la matrice a de cases. Reprenons l'exemple précédent :

$$S_1 = (\{C10AA01\}, \{EZAF001, I87.1, YYYY034\}, \{B01AC04\})$$

$$S_2 = (\{C10AA04\}, \{I87.1, EZAF001, YYYY200\}, \{B01AC04\})$$

Et écrivons les en séquences simples grâce à un ordonnancement alphanumérique des éléments des ensembles :

$$S'_1 = (C10AA01, EZAF001, I87.1, YYYY034, B01AC04)$$

$$S'_2 = (C10AA04, EZAF001, I87.1, YYYY200, B01AC04)$$

Le calcul par approche dynamique de la taille de la plus grande sous-séquence commune entre S'_1 et S'_2 est présenté par la matrice 5.4.

		$C10AA01$	$I87.1$	$EZAF001$	$YYYY034$	$B01AC04$
	0	0	0	0	0	0
$C10AA04$	0	0	0	0	0	0
$I87.1$	0	0	1	1	1	1
$EZAF001$	0	0	1	2	2	2
$YYYY200$	0	0	1	2	2	2
$B01AC04$	0	0	1	2	2	3

FIGURE 5.4 – Calcul de la taille de la plus grande sous-séquence commune entre S'_1 et S'_2 par une approche de programmation dynamique de complexité $\mathcal{O}(m \times n)$. La taille de la plus grande sous-séquence commune se trouve dans la dernière case en bas à droite, et a nécessité $m \times n = 5 \times 5 = 25$ opérations de comparaison deux-à-deux entre éléments des trajectoires.

Dans le cas où ces deux séquences sont représentées en tant que séquences d'ensembles, la matrice utilisée a pour taille $(n' + 1) \times (m' + 1)$ où n' et m' sont les nombres d'ensembles présents dans les deux trajectoires (matrice 5.5). 'A chaque case de la matrice correspond une opération d'intersection entre deux ensembles. Or cette opération peut se décomposer elle même en un nombre d'opérations inférieur ou égal au produit des tailles des deux ensembles. En particulier, dans le cas où cette intersection est non vide, le nombre d'opérations pour le calcul d'intersection est strictement inférieur à ce produit.

$$
\begin{array}{c}
 \\
\{C10AA04\} \\
\{I87.1,EZAF001,YYYY200\} \\
\{B01AC04\}
\end{array}
\begin{array}{c}
\begin{array}{cccc}
\{C10AA01\} & \{I87.1,EZAF001,YYYY034\} & \{B01AC04\}
\end{array} \\
\left(\begin{array}{cccc}
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 2 & 2 \\
0 & 0 & 2 & 3
\end{array}\right)
\end{array}
$$

FIGURE 5.5 – Calcul de la taille de la plus grande sous-séquence commune entre S_1 et S_2 par une approche de programmation dynamique. La taille de la plus grande sous-séquence commune se trouve dans la dernière case en bas à droite. L'approche nécessite $3 \times 3 = 9$ opérations d'intersection entre ensembles, qui se décomposent finalement en 22 opérations de comparaison deux-à-deux entre éléments atomiques.

Il en résulte que le calcul de la taille de la plus grande sous-séquence commune pour des séquences d'ensembles à n et m éléments demande possiblement moins de $m \times n$ opérations. Autrement dit, le formalisme de séquences d'ensemble demande moins d'opérations pour le calcul de la plus grande sous-séquence commune.

Nous avons mesuré les temps d'exécution du calcul de la plus grande sous-séquence commune, en considérant des trajectoires de soins comme des séquences d'ensembles et comme des séquences simples avec un ordonnancement alphanumérique pour traduire les ensembles en séquences simples (identiquement à l'exemple précédent). La table 5.3 présente cette comparaison pour l'ensemble des 14 441 patients retenus dans l'étude, et également pour le cas particulier des patients opérés pour une angioplastie.

	Angioplastie(n=88)	Ambulatoire(n=14441)
Séquences simples	120 sec	6132 sec
Séquences d'ensembles	60 sec	3362 sec

TABLE 5.3 – Temps d'exécution pour le calcul des similarités sémantiques entre patients deux à deux, pour deux échantillons. Les temps de calcul deux fois plus importants lorsque les séquences sont représentées en séquences simples reflète les différences de complexité algorithmique évoquées précédemment.

5.1.6.3 Classification de trajectoires de soins

Avant d'évaluer la pertinence de l'utilisation de similarités sémantiques dans la comparaison de trajectoires de soins, nous nous sommes assurés qu'une méthode de classification basée sur notre approche mènerait bien à la création de trois groupes, associés aux trois sous-groupes initiaux de patients. Seulement trois patients sont mal classés. Des analyses plus détaillées ont révélé que ces patients partagent des comorbidités fréquentes dans les autres groupes. Une visualisation de cette classification est représentée par la figure 5.6.

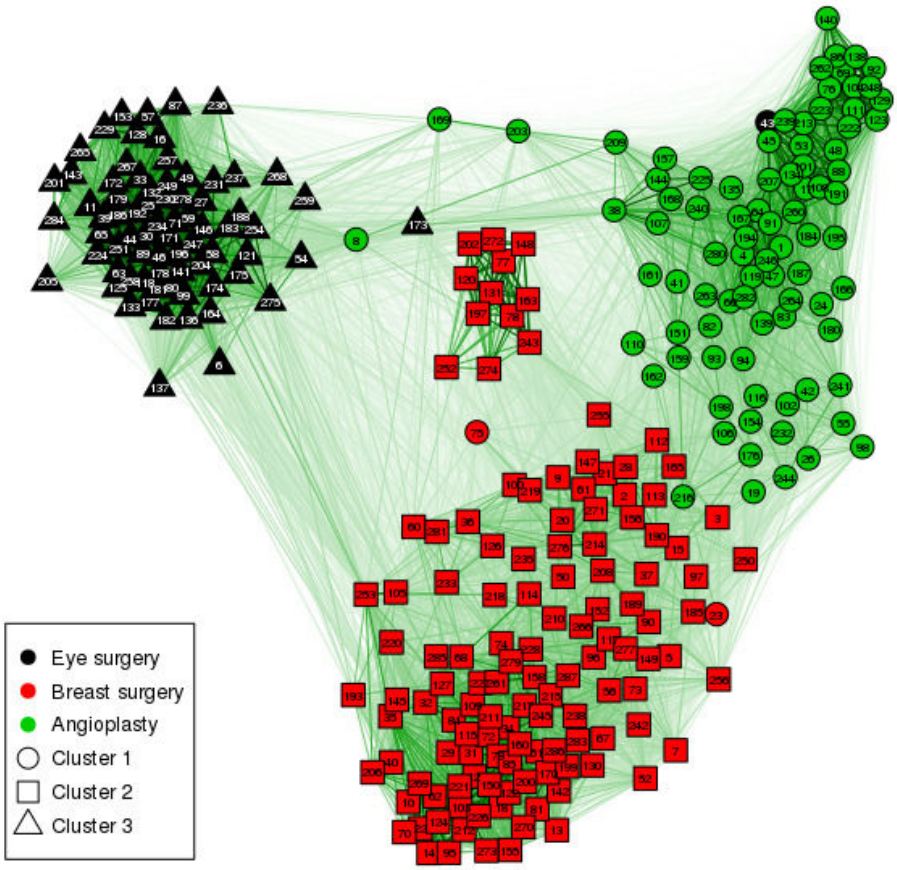

FIGURE 5.6 – Visualisation des clusters de trajectoires de soins. Les arcs du graphe donnent une idée de la similarité entre deux trajectoires de soins : la longueur des arcs est inversement proportionnelle aux similarités entre trajectoires, quand l'épaisseur est elle proportionnelle à ces similarités. Un arc petit et épais reflète une similarité proche de 1.

5.1.6.4 Amélioration des similarités entre trajectoires grâce aux similarités sémantiques

L'introduction de similarités sémantiques dans la mesure de similarité entre trajectoires de soins ne pouvant qu'augmenter ces dernières, nous nous sommes concentrés sur l'étude des similarités intra-groupe et des similarités inter-groupe. Les similarités intra-groupe sont les similarités entre une trajectoire et celle de son groupe, tandis que les similarités inter-groupes sont les similarités d'une trajectoire avec celles des autres groupes. Ces différentes similarités ont été comparées entre l'approche avec l'utilisation de similarités sémantiques et l'approche sans similarités sémantiques, dans le but d'identifier si les similarités intra-groupe augmentent plus que les similarités inter-groupe avec l'introduction de similarités sémantiques. Autrement dit, nous voulions déterminer si l'introduction des similarités sémantiques

conduisait à des groupes plus homogènes du point de vue de leurs similarités entre trajectoires de soins. Sur l'échantillon de 287 patients, donc de 287 trajectoires de soins, les similarités deux à deux entre trajectoires ont été calculées grâce aux deux approches, c'est à dire sans et avec l'intégration de similarités sémantiques. Dans les deux cas, pour chaque patient, nous avons calculé le ratio de la somme de ses similarités intra-groupes rapportée à la somme de ses similarités inter-groupes :

$$r(T_i^k) = \frac{\sum_{j \neq i} sim(T_i^k, T_j^k)}{\sum_{L \neq k} sim(T_i^k, T_j^l)}$$

Où T_i^k est la trajectoire i du cluster k.

À chaque trajectoire sont donc associés deux ratios, celui obtenu avec similarités sémantique, et celui obtenu sans. Pour chaque approche, on a donc un vecteur de ratios, de la taille du nombre de patients. Étant appariés, nous avons comparé ces deux vecteurs par un test des rangs signés de Wilcoxon (1945). Cette comparaison statistique a pu montrer qu'après l'introduction des similarités sémantiques, les ratios étaient significativement plus élevés (p=0,002). Autrement dit, après introduction des similarités sémantiques dans l'approche, les similarités d'un patient avec sa classe ont plus augmenté que ces similarités avec les autres classes.

5.1.7 Discussion et conclusion

Nous avons proposé un formalisme de représentation des trajectoires de soins en séquences d'ensembles, qui grâce à son expressivité est adapté au contexte des données médico-administratives. Suite à cette proposition, nous avons adapté la notion de la plus grande sous-séquence commune à ce formalisme, dans le but de mesurer la longueur de la part commune entre deux trajectoires de soins. Sur la base de cette mesure, nous avons proposé une première mesure de similarité entre trajectoires de soins. Ce formalisme de séquences d'ensemble, en plus du gain d'expressivité, a permis de réduire la complexité algorithmique de la méthode, et ainsi les temps d'exécution, de moitié dans notre cas d'étude.

Afin de ne pas sous-estimer la similarité entre trajectoires de soins, nous avons introduit dans l'approche des similarités sémantiques entre éléments des trajectoires à la place de comparaisons strictes. Cet enrichissement assouplit la méthode par l'introduction d'une logique floue. Les différences minimes entre certains éléments, dues aux grandes profondeurs des nomenclatures utilisées pour les coder, sont ainsi mieux prises en comptes, c'est-à-dire plus uniquement comme des éléments strictement différents. Bien que ne pouvant mener qu'à l'augmentation des similarités entres trajectoires de soins, cet enrichissement a montré qu'il rassemblait plus les patients avec leur groupe assigné *a priori* plutôt qu'avec les autres groupes. On suppose donc qu'il permettrait d'obtenir des groupes plus homogènes dans une méthode de classification de patients reposant sur des similarités entre trajectoires de soins.

Cependant, calculer des similarités sémantiques entre ensembles est bien plus complexe qu'entre éléments deux à deux. De fait, bien que le formalisme de séquences simples soit moins expressif que celui des séquences d'ensembles, il est le

formalisme pour lequel il est plus facile d'intégrer des similarités sémantiques. Dans la pratique, on peut alors être forcé à choisir entre (i) adopter le formalisme plus expressif des séquences d'ensembles (ii) et l'introduction de similarités sémantiques. Dans le cas où l'introduction de similarités sémantique est une priorité, le formalisme des séquences simples, avec un ordonnancement alphanumérique des codes de médicaments, actes et diagnostics des ensembles, semble le meilleur compromis. En parallèle aux outils popularisant l'analyse de séquences, des similarités sémantiques mesurées sur les hiérarchies des nomenclatures utilisées dans les bases de données médico-administratives pourrait permettre de rendre ces méthodes et outils plus adaptés à l'analyse des trajectoires de soins issues de ces bases de données. Dans la pratique, les coûts de transaction dans les méthodes d'édition de séquences pourrait par exemple se baser sur des similarités sémantiques mesurées sur l'ATC, la CIM-10 ou encore la CCAM. Dans cette perspectives, de nouvelles fonctions de *queryMed* (Rivault et al., 2018c) couplées à l'utilisation du package *TraMineR*[1] (Studer and Ritschard, 2016) pourraient permettre de faciliter la dissémination de notre approche. L'analyse de séquence *multichannel* (Pollock, 2007), implémentée dans *TraMineR*, peut être une piste pour considérer l'aspect multi-dimensionnel des trajectoires de soins issues des bases de données médico-administratives (Roux et al., 2018a). Des expérimentations avec une intégration de similarités sémantiques sont à réaliser dans ce sens.

Quand la comparaison de trajectoires permet de mesurer la part commune à deux trajectoires de soins, l'extraction de motifs permet elle d'identifier des « sous-trajectoires de soins » intéressantes, notamment car partagées par plusieurs patients.

1. Le site de *TraMineR*, une « boîte à outil pour l'analyse et l'exploration de données séquentielles » : `traminer.unige.ch`

5.2 Extraction de règles d'association à partir de trajectoires de soins : introduction de la hiérarchie des nomenclatures médicales

5.2.1 Introduction

Les règles d'association, faisant partie des méthodes d'extraction de motifs, peuvent permettre d'identifier des associations entre événements de santé dans des trajectoires de soins issues des bases de données médico-administratives. De telles associations, par exemples entre diagnostics et médicaments, sont décrites par différentes mesures statistiques (Geng and Hamilton, 2007). Néanmoins, des limites propres à la méthode ou aux données font que l'extraction de règles à partir de trajectoires de soins est loin d'être triviale.

Dans l'extraction de règles d'associations, des règles fréquentes et prédictives sont souvent qualifiées de fortes, ou d'intéressantes. En effet, le caractère prédictif d'une règle, souvent illustré par sa confiance, nous donne une idée quant à la valeur prédictive de l'antécédent sur le conséquent. Par exemple, la confiance d'une règle constituée d'un médicament à l'antécédent et d'un diagnostic au conséquent, mesure une estimation de la probabilité qu'un patient qui a pris ce médicament, ait également eu ce diagnostic. Une des limites de l'extraction de règles d'associations, est la capacité des méthodes et algorithmes dédiés à extraire seulement peu de règles fortes, c'est-à-dire à la fois fréquentes et prédictives. Cette limite s'explique tout d'abord par le fait que les algorithmes d'extraction de règles d'associations sont très souvent fondés sur un seuil minimal de fréquence observée de la règles dans un jeu de données (Agrawal and Srikant, 1994; Srikant and Agrawal, 1996; Zaki, 2001). Un seuil trop élevé écarte les règles peu fréquentes, quand bien même elles pourraient avoir une confiance élevée. Un seuil faible, choisi justement dans le but d'écarter le moins de règles possible, conduit à un volume de règles beaucoup plus important qu'il devient alors difficile d'analyser les résultats.

Enfin, la profondeur des nomenclatures médicales utilisées pour codifier les événements médicaux des trajectoires de soins induit une grande variabilité des données. Cette variabilité conduit à l'extraction de très nombreuses règles, très spécifiques, et souvent peu fréquentes. Or, des règles plus générales, notamment du point de vue de la hiérarchie des nomenclatures médicales, pourraient résumer l'information contenue dans plusieurs règles spécifiques, et ainsi réduire le nombre de règles à analyser. Une approche envisageable pourrait être de choisir un niveau de granularité plus élevé dans les hiérarchies médicales pour représenter les données avant l'extraction des règles d'association. Les trajectoires de soins sur lesquelles on extrait des règles d'association seraient ainsi constituées de familles de médicaments, d'actes médicaux, et de chapitres de maladies. Une telle approche écarte les règles spécifiques qui pourraient être très prédictives bien que peu fréquentes. Comme le propose Manda et al. (2012), l'extraction de règles d'association multi-niveaux, c'est à dire à différents degrés d'abstraction –ou de granularité–, permet d'extraire de nouvelles règles fortes qui sont des agrégation de règles spécifiques, sans pour autant écarter

les règles spécifiques intéressantes. L'extraction de règles multi-niveaux mène à une extraction de règles encore plus abondante et aggrave ainsi une des premières limites de la méthode. Également, si l'extraction de règles à un seul degré de granularité induit déjà le principe de règles redondantes (Bayardo et al., 2000), qui sont des cas particuliers d'autres règles plus générales, l'extraction de règles multi-niveaux complexifie la gestion de la redondance avec l'extension de cette notion aux hiérarchies des nomenclatures et ainsi la génération de bien plus de règles redondantes (Shaw et al., 2009; Chandanan and Shukla, 2015).

Dans le cadre de ma mobilité pour une collaboration avec Nicolas Jay et Aurélie Bannay du Laboratoire Lorrain de Recherche en Informatique et ses Applications (Loria) et du Département d'Information Médicale (DIM) de l'hôpital central de Nancy, nous nous sommes intéressés à une application de l'extraction de règles d'association pour l'aide au codage dans un établissement hospitalier. La vérification du codage des diagnostics réalisés à l'hôpital, lors d'un séjour hospitalier, nécessite de reprendre les fichiers et compte-rendus des patients un à un. De nombreux techniciens d'information médicale (TIM) réalisent ainsi cette tâche quotidiennement. Tout particulièrement, ils s'attachent à vérifier si des diagnostics de forte sévérité ne sont pas omis : une omission d'un de ces codes aurait pour conséquence une erreur de financement en défaveur de l'hôpital. Fournir aux TIM des règles d'association fortes –très prédictives– entre médicaments et diagnostics, peut les aider à prioriser la recherche de séjours sous-codés à ne pas omettre s'ils ont connaissance des médicaments pris par les patients. Dans cette application, adapter l'extraction de règles à différents degrés de granularité des nomenclatures médicales peut permettre de découvrir de nouvelles règles intéressantes qu'une extraction simple ne pourrait découvrir.

5.2.2 Objectifs

L'objectif de cette collaboration était d'étudier la faisabilité d'une extraction de règles multi-niveaux pour fournir à des TIMs un ensemble raisonnable de règles d'association les aidant dans leurs vérifications. Pour cela, de telles règles extraites sur des données hospitalières, doivent avoir un antécédent contenant des médicaments, qui soit très prédictif sur un conséquent contenant un diagnostic de forte sévérité. On s'intéresse donc particulièrement aux règles à forte confiance –proche de 1– et pour poursuivre cet objectif, on réalise une méthode d'extraction de règles multi-niveaux. La gestion de la redondance des règles ainsi obtenues doit pouvoir satisfaire l'objectif d'une extraction raisonnablement non-abondante.

5.2.3 Données

Les données utilisées dans le cadre de cette collaboration sont des données hospitalières issues de l'hôpital central de Nancy, dans tous les services de l'hôpital, pour tous les patients de plus de 60 ans, durant un mois de l'année 2018. La base de transactions pour l'extraction des règles se compose alors de trajectoires de soins,

séquences constituées des prises de médicaments et diagnostics réalisés à l'hôpital durant le séjour des patients (table 5.4). Les médicaments sont codés selon l'ATC, et les diagnostics selon la CIM-10.

Id transaction	Séquence
T_1	(atc:A02BC05, atc:N02BE01, icd10:M16.1)
T_2	(atc:N02BE01, atc:A02BC05, icd10:M17.1)
T_3	(atc:B01AC04, atc:A02BC01, icd10:M17.1)
T_4	(atc:B01AB05, atc:N02BE01, icd10:M16.1)
T_5	(atc:B01AC04, atc:A02BC05, icd10:M17.1)
T_6	(atc:A02BC05, icd10:M17.1)

TABLE 5.4 – Une base de transactions fictive constituées de médicaments codés selon l'ATC et de diagnostics codés selon la CIM-10.

5.2.4 Extraction de règles multi-niveaux

L'extraction de règles multi-niveaux s'est faite par la prise en compte de la structure hiérarchique de l'ATC. C'est donc uniquement à l'antécédent que les règles peuvent contenir des éléments à différents degrés de granularité. En effet, les codes à forte sévérité, auxquels on s'intéresse, sont des codes "feuilles" dans l'arbre de la CIM-10. De plus, la généralisation uniquement à l'antécédent permet de mieux gérer la redondance des règles. Cet ajout de la hiérarchie a été introduit avant l'extraction des règles d'association. Cette modification a consisté à rajouter les classes ancêtres des médicaments dans les données servant à l'extraction, c'est-à-dire dans la base de transactions (table 5.5). Pour ne pas obtenir un nombre de règle trop abondant, nous nous sommes restreint à ajouter aux transactions jusqu'à trois générations ancestrales. Suite à ce pré-traitement, l'extraction des règles d'association a été réalisée grâce à l'implémentation en *C* de l'algorithme *a priori* (Agrawal and Srikant, 1994) dans le package R *arules* (Hahsler et al., 2005). Les paramètres choisis imposent aux règles de contenir des médicaments à l'antécédent et un diagnostic à forte sévérité au conséquent, d'avoir un support minimal de 0,5%, une confiance minimale de 70%, une taille maximale de 5 éléments et une taille minimale de 2 éléments.

L'utilisation de cet algorithme d'extraction de règles d'association permet alors grâce au pré-traitement d'extraire des règles à différents degrés de granularité. Par exemple, en plus d'extraire la règle $R_1 : atc{:}A02BC05 \longrightarrow icd10{:}M17.1$, on va aussi extraire sa règles plus générales $R_2 : atc{:}A02BC \longrightarrow icd10{:}M17.1$.

Si grâce à cette généralisation on peut extraire plus de règles intéressantes, des fausses règles apparaissent dans les résultats. En effet, on peut extraire des règles avec à l'antécédent un médicament et une classe de médicaments à laquelle il appartient. Ces règles sont donc à écarter.

La règle $(atc{:}A02BC, atc{:}A02BC05) \longrightarrow icd10{:}M17.1$ est par exemple une "fausse" règle.

Si cette généralisation de l'extraction de règles multi-niveaux par ce pré-traitement

Id transaction	Séquence
T_1	(atc:A02BC, atc:N02BE, atc:A02BC05, atc:N02BE01, icd10:M16.1)
T_2	(atc:N02BE, atc:A02BC, atc:N02BE01, atc:A02BC05, icd10:M17.1)
T_3	(atc:B01AC, atc:A02BC, atc:B01AC04, atc:A02BC01, icd10:M17.1)
T_4	(atc:B01AB, atc:N02BE, atc:B01AB05, atc:N02BE01, icd10:M16.1)
T_5	(atc:B01AC, atc:A02BC, atc:B01AC04, atc:A02BC05, icd10:M17.1)
T_6	(atc:A02BC, atc:A02BC05, icd10:M17.1)

TABLE 5.5 – Base de transactions fictive après rajout d'une génération ancestrale de classes ATC

est relativement simple, elle s'accompagne cependant d'une augmentation considérable du nombre de règles découvertes, pouvant rendre leur analyse compliquée. Également, la notion de redondance entre règles d'association, déjà présente dans l'extraction de règles simples, doit être généralisée à l'extraction de règles multi-niveaux, qui en introduit encore plus. Une élagage des règles redondantes doit pouvoir limiter l'augmentation du nombre de règles ainsi obtenues.

5.2.5 Généralisation de la redondance aux règles multi-niveaux

La notion de la redondance des règles d'association a été formalisée par Bayardo et al. (2000) sous le nom de règles négatives ou de règles « *zero improvement* ». Sous cette appellation on peut alors comprendre que la notion de redondance de règle repose sur une mesure de qualité des règles : une règle « *zero improvement* » est littéralement une règle qui n'apporte pas d'amélioration d'après une mesure de qualité choisie, par rapport à une autre règle. Ces règles qui n'apportent pas d'amélioration peuvent alors être raisonnablement écartées de l'étude. Cette procédure d'élagage contribue à une meilleure clarté des résultats par un nombre de règles plus restreint à étudier.

Avant de s'intéresser à la redondance des règles multi-niveaux, il convient d'étudier la redondance entre règles simples, comme définie par Bayardo et al. (2000). Dans cette définition, les règles redondantes le sont par rapport à leurs règles plus générales. Définissons donc les principes de règles imbriquées.

Définition 20 *Règles imbriquées*
Soient $R : X \longrightarrow Y$ et $R' : X' \longrightarrow Y$ deux règles d'association. R' est plus générale que R si et seulement si $X' \subset X$. On dit alors qu'elles sont imbriquées, que R est une sous-règle de R' ou encore que R' est une sur-règle de R. R' est plus générale que R, et on note également $R' > R$.

Après cette définition, nous pouvons énoncer celle de la redondance d'une règle, généralement fondée sur la confiance :

Définition 21 *Redondance de règles imbriquées*
Une règle est redondante si elle possède une règle plus générale avec une confiance supérieure ou égale.

Plus formellement, une règle $R : X \longrightarrow Y$ *est dite redondante s'il existe une règle* $R' : X' \longrightarrow Y$ *avec* $X' \subset X$*, telle que* $confiance(R') \geq confiance(R)$*. Pour cette raison on peut aussi dire qu'une règle plus spécifique est redondante si elle est seulement autant ou moins prédictive qu'une de ces règles plus générales.*

Exemple 1 *La règle* R_1 : $atc{:}A02BC05 \longrightarrow icd10{:}M17.1$ *est plus générale que* $R_3 : (atc{:}A02BC05, atc{:}N02BE01) \longrightarrow icd10{:}M17.1$*. Pour en revenir au principe des règles « zero improvement », la redondance consiste ici à savoir si le retrait de* $atc{:}N02BE01$ *à l'antécédent de la règle* R_3 *améliore ou non sa confiance.*
En l'occurrence, on peut calculer la confiance de ces règles dans la base de transaction 5.4 :

$$confiance(R_1 : atc{:}A02BC05 \longrightarrow icd10{:}M17.1) = 0,75$$

$$confiance(R_3 : (atc{:}A02BC05, atc{:}N02BE01) \longrightarrow icd10{:}M17.1) = 1/2 = 0,5$$

Comme $R_1 > R_3$ *et* $confiance(R_1) > confiance(R_3)$*, la règle* R_3 *est redondante.*

Avec la généralisation de l'extraction de règles multi-niveaux, on peut introduire une nouvelle façon de voir le principe de sous-règles. En effet, le fait d'ajouter des classes ancêtres de médicaments nous amène à étendre le principe de sous-règles aux relations de subsomption entre classes de médicaments. Pour définir cette généralisation du principe de règles imbriquées, nous proposons d'étendre celui des ensembles imbriqués :

Définition 22 *Ensembles imbriqués*
Soient $X = (x_1, x_2, ..., x_n)$ *et* $X' = (x'_1, x'_2, ..., x'_m)$ *deux ensembles avec* $m \leq n$*.* X' *est un ensemble plus général que* X *si et seulement si* x'_i *est présent dans* X *ou bien subsume un élément de* X*, pour tout* $x'_i \in X'$*. On note* $X' \subset_* X$*.*

Définition 23 *Règles multi-niveaux imbriquées*
Soient $R : X \longrightarrow Y$ *et* $R' : X' \longrightarrow Y$ *deux règles d'association multi-niveaux.* R' *est plus générale que* R *si et seulement si* $X' \subset_* X$*. On dit alors qu'elles sont imbriquées, et que* R *est une sous-règle de* R'*, et que* R' *est une sur-règle de* R*. On note également* $R' >_* R$*.*

Exemple 2 R_1 : $(atc{:}A02BC05) \longrightarrow icd10{:}M17.1)$ *est une sous-règle de* R_2 : $(atc{:}A02BC) \longrightarrow icd10{:}M17.1$.

Il est assez commode avec cette généralisation de la notion de sous-règles d'utiliser la confiance pour comparer les qualités de deux règles imbriquées, et ainsi

de généraliser la définition 21. Cependant, l'opérateur $\geq$ dans la comparaison des confiances des règles, mènerait à toujours préférer les règles plus générales lorsque leur confiance est égale. C'est plutôt souhaité dans le cas des règles simples, à un seul niveau de hiérarchie, car cela réduit la taille des règles retenues (les règles redondantes, plus longues, ayant été écartées). En revanche, dans le cas des règles d'association multi niveaux, cela conduirait à constamment préférer une règle plus générale à sa règle spécifique lorsque leur confiance sont égales. On privilégierait ainsi dans de nombreux cas des règles qui perdraient en information, constituées de familles de médicaments et non plus de médicaments, sans pour autant gagner en prédictivité par cette généralisation. La redondance doit finalement s'adapter au domaine d'application et à l'étude. Nous définissons ainsi pour ce travail la redondance des règles multi-niveaux avec conséquent fixe de la façon suivante :

Définition 24 *Redondance de règles multi-niveaux imbriquées*
Une règle multi-niveaux est redondante si elle possède une règle plus générale avec une confiance strictement supérieure. Plus formellement, une règle multi-niveaux $R : X \longrightarrow Y$ *est dite redondante s'il existe une règle* $R' : X' \longrightarrow Y$ *avec* $X' \subset X$*, telle que* $confiance(R') > confiance(R)$*. Pour cette raison on peut aussi dire qu'une règle plus spécifique est redondante si elle est moins prédictive qu'une de ces règles plus générales.*

Exemple 3 *Si on revient à l'exemple de* R_1 *et* R_2*, on peut calculer leur confiance et support :*

$$support(R_1 : (atc{:}A02BC05) \longrightarrow icd10{:}M17.1) = 0,5$$

$$confiance(R_1 : (atc{:}A02BC05) \longrightarrow icd10{:}M17.1) = 0,75$$

$$support(R_2 : (atc{:}A02BC) \longrightarrow icd10{:}M17.1) = 4/6 = 2/3$$

$$confiance(R_2 : (atc{:}A02BC) \longrightarrow icd10{:}M17.1) = 4/5$$

R_2 *est plus générale que* R_1*, et sa confiance est meilleure.* R *est donc redondante.*

L'élagage des règles ne se fait pour le moment uniquement sur des règles spécifiques par rapport à leurs règles générales. Or, il est possible qu'une règle générale soit moins bonne que toutes ses règles plus spécifiques (du point de vue de leur confiance). Dans cette situation, on peut préférer étudier toutes les règles spécifiques sans cette règle moins bonne, jugée redondante.
La définition d'une règle redondante devient alors :

Définition 25 *Une règle multi-niveaux est redondante si elle possède une règle plus générale avec une confiance strictement supérieure ou si toutes ses sous-règles ont une confiance strictement supérieure à la sienne.*

Plus formellement, une règle multi-niveaux $R : X \longrightarrow Y$ *est dite redondante dans les cas suivants :*

1. *S'il existe une règle* $R' : X' \longrightarrow Y$ *avec* $X' \subset_* X$*, telle que* $confiance(R') > confiance(R)$ *;*
2. *Ou bien si pour toutes les règles* R^* *telles que* $R' >_* R^*$*, on a* $confiance(R^*) > confiance(R)$*.*

5.2.5.1 Elagage de la redondance

Nous avons donc étendu la notion de redondance simple de une notion de redondance hiérarchique. Également, avec cette nouvelle définition 25, des règles peuvent être redondantes par comparaison à leurs sur-règles comme à leurs sous-règles. Cette distinction a pour conséquence de complexifier l'étape d'élagage des règles redondantes. En effet, il peut exister certaines dépendances entre les différentes sortes de redondances. Retirer une règle peut par exemple rendre une de ses sur-règles redondantes par rapport à toutes ses sous-règles restantes. Nous avons choisi de gérer la redondance des règles grâce aux algorithmes 2, 3, et 4, dans l'ordre suivant de présentation.

Dans un premier temps, les règles redondantes simples, c'est-à-dire sans considérer que des règles peuvent être imbriquées du fait des hiérarchies, sont élaguées.

Algorithme 2 : Élagage des règles spécifiques redondantes simples.

Données : R un ensemble de règles d'associations.
Résultat : R duquel sont retirées les règles spécifiques redondantes simples.

```
pour r dans R faire
    pour r' dans R faire
        si r ≠ r' et conf(r) ≥ conf(r') et r > r' alors
            suppression de r'
```

Un algorithme très proche permet ensuite d'élaguer les règles redondantes hiérarchiques. L'opérateur $>$ est ici préféré pour la comparaison de la confiance des règles.

Algorithme 3 : Élagage des règles spécifiques redondantes hiérarchiques

Données : R un ensemble de règles d'associations.
Résultat : R duquel sont retirées les règles spécifiques redondantes hiérarchiques.

```
pour r dans R faire
    pour r' dans R faire
        si r ≠ r' et conf(r) > conf(r') et r >* r' alors
            suppression de r'
```

Enfin, un dernier algorithme permet de retirer les règles redondantes générales par rapport à leurs sous-règles.

Algorithme 4 : Élagage des règles générales redondantes hiérarchiques

Données : R un ensemble de règles d'associations.
Résultat : R duquel sont retirées les règles générales redondantes hiérarchiques.

```
pour r dans R faire
    a_supprimer = Vrai
    pour r' dans R faire
        si r ≠ r' alors
            si conf(r) > conf(r') alors
                a_supprimer = Faux
                Stop
            si r >* r' alors
                a_sous_regle = Vrai
    si a_sous_regle = Vrai et a_supprimer = Vrai alors
        suppression de r
```

5.2.6 Résultats

Par principe, la généralisation de l'extraction de règles d'associations aux règles multi-niveaux ne peut qu'augmenter le nombre de règles obtenues. Notamment, les règles obtenues par toute autre extraction simple à un seul degré de granularité, seront présentes dans les règles multi-niveaux. De plus, cette généralisation pourrait amener à découvrir des règles qui ne seraient pas plus intéressantes (du point de vue de la confiance) que les règles obtenues avec une extraction simple. En effet, suivant l'étude, son domaine d'application et les hiérarchies utilisées, il se pourrait que les règles les plus spécifiques, au degré de granularité le plus fin, soient les règles les plus prédictives.

La table 5.2.6 présente le nombre de règles obtenues dans le cas d'une extraction simple au degré de granularité le plus fin, celui des règles obtenues avec une extraction multi-niveaux (avec trois degrés de granularité), et celui de ces règles après l'élagage des règles redondantes. L'extraction de règles multi-niveaux permet dans notre cas de découvrir des règles sur un plus grand nombre de diagnostics à forte sévérité. Cette extraction mène aussi à un nombre de règles trop important pour qu'elles soient toutes lues par un expert. L'élagage des règles redondantes ramène le nombre de règles à étudier à un nombre bien plus raisonnable. Cette réduction du nombre de règles à analyser a ainsi permis à Nicolas Jay et Aurelie Bannay, médecins de santé publiques au DIM de Nancy et habitués aux problématiques de codage des actes et diagnostics en milieu hospitalier, d'être en capacité d'analyser les règles obtenues pour la majorité des diagnostics étudiés. Il reste cependant que certains diagnostics sont encore associés à plusieurs centaines de règles, même après l'élagage de la redondance.

Codes	Nombre de règles simples	Nombre de règles simples après élagage de la redondance selon Bayardo et al. (2000)	Nombre de règles multi-niveaux	Nombre de règles multi-niveaux après élagage de la redondance
A46 Érysipèle	0	0	81	1
B95.6 Staphylococcus aureus	2	1	482	12
D61.1 Aplasie médullaire médicamenteuse	4	2	504	17
E43 Malnutrition protéino-énergétique grave	70	33	12902	532
J69.0 Pneumopathie due à des aliments et des vomissements	0	0	13	4
N17.8 Autres insuffisances rénales aiguës	4	2	2398	80
R02 Gangrène	0	0	39	3
R57.2 Choc septique	0	0	494	10
R65.1 Syndrome de réponse inflammatoire systémique d'origine infectieuse avec défaillance d'organe	0	0	57	11
Z51.1 Séance de chimiothérapie pour tumeur	106	21	94219	439
Z51.5 Soins palliatifs	46	10	13073	102
Z74.2 Besoin d'assistance à domicile	1	1	39	6

TABLE 5.6 – Dénombrements des règles extraites grâce à une extraction simple au degré de granularité le plus fin, à une extraction multi-niveaux à trois degrés de granularité, et après élagage des règles multi-niveaux redondantes.

Comparer la confiance moyenne (table 5.7) avant et après l'élagage des règles redondantes n'est pas aussi pertinent. En effet, l'élagage des règles redondantes amène parfois à retirer des règles dont la confiance est supérieure à la confiance moyenne. Dans ce cas, la confiance moyenne des règles diminue. Pour autant, la

règle a été écartée car redondante : une ou plusieurs autres règles capturent donc l'information que cette règles apportait, avec une meilleur confiance. Par principe, la gestion de la redondance ne mène donc pas forcément à une confiance moyenne plus élevée, mais de façon sûre à une diminution du nombre de règles d'association à analyser. Également, la confiance maximale des règles ne peut qu'augmenter (ou rester la même) lorsque l'on extrait des règles multi-niveaux, par rapport à des règles simples.

Codes	Confiances moyennes des règles simples	Confiance maximale des règles simples	Confiance moyenne des règles multi-niveaux	Confiance moyenne des règles multi-niveaux après gestion de la redondance	Confiance maximale des règles multi-niveaux après gestion de la redondance
A46	\	\	0,717	0,778	0,778
B95.6	0,7	0,7	0,715	0,721	0,875
D61.1	0,777	0,875	0,735	0,736	0,875
E43	0,78	1	0,76	0,765	1
J69.0	\	\	0,7	0,7	0,7
N17.8	0,7	0,727	0,754	0,757	1
R02	\	\	0,731	0,758	0,875
R57.2	\	\	0,802	0,803	1
R65.1	\	\	0,744	0,744	0,875
Z51.1	0,8	1	0,816	0,808	1
Z51.5	0,832	1	0,82	0,802	1
Z74.2	0,7	0,7	0,747	0,756	0,8

TABLE 5.7 – Confiances moyennes des règles extraites grâce à une extraction simple au degré de granularité le plus fin, à une extraction multi-niveaux à trois degrés de granularité, et après élagage des règles multi-niveaux redondantes.

5.2.7 Perspectives

L'apport de l'extraction de règles multi-niveaux est d'augmenter le nombre de règles intéressantes (notamment fréquentes et prédictives) découvertes. Dans le cadre de cette application à l'aide au codage, cette approche permet de découvrir plus de règles très prédictives. L'apport de la gestion de la redondance adaptée à cette généralisation de l'extraction est elle de rendre les résultats plus lisibles par des spécialistes, en réduisant le nombre de règles à analyser. Tout d'abord, la redondance se base uniquement sur la confiance. On pourrait vouloir intégrer dans la définition

de la redondance d'autres mesures de qualité des règles, par exemple le support. Ensuite, dans le cadre d'une application pour l'aide au codage, on peut se demander si la gestion de la redondance est essentielle. En effet, si l'application est d'aider au codage en proposant des séjours hospitaliers à vérifier en priorité, fournir aux TIMs un ordre de séjours à vérifier plutôt qu'un ensemble de règles est plus avisé. Enfin, un ordre de séjour pourrait se baser sur un ensemble de règles, sans que les TIMs n'aient à analyser ces associations. Dans cette situation, découvrir peu ou beaucoup de règles n'importerait alors peu, tant que ces règles mèneraient à une bonne prédiction des séjours sous-codés. Néanmoins, la gestion de la redondance que nous proposons permet d'écarter certaines règles trop générales en préférant leurs sous-règles plus spécifiques. On peut faire l'hypothèse que la suppression de ces règles peut améliorer la qualité de prédiction de l'approche, notamment en réduisant le cas de faux négatifs, c'est à dire les cas où l'approche soupçonnerait un sous-codage à tort. Une mise en application sur des données hospitalières, en collaboration avec des TIMs, permettraient d'infirmer ou de confirmer ces hypothèses.

Si l'extraction de règles multi-niveaux et la gestion des règles redondantes permettent d'obtenir pour cette application des règles intéressantes, qui ne sont pas noyées parmi un déluge de règles inintéressantes, il n'en ai pas forcément de même lorsque l'on considère des paramètres différents. Notamment, avec une confiance minimale plus faible, la gestion de la redondance devient longue et difficile du fait du nombre important de règles extraites. Dans une extraction plus large de règles d'association, des outils d'exploration des règles pourrait permettre de filtrer des règles inintéressantes, ou bien de sélectionner des sous-ensembles de règles d'intérêt. La section 5.3 propose une telle approche reposant sur l'utilisation des technologies du Web Sémantique et des ontologies du *Linked Data*.

5.3 Exploration de règles d'associations : utilisation des technologies du Web Sémantique et des ontologies du *Linked Data*

5.3.1 Introduction

Toujours dans le cadre de la collaboration avec Nicolas Jay et Aurelie Bannay, nous nous sommes intéressés à l'utilisation des technologies du Web Sémantique et d'ontologies du *Linked Data* pour l'exploration de règles d'associations. Nous l'avons déjà évoqué en section 5.2, une des principales limites aux méthodes d'extraction de règles d'association est leur capacité à extraire un nombre de règles si important qu'il devient difficile d'analyser les résultats. De plus, comme l'a souligné Chen et al. (2008), toutes les règles extraites ne sont pas forcément pertinentes. Certaines peuvent être triviales car déjà connues. Par exemple dans une application en pharmaco-épidémiologie, découvrir des règles décrivant une association entre un médicament et un état de santé déjà connue, peut alors sembler peu utile. Dans ce cas, ces règles sont jugées peu intéressantes du fait de l'apport d'une connaissances externe, et non du fait d'une mesure de qualité statistique associée à la règle. De la même façon, des connaissances sur les relations entre éléments d'une règle pourrait la rendre intéressante. Pour une application en pharmaco-vigilance, il peut ainsi être pertinent d'extraire des associations entre médicaments et événements indésirables. Les règles ainsi ciblées permettraient alors de caractériser ces associations particulières –signaux de pharmacologie– par des mesures statistiques propres au domaine des règles d'association. Dans la sécurité des soins, les règles associées à des complications ou à des consommations de soins innatendues, comme une interaction médicamenteuse, une contre-indication entre médicaments, ou entre médicament et maladie, sont des règles pertinentes à étudier. Nous proposons dans ce travail d'utiliser des ontologies médicales et pharmacologiques du *Linked Data* et les technologies du Web Sémantique pour filtrer et cibler des ensembles de règles d'association médicales.

5.3.2 Objectifs

L'approche doit pouvoir permettre de se focaliser sur des groupes de règles, comme d'en écarter certaines. Si un des objectif est la réduction du nombre de règles à analyser, l'objectif principal est de pouvoir proposer une méthode d'exploration de règles reposant sur l'apport de connaissances externes. Ce travail se concentre essentiellement sur les indications des médicaments, les contre-indications et les interactions. La méthode doit cependant pouvoir être généralisable à toute autre relation entre les éléments des règles.

5.3.3 Données et extraction de règles

Cette étude a été réalisée sur la base de données de patient opérés pour une arthroplastie (paragraphe 4.1.2). Les médicaments et diagnostics, sous la forme de

codes ATC et respectivement de codes CIM-10, ont été extraits dans un intervalle de six mois centré sur la première hospitalisation. Les hospitalisations retenues sont donc les premiers séjours hospitaliers pour une arthroplastie, d'avril à décembre durant l'année 2012, afin d'avoir des historiques de soins de taille similaire. Cet échantillon représente 1293 patients.

L'extraction des règles d'association a été réalisée grâce à l'algorithme *a priori* (Agrawal and Srikant, 1994) disponible dans le package R *arules* (Hahsler et al., 2005). Les paramètres choisis imposent aux règles de contenir des médicaments à l'antécédent et un diagnostic au conséquent, d'avoir un support minimal de 1%, une confiance minimale de 1%, une taille maximale de 5 éléments et une taille minimale de 2 éléments. 928 règles sont ainsi extraites.

5.3.4 Représentation de règles d'association en RDF

Lier les résultats, c'est-à-dire les règles extraites, à d'autres connaissances du domaine, permet de les explorer efficacement. Nous avons suivi notre approche de représentation, d'intégration et d'exploration des trajectoires de soins basée sur les technologies du Web Sémantique et les ontologies biomédicales (chapitre 4), mais cette fois appliquée à des règles d'association. Ainsi, une fois les règles extraites, elles ont été transformées aux standards du Web Sémantique, en RDF (code source 10).

```
#Une règle et ses mesures de qualité :
:Association_1 :has_antecedent :antecedent_1 ;
        :has_consequent :consequent_1 ;
        :has_support "0.01"^^xsd:float ;
        :has_confidence "0.132"^^xsd:float ;
        :has_lift "1.45"^^xsd:float .
#Son antécédent et ses éléments :
:antecedent_1 :has_item atc:A10BA02 ;
        :has_item atc:N02BE01 .
#Son conséquent et son élément :
:consequent_1 :has_item icd10:E11.98 .
```

Code source 10: Représentation d'une règle d'association en un graphe de triplets RDF, sérialisé en *turtle*.

La figure 5.7 offre une représentation graphique de la règle décrite par le code source ci-dessus.

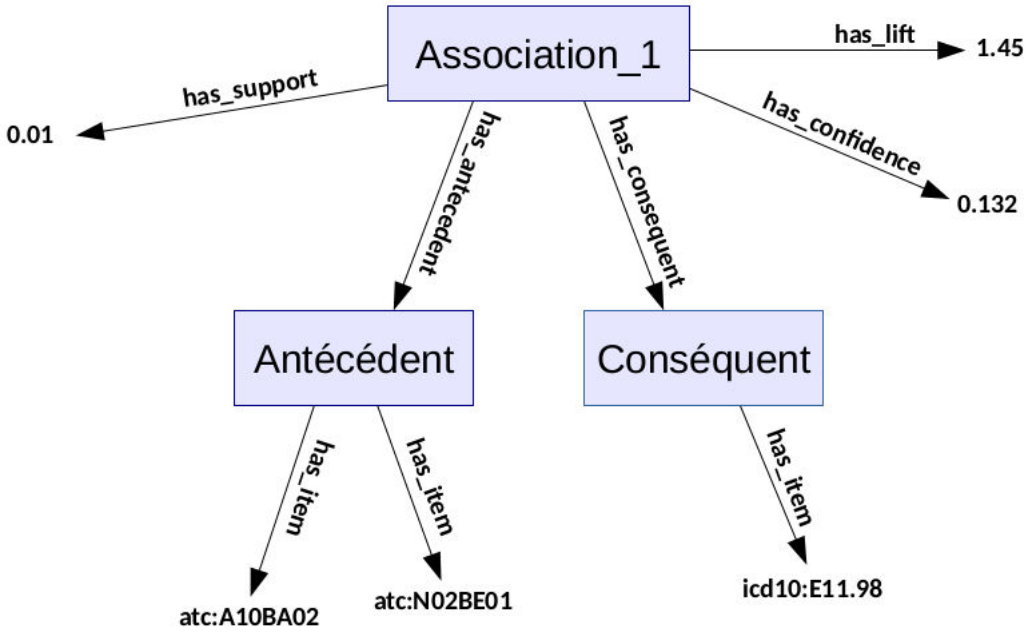

FIGURE 5.7 – Visualisation du graphe RDF associé à la règle d'association `Association_1`.

Les ontologies et bases de connaissances présentées en section 4.1.4.1 ont ensuite été reliées aux éléments des antécédents et conséquents, et ainsi aux règles d'association. Les bases de connaissances DID et DIKB ont été transformées en RDF pour l'application. Cette intégration d'ontologies médicales et pharmacologiques a alors permis d'explorer les trajectoires de soins selon divers critères nécessitant des connaissances du domaine médical et pharmacologique.

5.3.5 Exploration de règles d'association avec SPARQL

Nous nous sommes particulièrement intéressés à pouvoir déterminer les règles qui décrivaient une situation particulière, parmi les suivantes :

— Les règles contenant une interaction entre les médicaments à l'antécédent ;

— Les règles contenant une contre-indication médicamenteuse (à l'antécédent) ;

— Les règles contenant une contre-indication entre un médicament et un diagnostic (soit entre un élément de l'antécédent et l'élément du conséquent) ;

— Les règles contenant un médicament à l'antécédent et son indication au conséquent.

Le graphe RDF des règles extraites et les ontologies ont été chargées dans le *triplestore* FUSEKI. Des requêtes SPARQL permettent ensuite de mener à bien les explorations ci-avant. Le code source 11 permet ainsi de déterminer les règles

d'association dont un médicament à l'antécédent est indiqué dans le cadre de la maladie au conséquent, selon le graphe de connaissances DID.

```
SELECT DISTINCT ?regle
WHERE{
#La règle a un antécédent et un conséquent :
?regle :has_antecedent ?antecedent .
?regle :has_consequent ?consequent .
#Les éléments de l'antécédent et du conséquent :
?antecedent :has_item ?drug .
?consequent :has_item ?diagnostic .
#Ces éléments ont des classes ancêtres :
?drug rdfs:subClassOf* ?drug_ancestor .
?diagnostic rdfs:subClassOf* ?diagnostic_ancestor .
#Existance d'une relation d'indication entre médicament
#et diagnostic, ou entre leurs classes ancêtres :
?drug_ancestor DID:has_indication ?diagnostic_ancestor .}
```

Code source 11: Requête SPARQL pour l'identification des règles d'association contenant une relation d'indication entre l'antécédent et le conséquent, d'après DID.

Le code source 12 présente la requête SPARQL pour la recherche de règles d'association contenant une interaction médicamenteuse selon DIKB.

```
SELECT DISTINCT ?regle ?source ?CI
WHERE{
#La règle a un antécédent :
?regle :has_antecedent ?antecedent .
#L'antécédent est constitué de deux codes ATC différents :
?antecedent :has_item ?drug_1 .
?antecedent :has_item ?drug_2 .
FILTER(?drug_1 != ?drug_2)
#Ces codes ont des ancêtres :
?drug_1 rdfs:subClassOf* ?drug_1_ancestor .
?drug_2 rdfs:subClassOf* ?drug_2_ancestor .
#Qui interviennent dans une interaction médicamenteuse :
?drug_1_ancestor DIKB:ddi_interactor_in ?interaction .
?drug_2_ancestor DIKB:ddi_interactor_in ?interaction .
?interaction DIKB:has_source ?source .
?interaction DIKB:is_contraindicated ?CI .
}
```

Code source 12: Requête SPARQL pour l'identification des règles d'association contenant une interaction médicamenteuse d'après DIKB.

Cette requête permet également d'identifier les règles d'association contenant une contre-indication médicamenteuse à l'antécédent grâce à la variable `?CI`. Cette variable est une valeur booléenne, vraie lorsqu'une source présente dans DIKB juge l'interaction contre-indiquée.

La code source 13 présente la requête SPARQL pour l'identification des règles d'association contenant une contre-indication entre un médicament à l'antécédent et un diagnostic au conséquent, en utilisant les connaissances apportées par l'ontologie NDF-RT. Elle se construit de la même façon que la requête 9, présentée dans le chapitre 4.

```
SELECT DISTINCT *
WHERE {
#Contre-indications dans NDF-RT :
?ndf_med rdfs:subClassOf ?CI . #Médicament NDF
?CI owl:onProperty ndf:CI_with .
?CI owl:someValuesFrom ?ndf_diag . #Diagnostic ou état de santé NDF
#En plus de ?ndf_diag et ?ndf_med, toutes leurs sous-classes supportent la CI :
?subclass_ndf_diag rdfs:subClassOf* ?ndf_diag .
?subclass_ndf_med rdfs:subClassOf* ?ndf_med .
#Correspondance à l'ATC et à la CIM-10 via les CUI :
?subclass_ndf_diag ndf:UMLS_CUI ?cui_diag . #CUI
?ATC umls:cui ?cui_med . #ATC
?subclass_ndf_med ndf:UMLS_CUI ?cui_med . #CUI
?CIM10 umls:cui ?cui_diag . #CIM-10
#En plus de ?ATC et ?CIM10, toutes leurs sous-classes supportent la CI :
?atc_med rdfs:subClassOf* ?ATC .
?cim10_diag rdfs:subClassOf* ?CIM10 .

#On peut maintenant chercher les règles qui ont ?atc_med et ?cim10_diag :

#La règle a un antécédent et un conséquent :
?regle :has_antecedent ?antecedent .
?regle :has_consequent ?consequent .

#Les éléments de l'antécédent et du conséquent :
?antecedent :has_item ?atc_med .
?consequent :has_item ?cim10_diag .
}
```

Code source 13: Requête SPARQL pour l'identification des règles d'association contenant une contre-indication entre médicament et diagnostic selon l'ontologie NDF-RT.

5.3.6 Résultats

Comme le montre la figure 5.8, aucune règle contenant une contre-indication, que ce soit entre médicaments ou entre médicament et diagnostic, n'a été identifiée. Comme nous avons pu le montrer en section 4.1.5, des contre-indications sont bien présentes dans le jeu de données utilisé (paragraphe 4.1.2). Cependant, le paramètre d'un support minimal à 1% ne permet pas de les retrouver dans des règles d'association. On pourrait juger de triviales les 400 règles d'association contenant une relation d'indication entre un médicament à l'antécédent et un diagnostic au conséquent, et les écarter de l'analyse des résultats.

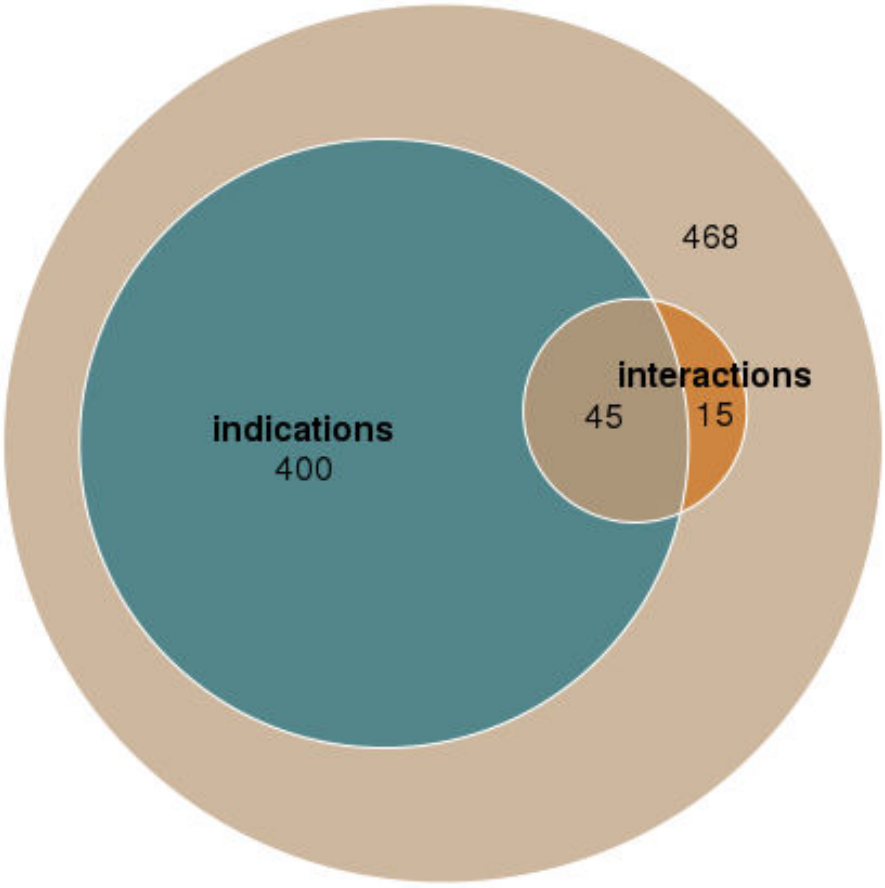

FIGURE 5.8 – Diagramme de Venn associé aux résultats des requêtes 11, 12 et 13.

En effet, ces règles décrivent en partie une association déjà connue. De plus, si on choisit de les écarter, la part non-triviale de l'association correspond à une autre règle jugée elle non-triviale. L'exemple présenté par la table 5.8 illustre bien cette caractéristique, pouvant d'ailleurs s'apparenter à de la redondance de règle (définie en section 5.2).

Antécédent	Conséquent	Support	Conf.
$\{atc{:}A02BC05, atc{:}N02BE01, atc{:}N05CF02\}$	$\{icd10{:}M17.1\}$	0.01005	0.52
$\{atc{:}A02BC05, atc{:}N05CF02\}$	$\{icd10{:}M17.1\}$	0.01315	0.53125

TABLE 5.8 – Deux règles imbriquées parmi les 928 règles extraites. La première appartient aux règles contenant une relation d'indication entre l'antécédent et le conséquent.

Le paracétamol ($atc{:}N02BE01$) est indiqué dans la prise en charge de l'arthrose du genou ($icd10{:}M17.1$) selon la base de connaissances DID. La deuxième appartient aux règles restantes. Bien que la première puisse être jugée triviale, la seconde décrit la part non-triviale de la première. Néanmoins, les règles contenant une indication et une interaction sont à conserver. En effet, la part déjà connue de la règle, c'est à dire l'indication, peut elle même constituer une partie de l'interaction. En l'écartant, on écarte alors aussi l'interaction. Le ciblage des règles contenant une interaction nous permet d'étudier les interactions médicamenteuses, et leurs associations avec des états de santé. Ces règles ainsi ciblées représentent de plus un ensemble restreint de règles, qu'il est possible d'analyser par un œil humain dans un temps raisonnable. Il est ainsi assez facile de regrouper les règles par interaction et/ou par groupes de diagnostics similaires. Deux règles contenant une interaction entre un paracétamol ($atc{:}N02BE01$) et de l'énoxaparine ($atc{:}B01AB05$) ont particulièrement retenu notre attention (figure 5.9 et 5.10).

$$\{atc{:}B01AB05, atc{:}N02BE01\} \Longrightarrow \{icd10{:}D64.8\}$$
$$\{atc{:}B01AB05, atc{:}N02BE01\} \Longrightarrow \{icd10{:}D62\}$$

FIGURE 5.9 – Règles d'association contenant une interaction entre un paracétamol ($atc{:}N02BE01$) et de l'énoxaparine ($atc{:}B01AB05$).

Les diagnostics associés à ces deux règles, une anémie post-hémorragique aiguë ($icd10{:}D62$) et une autre anémie précisée ($icd10{:}D64.8$) sont des diagnostics relativement similaires, et d'ailleurs proches dans la hiérarchie de la CIM-10. Si on peut être relativement sûr que $icd10{:}D62$ indique une hémorragie, le code $icd10{:}D64.8$ est plus incertain en ce qui concerne des éventuels saignements. La base de connaissances DIKB répertorie une interaction entre le paracétamol et l'énoxaparine, sans pour autant lui donner une explication. L'interaction entre le paracétamol et les anticoagulants est en effet encore peu connue, car peu répertoriée dans les dictionnaires de médicaments (Ornetti et al., 2005), par exemple le dictionnaire français Vidal. À l'inverse, l'interaction entre anticoagulants et l'aspirine ou les anti-inflammatoire non stéroïdien, bien connue pour augmenter le risque hémorragique chez le patient, a fait du paracétamol le traitement antalgique et antipyrétique de référence chez les patients recevant des anticoagulants (Mahé, 2004). Hylek et al. (1998) ont pourtant démontré que l'effet anticoagulant du paracétamol ainsi que son association avec d'autres anticoagulants, étaient sous-estimés.

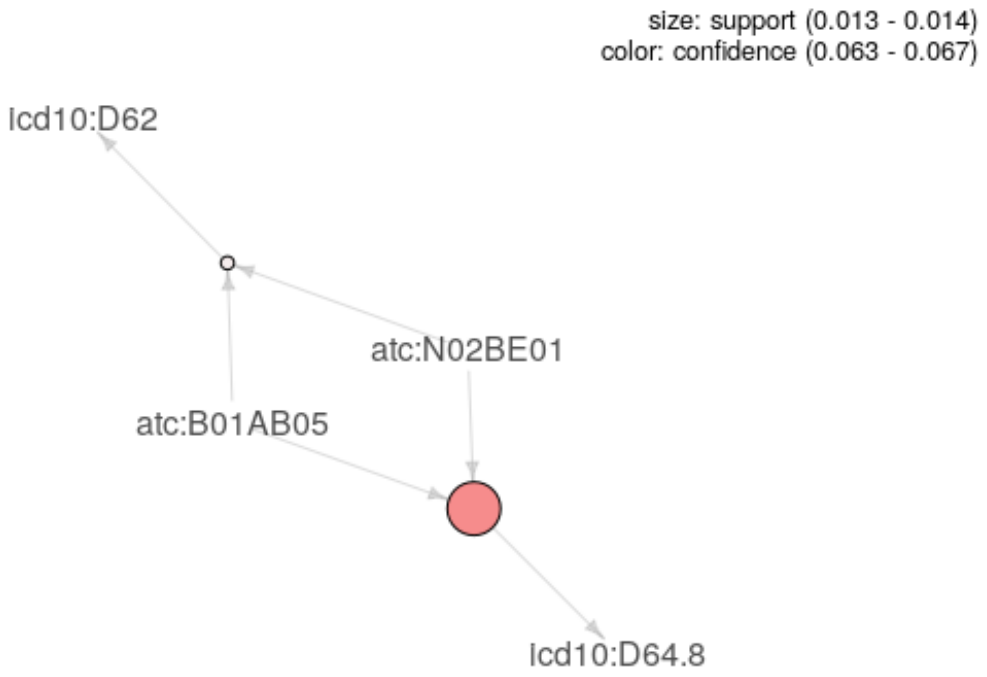

FIGURE 5.10 – Représentation de deux règles proches, contenant la même interaction à l'antécédent. Un cercle représente une règle. La taille des cercles reflète le support des règles, l'intensité du rouge sa confiance. Les flèches indiquent la direction des règles, des éléments de l'antécédent vers le conséquent.

L'étude de telles règles et de leurs mesures de qualité (figure 5.10), et notamment la comparaison à d'autres règles proches, pourrait permettre de mesurer le risque associé à certaines interactions mal connues. Dans notre exemple, les anticoagulants administrés seuls pouvant exposer à des risques de saignements, on pourrait vouloir comparer les deux règles (et leur confiance) présentées ci-avant aux règles suivantes :

$$\{atc{:}B01AB05, \neg atc{:}N02BE01\} \Longrightarrow \{icd10{:}D64.8\}$$
$$\{atc{:}B01AB05, \neg atc{:}N02BE01\} \Longrightarrow \{icd10{:}D62\}$$

FIGURE 5.11 – Règles "disjointes" aux règles présentées par la figure 5.10, contenant l'élément négatif $\neg atc{:}N02BE01$. La comparaison de certaines mesures de qualité de ces règles avec les précédentes pourrait permettre d'étudier le risque hémorragique de l'association entre l'énoxaparine et le paracétamol par rapport au risque hémorragique de l'énoxaparine seule.

5.3.7 Conclusion et discussion

Les technologies du Web Sémantique sont adaptées à la représentation et l'exploration de règles d'association médicales. L'utilisation d'ontologies biomédicales

du *Linked Data* permettent de baser cette exploration sur des connaissances du domaine médical et pharmacologique, et par exemple de rechercher dans un ensemble de règles des relations entre médicaments et états de santé. On peut d'ailleurs facilement imaginer une ontologie répertoriant plus largement les associations déjà répertoriées et étudiées, que ce soit entre médicaments, états de santé, actes médicaux, ou tout autre concept médical. A l'image de *Gene Ontology* (Carbon et al., 2017), les annotations d'une telle ontologie pourraient à la fois se baser sur la littérature disponible propre aux associations médicales (Lim et al., 2018), comme sur la collaboration de la communauté scientifique médicale, les pharmacologues, épidémiologistes et pharmaco-épidémiologistes par exemple.

L'approche étudiée dans ce travail permet de réduire le nombre de règles à analyser, soit en en écartant certaines déjà connues, soit en en ciblant certaines par la présence d'une relation sémantique d'intérêt. L'exploration proposée permet donc de réduire une des limites de l'extraction de règles d'association, celle d'en obtenir un nombre trop abondant. Ces relations entre concepts, telles les interactions, les indications et les contre-indications, sont cependant généralement répertoriées entre des concepts à la granularité fine. DIKB, ou encore le thésaurus des interactions médicamenteuses de l'ANSM, répertorient des interactions médicamenteuses majoritairement entre des codes de médicaments au plus bas de la hiérarchie de l'ATC. Dès lors, si combiner cette approche à celle de la section 5.2 –par l'extraction de règles multi-niveaux– semble être une bonne idée, très peu de relations particulières entre concepts médicaux seraient identifiées.

Après avoir étudié l'extraction de motif et l'exploration de motifs extraits, appliquées aux données médico-administratives, la section 5.4 étudie la recherche de motifs dans des ensembles de trajectoires de soins.

5.4 Reconnaissances de chroniques

5.4.1 Introduction

Lorsqu'elle est appliquée aux trajectoires de soins, l'extraction de chroniques cherche à extraire des motifs intéressants d'événements de soins, souvent des motifs partagés par un grand nombre de patients, c'est à dire des motifs fréquents. Cependant, des motifs bien moins fréquents, même rares, peuvent être intéressants. C'est particulièrement le cas en pharmaco-épidémiologie, en sécurité des soins ou encore en pharmacovigilance, où certaines maladies, des consommations de soins ou de médicaments inattendues, peuvent être des événements relativement rares. L'extraction de motifs rares et intéressants est cependant très limitée, comme on a pu le voir en section 5.2 avec l'extraction de règles d'association. Les algorithmes pour l'extraction de motifs, basés sur le support, impliquent un compromis entre l'extraction d'un nombre raisonnable de motifs et la découverte de motifs rares. Face aux limites de l'extraction de motifs, (Samet et al., 2017) se sont intéressés à la reconnaissance de motifs rares fournis par des experts plutôt qu'à leur extraction non-supervisée. Dans une base de données de séquences d'événements, des motifs rares définis par des experts du domaine sont ainsi recherchés. Le formalisme des chroniques, avec notamment la présence de contraintes temporelles entre événements, semble assez expressif pour représenter les motifs fournis par des experts. Les méthodes que l'on pourrait utiliser pour reconnaître des chroniques dans des ensembles de séquences doivent elles être assez expressives pour pouvoir traduire ces chroniques en requêtes, ainsi qu'assez rapides pour que la recherche puisse se faire de façon intéractive.

Comme souligné en section 4.1.3.2, les technologies du Web Sémantique sont appropriées à la recherche de séquences d'événements. RDF est adapté à la représentation de séquences d'événements, et SPARQL l'est tout autant pour la recherche de motifs temporelles tels que des chroniques.

Dans le cadre d'une collaboration avec Thomas Guyet, chercheur au sein du consortium PEPS, nous nous sommes intéressés à explorer différentes méthodes pour la reconnaissance de chroniques dans des bases de données de séquences d'événements, constituées à partir des bases médico-administratives françaises. Thomas Guyet apportait une première méthode basée sur le paradigme de programmation déclarative d'*Answer Set Programming* (ASP) (Guyet et al., 2017), ainsi qu'un algorithme dédié à la reconnaissance de chroniques. Nous présentons dans cette section l'approche que nous avons apportée à cette collaboration : une méthodologie basée sur l'utilisation des technologies du Web Sémantique, en l'occurrence RDF et SPARQL.

5.4.2 Objectifs

L'objectif de cette collaboration est de comparer différentes approches dans la reconnaissance de chroniques sur une base de données d'événements pouvant être issus des bases médico-administratives. La comparaison devait tout d'abord traiter de l'efficacité des méthodes, en terme de résultat et de rapidité d'exécution. Enfin,

nous souhaitions également pouvoir comparer l'expressivité des méthodes compte tenu du contexte particulier des données médico-administratives.

5.4.3 Données et outils

Dans cette comparaison, nous considérons les suppositions suivantes :

— La reconnaissance d'une chronique dans un ensemble de trajectoires a un temps d'exécution qui est proportionnel au nombre de trajectoires ;

— Une optimisation de la reconnaissance de chronique n'est pas rendue possible par la recherche de plusieurs chroniques.

En conséquence à ces *a priori*, nous nous intéressons à reconnaître des chroniques uniques dans des trajectoires également uniques, de taille variable. Les données qui ont permis de réaliser cette étude, les trajectoires comme les chroniques, ont été simulées, comme présenté ci-après.

Chroniques Les éléments des chroniques (figure 5.12) sont constituées de classes de médicaments de l'ATC. Deux comparaisons ont été réalisées. Dans un premier temps, les chroniques étaient constituées uniquement de classes feuilles dans la hiérarchie de l'ATC. Dans la deuxième comparaison, le niveaux de granularité des éléments des chroniques varient. Il est tiré aléatoirement selon un loi dont les probabilités sont inversement proportionnelles au degré de granularité de l'ATC. Une fois le degré de granularité fixé, une classe de l'ATC pour ce niveau de hiérarchie est tirée aléatoirement selon une loi de probabilité dont la distribution suit des fréquences de classes ATC observées dans un jeu de données issu du SNDS. Des contraintes temporelles sont générées selon une distribution uniforme. Les paramètres de la simulation de chroniques sont la taille moyenne des chroniques (*i.e.* le nombre d'éléments) et le taux de contraintes temporelles des chroniques générées.

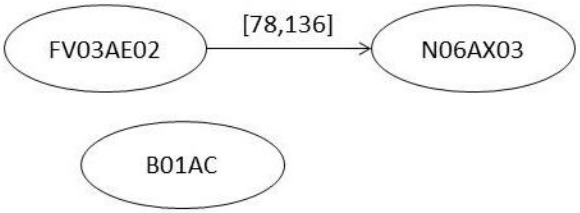

FIGURE 5.12 – Chronique constituées de trois médicaments codés selon l'ATC, à différents degrés de granularité, et d'une contrainte temporelle.

Trajectoires Les trajectoires dans lesquelles on recherche des chroniques sont ensuite générées en fonction des chroniques générées. Des chroniques générées sont en effet dissimulées dans une partie des trajectoires générées. Les éléments constituant les trajectoires sont toutes des feuilles de la hiérarchie de l'ATC, générées aléatoirement selon une loi de probabilité dont la distribution suit encore une fois

des fréquences de classes ATC observées dans un jeu de données issu du SNDS. Les délais entre éléments sont générés selon une distribution de type gaussien. Les paramètres de cette simulation de trajectoires sont la taille moyenne des séquences (*i.e.* le nombre d'éléments), leur taille temporelle et le taux minimum de trajectoires dissimulant une chronique.

5.4.3.1 Les technologies du Web Sémantique pour la reconnaissance de chroniques

Une fois générées, les trajectoires sont transformées en graphe RDF (figure 14).

```
patdb:patient0 patdb:hasEvent patdb:patient0evt0 .
patdb:patient0evt0 rdf:type patdb:DrugDelivery ;
        patdb:drugDelivered atc:C03EB01 ;
        patdb:deliveryDate '38'^^xsd:integer .

patdb:patient0 patdb:hasEvent patdb:patient0evt1 .
patdb:patient0evt1 rdf:type patdb:DrugDelivery ;
        patdb:drugDelivered atc:B01AC06 ;
        patdb:deliveryDate '47'^^xsd:integer .

patdb:patient0 patdb:hasEvent patdb:patient0evt2 .
patdb:patient0evt2 rdf:type patdb:DrugDelivery ;
        patdb:drugDelivered atc:B03BB01 ;
        patdb:deliveryDate '76'^^xsd:integer .

patdb:patient0 patdb:hasEvent patdb:patient0evt3 .
patdb:patient0evt3 rdf:type patdb:DrugDelivery ;
        patdb:drugDelivered atc:N05BA01 ;
        patdb:deliveryDate '113'^^xsd:integer .
```

Code source 14: Graphe RDF d'une trajectoire, écrit en *turtle*.

Les chroniques sont traduites par des requêtes SPARQL (code source 15). Elles ont alors été exécutées en utilisant la librairie java de *Jena*[2], ainsi que la librairie java *Virtuoso Jena Provider*[3]. Cette deuxième permet d'exécuter des requêtes SPARQL sur un triplestore de Virtuoso.

2. Le site du projet *Jena* : https://jena.apache.org/
3. *Virtuoso Jena Provider* sur le site de Virtuoso : http://vos.openlinksw.com/owiki/wiki/VOS/VirtJenaProvider

```
SELECT DISTINCT ?patient WHERE{
        ?patient patdb:hasEvent ?evt0 .
        ?evt0 rdf:type patdb:DrugDelivery .
        ?evt0 patdb:deliveryDate ?date0 .
        ?evt0 patdb:drugDelivered atc:V03AE02 .

        ?patient patdb:hasEvent ?evt1 .
        ?evt1 rdf:type patdb:DrugDelivery .
        ?evt1 patdb:deliveryDate ?date1 .
        ?evt1 patdb:drugDelivered ?atc1 .
        ?atc1 rdfs:subClassOf* atc:B01AC .

        ?patient patdb:hasEvent ?evt2 .
        ?evt2 rdf:type patdb:DrugDelivery .
        ?evt2 patdb:deliveryDate ?date2 .
        ?evt2 patdb:drugDelivered atc:N06AX03 .

        FILTER ( ?date2 - ?date0 >= 78.0)
        FILTER ( ?date2 - ?date0 <= 136.0)
}
```

Code source 15: La chronique 5.12, écrite comme une requête SPARQL.

5.4.4 Résultats

Reconnaissance les résultats obtenus avec la librairie java *Jena* sont identiques à ceux obtenus avec les autres approches, les résultats obtenus avec la librairie java *Virtuoso Jena Provider* varient. En effet, *Virtuoso Jena Provider* n'identifie pas toutes les chroniques qui sont à reconnaître. Ces chroniques manquées sont toutes des chroniques avec au moins une classe de l'ATC qui ne soit pas une feuille de l'arbre de l'ATC (telles que la chronique présentée en figure 5.12). Janke et al. (2017) ont montré que l'implémentation de SPARQL dans *Virtuoso* ne gérait pas parfaitement le *property path* de SPARQL, c'est-à-dire la propriété permettant aux requêtes de parcourir de façon transitive un graphe RDF (comme présenté par le code source 7 du chapitre 4). Nous avons par conséquent choisi de continuer les comparaisons avec la librairie java de *Jena* uniquement.

Temps d'exécution Les comparaisons ont été réalisées selon différentes valeurs des paramètres, de la taille moyenne des séquences et de la taille temporelle moyenne des séquences. Les chroniques sont pour le moment constituées de quatre codes ATC, feuilles de l'arbre de l'ATC. Dans ces comparaisons, les temps d'exécution sont meilleurs avec les méthodes apportées par Thomas Guyet qu'avec SPARQL (figure 5.13). Son algorithme spécialement dédié à la reconnaissance de chroniques est légèrement meilleur que sa méthode basée sur le paradigme ASP.

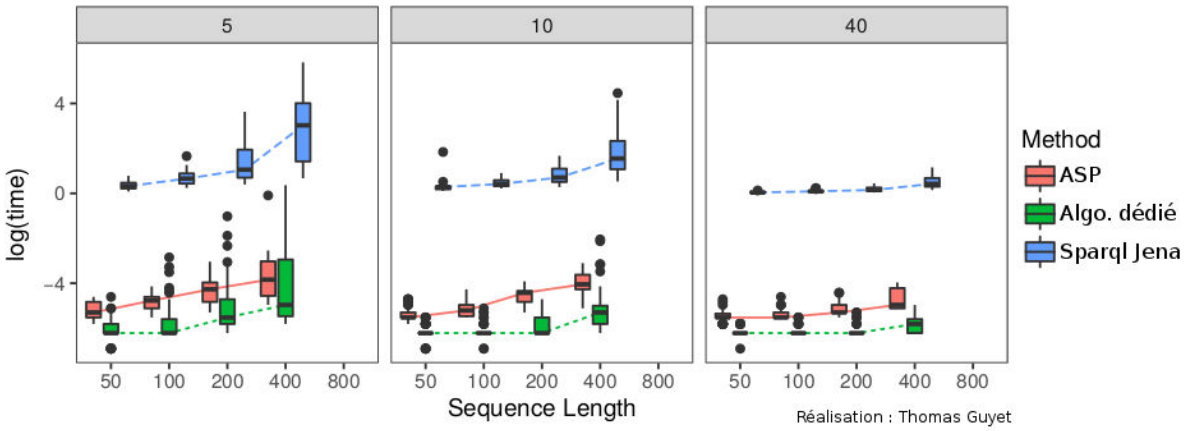

FIGURE 5.13 – Diagrammes en boîte des temps d'exécution pour la reconnaissance de chroniques. Plusieurs exécutions sont réalisées selon la taille temporelle moyenne des séquences (en ordonnée) et la taille moyenne des séquences (5, 10 et 40). Les chroniques sont en revanche toutes composées de 4 médicaments codés selon l'ATC. Le logarithme des temps d'exécution (en secondes) est représenté en ordonnée.

Chroniques et hiérarchie La deuxième partie de cette collaboration, toujours en cours, implique la reconnaissance de chroniques lorsque leurs éléments ne sont pas nécessairement au degré le plus fin de la nomenclature de l'ATC. La reconnaissance de telles chroniques requiert des modifications de l'algorithme dédié et de la méthode basée sur ASP. Pour le moment, les chroniques ont été simplement "aplaties", dans le sens où chaque chronique contenant un code nœud de l'ATC a été traduite en toutes ses sous-chroniques composées uniquement de feuilles de l'ATC. Une chronique avec plusieurs code nœuds de l'ATC, à des degrés de granularité plus ou moins élevés, peut alors mener à une explosion du nombre de chroniques à reconnaître. La méthode basée sur les technologies du Web Sémantique permet elle de gérer bien plus naturellement la reconnaissance de chroniques composées de code nœuds de l'ATC, avec une seule requête, comme le montre ci-avant le code source 15.

5.4.5 Conclusion et perspectives

Cette collaboration devrait se poursuivre par une adaptation de l'algorithme dédié par Thomas Guyet pour une meilleur gestion des hiérarchies médicales. Cette adaptation est en effet complexe à implémenter en ASP. En ce qui concerne l'approche basée sur les technologies du Web Sémantique, bien qu'elle ne se soit pas illustrée par sa rapidité, elle pourrait s'avérer la plus performante dans la reconnaissance de chroniques particulières, nécessitant un apport de connaissances externes. Un outil interactif de reconnaissance de chroniques à destination de professionnels de la santé, ou d'analystes de données santé, pourrait ainsi être enrichi de la re-

cherche de relations sémantiques entre concepts, telles que des interactions médicamenteuses, des contre-indications ou encore des relations d'indications. Il serait alors pertinent de réaliser des expérimentations supplémentaires en intégrant d'autres ontologies que l'ATC. De telles perspectives demanderaient cependant des adaptations majeures aux autres méthodes

5.5 Synthèse

Face aux nombreuses complexités à traiter des trajectoires de soins issues des bases de données médico-administratives, la deuxième partie de la thèse a permis de démontrer l'intérêt de considérer l'enrichissement de méthodes d'analyse et de fouille de données par des connaissances externes. En effet, des limites sont propres aux méthodes, et d'autres le sont aux trajectoires de soins et également aux données médico-administratives. Si ces limites sont généralement contournées par l'utilisation de pré-traitement des données, ce travail s'est plutôt focalisé sur la modification de méthodes d'analyse et de fouille de séquences pour les rendre adaptées aux trajectoires de soins et aux données médico-administratives. Ces approches fournissent des pistes pour s'affranchir des pré-traitements habituels, qui sont généralement réalisés au détriment de l'information portée par les trajectoires de soins, ou qui renforcent le caractère supervisé des méthodes utilisées.

Particulièrement, les différentes approches proposent des modifications de méthodes de fouille et d'analyse de séquences pour prendre en compte les structures hiérarchiques des nomenclatures servant à la codification des données médico-administratives. L'approche ainsi décrite dans la section 5.1 a été présentée à plusieurs reprises lors de conférences scientifiques (Rivault et al., 2017b,a). Un article de cinq pages (Rivault et al., 2017b) a également été publié dans les *Lecture Notes in Computer Science*[4] (annexe 6). Outre cet enrichissement de méthodes existantes, nous proposons d'appliquer la méthodologie présentée dans le chapitre 4 pour explorer les résultats, parfois abondants avec l'utilisation telles méthodes, selon des critères nécessitant l'apport de connaissances externes aux données. Les travaux entamés lors de collaborations nécessitent d'être approfondis. Leur valorisation scientifique est en cours.

4. LNCS sur le site de Springer : https://www.springer.com/gp/computer-science/lncs

Chapitre 6

Conclusion

Dans l'objectif de développer des outils et méthodes génériques pour la réutilisation des données médico-administratives à des fins de recherche en santé publique, les travaux menés dans le cadre de cette thèse se sont concentrés sur l'intégration de connaissances du domaine médical aussi bien lors de l'exploration que lors de l'analyse de ces données. Les approches mises en œuvre s'attachent également à considérer les données de chaque patient selon le paradigme récent des trajectoires de soins. Cette vision d'une trace globale des consommations et états de santé des patients requiert des méthodes adaptées, afin de considérer au mieux le caractère multi-disciplinaire de la prise en charge des patients dans divers services de santé.

Les technologies du Web Sémantique sont pertinentes pour mettre en œuvre l'exploration de trajectoires de soins, lorsqu'elles sont issues des données volumineuses des bases médico-administratives (Rivault et al., 2015, 2016a). L'intégration de plusieurs ontologies médicales et pharmacologiques a permis d'élargir le champs de l'exploration de données à des critères complexes nécessitant un apport de connaissances externes. Ce travail a notamment permis d'identifier des relations d'interaction entre médicaments, d'indication entre médicaments et diagnostics, de contre-indication entre médicaments et diagnostics ou encore entre médicaments (Rivault et al., 2018a). Cette exploration enrichie de connaissances ouvre de nouvelles voies dans la réutilisation des données médico-administratives pour la recherche en santé publique. En pharmacovigilance, elle pourrait en effet permettre d'identifier des consommations de soins non conformes ou inattendues, et ce sur l'ensemble de la population bénéficiaire de l'Assurance Maladie. En pharmaco-épidémiologie, une telle exploration peut permettre de mieux prendre en compte les relations entre médicaments, actes médicaux et états de santé, dans l'étude des potentiels risques et bénéfices des produits de santé et de leurs conditions d'usage. L'approche présentée repose sur le développement continu d'ontologies décrivant le domaine médical et pharmacologique, mais c'est aussi la multitude de ces sources de connaissances qui peut constituer une limite. La dispersion de ces sources sur le *Linked Data* et leur hétérogénéité rendent leur accès et leur utilisation conjointe compliquée. Dans la réutilisation des bases de données médico-administratives françaises, on peut également déplorer un manque d'ontologies pensées spécifiquement autour des nomenclatures utilisées, telles que la CCAM ou la NABM.

Si l'amélioration des études de santé publique réutilisant les données médico-administratives peut donc se faire grâce à une exploration des trajectoires de soins

enrichie de connaissances médicales et pharmacologiques, nous proposons également des pistes pour améliorer l'étape d'analyse des trajectoires de soins. Dans le cadre de cette thèse et de plusieurs collaborations, nous avons proposé à nouveau de considérer l'intérêt d'introduire des connaissances médicales et pharmacologiques. Nous avons adapté plusieurs méthodes d'analyse et de fouille de séquences pour prendre en compte la connaissance relative aux structures hiérarchiques des nomenclatures médicales codant les données médico-administratives. Ces méthodes, de plus en plus utilisées pour comparer des trajectoires de soins, les regrouper, ou pour en extraire des motifs de trajectoires, se heurtent souvent à la complexité des données médico-administratives, volumineuses et très variables. Des méthodes de comparaison de séquences, appliquées aux trajectoires de soins, ont notamment tendance à sous-estimer la ressemblance entre deux trajectoires. Les méthodes d'extraction de motifs ont elles tendance à extraire des motifs peu intéressants, en plus d'être abondants. Nos expérimentations ont montré que la prise en compte de la structure hiérarchique des nomenclatures médicales utilisées dans les bases de données médico-administratives, par exemple sous la forme de similarités sémantiques entre les composantes des trajectoires de soins, permet de réduire ces limites (Rivault et al., 2017a,b). Nous avons également utilisé les technologies du Web Sémantique et des ontologies biomédicales afin d'explorer les résultats de telles méthodes. Ces travaux montrent ainsi que des méthodes d'analyse et de fouille de séquences peuvent être adaptées aux objets que sont les trajectoires de soins, mais que des modifications –pouvant reposer sur l'intégration de connaissance externes aux données– sont nécessaires pour prendre en compte leurs complexités.

Durant cette thèse, et notamment grâce à la valorisation scientifique de ces deux axes de recherche, nous avons pu constater que ces approches d'exploration et d'analyse des trajectoires de soins n'étaient que peu envisagées par les chercheurs de santé publique utilisant les données médico-administratives. Bien que le paradigme des trajectoires de soins soit de plus en plus utilisé, l'adaptation des méthodes aux complexités des données n'est que peu fréquente. Actuellement, des explorations de données laborieuses et limitées ainsi que des étapes de pré-traitement des données leurs sont encore préférées. Dans l'optique de favoriser la réutilisation de ces approches par les chercheurs en santé publique, nous avons développé le package R *queryMed* (Rivault et al., 2018c). Il permet à des non-experts des technologies du Web Sémantique et des ontologies de récupérer des connaissances médicales du *Linked Data*, de les lier à des données de santé, et d'explorer ces données en utilisant les connaissances récupérées (Rivault et al., 2018b,a). Si *queryMed* rend ainsi plus accessible l'enrichissement de l'exploration des trajectoires de soins issues des bases de données médico-administratives, il ne facilite pas encore l'enrichissement de méthodes pour leur analyse. Les futurs développements pourraient fournir des clefs pour une adaptation plus aisée des méthodes d'analyse et de fouille de séquences en pharmaco-épidémiologie. En effet, bien que les utilisateurs des données médico-administratives ne soient majoritairement pas prêts à modifier en profondeur des méthodes existantes, ou à en développer de nouvelles, des mesures de similarité sé-

mantique entre concepts médicaux pourraient permettre des enrichissements plus accesibles. Des prochaines fonctionnalités de *queryMed*, en fournissant différentes mesures de similarités sémantiques calculées sur les nomenclatures médicales utilisées dans les bases médico-administratives, pourraient favoriser la diffusion de ces approches.

Les bases de données médico-administratives, du fait de leur objectif financier et managérial, n'étaient initialement pas destinées pour la recherche en santé publique. La couverture à l'échelle d'un pays en font néanmoins une source d'information sans égal. L'annonce de la création d'un « Health Data Hub » [1] par la Ministre des solidarités et de la santé va dans ce sens. Dans le rapport de ce « Health Data Hub », les données de santé y sont décrites comme des « gisements », riches mais difficiles à atteindre et à utiliser. Des acteurs des données de santé y décrivent ainsi de nombreuses difficultés à les analyser. L'hétérogénéité et la qualité des données y sont notamment critiquées. Une faible interopérabilité sémantique des données ainsi qu'un manque de terminologies de référence y sont également décrites. Ces limites font que les approches exploitées dans le cadre de cette thèse sont finalement peu envisagées sur des données de santé, et particulièrement sur les données médico-administratives. Le « Health Data Hub », dans son ambition d'appuyer la définition et l'implémentation de terminologies, formats et standards communs, et de partager des librairies et outils pour le traitement sémantique des données de santé, pourrait permettre la dissémination de ces approches.

1. Le rapport du « Health Data Hub » : https://solidarites-sante.gouv.fr/IMG/pdf/181012_-_rapport_health_data_hub.pdf

Valorisation scientifique

Rivault, Y., Dameron, O., and Le Meur, N. Une infrastructure générique basée sur les apports du Web Sémantique pour l'analyse des bases médico-administratives. In *Plate-forme Intelligence Artificielle 2015, conférence Ingénierie des Connaissances*, Rennes, France, June 2015. (Cité en pages 69 et 115.)

Rivault, Y., Dameron, O., and Le Meur, N. La gestion de données médico-administratives grâce aux outils du Web sémantique. In *Journées ADELF-EMOIS " Système d'information hospitalier et Epidémiologie "*, volume 64 of *Revue d'Épidémiologie et de Santé Publique*, page S15, Dijon, France, March 2016a. doi : 10.1016/j.respe.2016.01.051. (Cité en pages 69 et 115.)

Rivault, Y., Le Meur, N., and Dameron, O. Complications post-opératoires et mode de prise en charge en angioplastie : apport du Programme de Médicalisation des Systèmes d'Information (PMSI). In *Congrès Adelf-Epiter 2016*, Congrès Adelf-Epiter 2016, Rennes, France, September 2016b. (Cité en page 47.)

Rivault, Y., Le Meur, N., and Dameron, O. *Mesures de similarité entre trajectoires de soins issues de bases de données médico-administratives*. March 2017a. Published : 8 èmes rencontres scientifiques du réseau doctoral en santé publique. (Cité en pages 114 et 116.)

Rivault, Y., Le Meur, N., and Dameron, O. A Similarity Measure Based on Care Trajectories as Sequences of Sets. In *Artificial Intelligence in Medicine*, Lecture Notes in Computer Science, pages 278–282. Springer, Cham, June 2017b. ISBN 978-3-319-59757-7 978-3-319-59758-4. doi : 10.1007/978-3-319-59758-4_32. (Cité en pages 114, 116 et 133.)

Rivault, Y., Dameron, O., and Le Meur, N. Ontologies biomédicales et Web Sémantique pour la réutilisation des bases de données médico-administratives en pharmaco-épidémiologie. In *7ème Journées Francophones sur les Ontologies*, Hammamet, Tunisia, November 2018a. (Cité en pages 69, 115 et 116.)

Rivault, Y., Dameron, O., and Le Meur, N. queryMed : enrichissement des données de recherche en pharmaco-épidémiologie. In *Septièmes rencontres R*, pages 1–2, Rennes, France, July 2018b. (Cité en pages 69 et 116.)

Rivault, Y., Dameron, O., and Le Meur, N. queryMed : Release 0.5.1, September 2018c. (Cité en pages 42, 66, 69, 87 et 116.)

Annexes

queryMed

Vignette de queryMed

Incluse dans le package, ce document vise à expliquer aux nouveaux utilisateurs de queryMed toutes ses fonctionnalités. Disponible au format .Rmd [2], la vignette contient du code R relatifs aux principales fonctions du package, ainsi que du texte expliquant son exécution.

2. La vignette de queryMed au format .Rmd sur github : https://github.com/yannrivault/queryMed/blob/master/queryMed/vignettes/How-to-queryMed.Rmd

queryMed package: how to annotate medicine and pathology codes for pharmaco-epidemiological studies

Y. Rivault, O.Dameron, and N. Le Meur

03/08/2018

Introduction :

Because medical data, for example drugs and diseases, is often codified according to international nomenclatures, it can be linked to knowledge representations from medical and pharmacological domains. This can help improving data analysis by enriching the information it contains, for example by mining drug-drug interactions in a database of drug consumption (Pathak, Kiefer, and Chute 2013).

Semantic Web technologies and Linked Data initiatives have led to the spread of knowledge representations through ontologies, thesauri, taxonomies and nomenclatures. By providing standards and technologies for knowledge representation, integration and interrogation, the Semantic Web supports both technical and semantic interoperability for knowledge sharing and reuse. But if several Linked Data initiatives have published medical and pharmacological ontologies, the use of these standards, technologies and knowledge representations is still hesitant by the statisticians who deal with healthcare data (Ferreira et al. 2013).

The queryMed package purpose is to provide a user-friendly way to accesss the main medical and pharmacological knowledge sources from the Linked Data, through R, and linking them to healthcare data, so that the biostatisticians, epidemiologist and pharmaco-epidemiologists could enrich the data they analyze.

Installing queryMed

To retrieve and install queryMed, for the first time through github, you can use *devtools* R package:

```
install.packages("devtools")
devtools::install_github("yannrivault/queryMed/queryMed@queryMed_0.5.1")
```

To load queryMed call the *library()* function:

```
library(queryMed)
```

SPARQL

SPARQL is one of the standards from the Semantic Web. It allows to query knowledge and data written in the Semantic Web representation standards (e.g. RDF and OWL). Some remote servers, called SPARQL endpoints, give access to such data and knowledge. As you might have already guessed, it can be queried with SPARQL. There are many SPARQL endpoints that are fully or partly dedicated to biomedical knowledge : BioPortal(Salvadores et al. 2013), Bio2rdf (Callahan et al. 2013), Ontobee (Ong et al. 2017) or also DB-pedia (Lehmann et al. 2015).

queryMed offers an elementary function to send SPARQL queries over SPARQL endpoints from the Web.

Here is an example of a SPARQL query, sent on bio2rdf :

```
query=
"SELECT DISTINCT *
  WHERE  {
    ?db <http://bio2rdf.org/drugbank_vocabulary:x-atc> ?atc .
```

```
    ?db dcterms:title ?title .
    ?db rdfs:label ?label .
    ?db dcterms:description ?description .
    ?db <http://bio2rdf.org/drugbank_vocabulary:category> ?category .
  }
limit 5
"

res=sparql(query,url="http://bio2rdf.org/sparql")
```

```
## Querying http://bio2rdf.org/sparql
```

```
res
```

```
## # A tibble: 5 x 6
##   db      atc     title label  description                      category
##   <chr>   <chr>   <chr> <chr>  <chr>                            <chr>
## 1 http:/~ http:/~ Algl~ Alglu~ Human Beta-glucocerebrosidase o~ http://bi~
## 2 http:/~ http:/~ Laro~ Laron~ Human recombinant alpha-L-iduro~ http://bi~
## 3 http:/~ http:/~ Cycl~ Cyclo~ A cyclic undecapeptide from an ~ http://bi~
## 4 http:/~ http:/~ Cycl~ Cyclo~ A cyclic undecapeptide from an ~ http://bi~
## 5 http:/~ http:/~ Cycl~ Cyclo~ A cyclic undecapeptide from an ~ http://bi~
```

If Uniform Resource Identifier (URI) is a standard in the Semantic Web, it is not so convenient from a statistician point of vue. Let's turn it into normal data with *uri2norm()*.

```
uri2norm(res)
```

```
## # A tibble: 5 x 6
##   db      atc     title        label  description                 category
##   <chr>   <chr>   <chr>        <chr>  <chr>                       <chr>
## 1 DB00088 A16AB01 Alglucerase  Alglu~ Human Beta-glucocerebrosid~ Enzyme~
## 2 DB00090 A16AB05 Laronidase   Laron~ Human recombinant alpha-L-~ Enzyme~
## 3 DB00091 L04AD01 Cyclosporine Cyclo~ A cyclic undecapeptide fro~ Antifun~
## 4 DB00091 L04AD01 Cyclosporine Cyclo~ A cyclic undecapeptide fro~ Enzyme~
## 5 DB00091 L04AD01 Cyclosporine Cyclo~ A cyclic undecapeptide fro~ Antirhe~
```

The query results give some informations about drugs that are both codified according to DrugBank and the Anatomical Therapeutic and Chemical classification (ATC).

But querying knowledge and data through SPARQL endpoints on the Web requires an expertise in SPARQL syntax, the knowledge of potential useful SPARQL endpoints and also the representation of the knowledge they contain. This is probably why their use remains shy in some domains, for example in epidemiology and more generally in public health.

queryMed provides predefined SPARQL queries dedicated to medical and pharmacological domains –drugs and diseases– embeded in R functions.

Could we retrieve some information about the drugs present in a healthcare database ?

The example dataset *drug_set* is a dataframe that contains patients Id and prescribed drugs, codified according to the ATC.

```
data(drug_set)
drug_set[1:5,1:2]
```

```
##   patient     ATC
## 1       1 B01AC04
## 2       1 B01AC04
## 3       1 B01AC06
## 4       1 B01AC06
## 5       1 B03AA02
```

To retrieve some information about drugs we could call *bio2rdf()* or *dbpedia()*. These functions send predefined queries on Bio2RDF and DBpedia.

```
bio2rdf <- uri2norm(bio2rdf_db(lang="en"))
```

```
## Querying http://bio2rdf.org/sparql
```

```
dbpedia <- uri2norm(dbpedia_drug(lang="en"))
```

```
## Querying https://dbpedia.org/sparql
```

And then we could apply a filter on the drug present in our database :

```
drug_set_bio2rdf <- bio2rdf[bio2rdf$atc %in% drug_set$ATC,]
drug_set_dbpedia <- dbpedia[dbpedia$atc %in% drug_set$ATC,]
head(drug_set_bio2rdf)
```

```
## # A tibble: 6 x 6
##   db      atc     title        label  description                category
##   <chr>   <chr>   <chr>        <chr>  <chr>                      <chr>
## 1 DB00091 L04AD01 Cyclosporine Cyclo~ A cyclic undecapeptide fro~ Antifun~
## 2 DB00091 L04AD01 Cyclosporine Cyclo~ A cyclic undecapeptide fro~ Enzyme-~
## 3 DB00091 L04AD01 Cyclosporine Cyclo~ A cyclic undecapeptide fro~ Antirhe~
## 4 DB00091 L04AD01 Cyclosporine Cyclo~ A cyclic undecapeptide fro~ Dermato~
## 5 DB00091 L04AD01 Cyclosporine Cyclo~ A cyclic undecapeptide fro~ Immunos~
## 6 DB00099 L03AA02 Filgrastim   Filgr~ Filgrastim is a recombinan~ Hematop~
```

```
head(drug_set_dbpedia)
```

```
## # A tibble: 6 x 6
##   drug          atc     db      abstract                 comment     label
##   <chr>         <chr>   <chr>   <chr>                    <chr>       <chr>
## 1 Clozapine     N05AH02 DB00363 Clozapine, sold under t~ Clozapine,~ Cloz~
## 2 Valproate     N03AG01 DB00313 Valproate (VPA), and it~ Valproate ~ Valp~
## 3 Sulfasalazine A07EC01 DB00795 Sulfasalazine (SSZ), ma~ Sulfasalaz~ Sulf~
## 4 Amitriptyline N06AA09 DB00321 Amitriptyline, sold und~ Amitriptyl~ Amit~
## 5 Ergotamine    N02CA02 DB00696 Ergotamine is an ergope~ Ergotamine~ Ergo~
## 6 Itraconazole  J02AC02 DB01167 Itraconazole (code name~ Itraconazo~ Itra~
```

But drugs are not always codified according to the ATC nomenclature. Linked Data initiatives have made significant efforts to provide links –or mappings– between the main nomenclatures. For example, the Concept Unique Identifier(CUI) from the Unified Medical Language System has been used to annotate codes of drugs and diagnoses from several nomenclatures. This kind of mapping is not always easy to use. We provide a function, *mapping_cui()*, that allows to search for a CUI mapping between medical terms. Because the function programmatically accesss to BioPortal API (Whetzel et al. 2011) to search for a potential mapping, it needs an API key. To have one, you need to register at BioPortal.

```
drug_ATC_NDFRT <- mapping_cui(codes=drug_set$ATC,
                              ontologies_source="ATC",
                              ontologies_target="NDFRT",
                              api_key="your_api_key")
head(drug_ATC_NDFRT)
```

It gives you a mapping between ATC codes from *drug_set* and the National Drug File - Reference Terminology (NDF-RT), using CUI, when it exists (an ATC or NDF-RT code usually leads to at least one CUI code, but a term in NDF-RT could not exist in ATC, or vice versa).

Then we can merge this mapping table to our initial database :

```
drug_set_ATC_CUI_NDFRT <- merge(drug_set[,c("patient","ATC")],
                                drug_ATC_NDFRT,by.x="ATC",
                                by.y="source",
                                all.x=T)
head(drug_set_ATC_CUI_NDFRT)
```

This allows to extend the approach to the main medical and pharmacological nomenclatures. For the next examples, we added NDF-RT mapping in *drug_set* and *disease_set* databases.

This is for the moment quite simple annotation. queryMed offers also the possibility to retrieve more complex informations, such as drug interactions, drug-disease contraindications and drug indications.

Drug-disease contraindications from the National Drug File - Reference Terminology

The *NDFRT_CI_with()* function send a SPARQL query on Ontobee SPARQL endpoint to retrieve contraindications between drugs and diseases :

```
NDFRT_CI <- NDFRT_CI_with()
```

```
## Querying http://sparql.hegroup.org/sparql/
```

```
NDFRT_CI <- uri2norm(NDFRT_CI)
head(NDFRT_CI)
```

```
## # A tibble: 6 x 6
##   ndf_drug    cui_drug label_drug       ndf_diag cui_diag label_diag
##   <chr>       <chr>    <chr>            <chr>    <chr>    <chr>
## 1 N0000020091 C0014704 ERGONOVINE       N000000~ C0000821 Abortion, Threa~
## 2 N0000145814 C0059514 ERGONOVINE MALE~ N000000~ C0000821 Abortion, Threa~
## 3 N0000023156 C1572765 WARFARIN SODIUM~ N000000~ C0000821 Abortion, Threa~
## 4 N0000022035 C0244656 FOSPHENYTOIN     N000000~ C0001396 Adams-Stokes Sy~
## 5 N0000022099 C0733758 FOLLITROPIN      N000000~ C0001621 Adrenal Gland D~
## 6 N0000145817 C0012258 DIGITOXIN        N000000~ C0002726 Amyloidosis [Di~
```

If SPARQL endpoints and medical ontologies are quite dispersed over the Web, some initiatives have tried to gather similar knowledge from different sources from the Linked Data. Hence, the Drug Indication Database (DID) have pooled twelve sources of knowledge about drug indications (Sharp 2017). Similarly, the Drug Interaction Knowledge Base (DIKB) have collected fourteen sources of knowledge about potential drug interactions (Ayvaz et al. 2015).

DID and DIKB

Curated versions of DID and DIKB are available in *queryMed* as build-in datasets.

```
data(DIKB)
data(DID)
```

We have now simple knowledge (e.g. definitions, synonyms, comments) as well as complex knowledge to annotate health data. If the simple knowledge is easy to merge with a health database of diseases or drugs,

complex knowledge such as contraindications, interactions or indications, needs a more complex function to search for semantic relations (here specifically pairs of codes) in a database.

find_relations() function aims to perform this kind of mining. And with the appropriate knowledge, it can help to answer the following questions :

- Do patients have drug-disease contraindications ?
- Do patients have drug interaction ?
- Do patients have drug indicated for their disease or health status ?

Let us answer to these questions on the test databases present in *queryMed* : *drug_set* and *disease_set*. Similarly to *drug_set*, *disease_set* is a test dataframe that contains diseases codes for patients, codified according to the International Classification of Diseases - 10th revision (ICD10), and mapped to CUI and NDF-RT.

```
data(disease_set)
head(disease_set)
```

```
##   patient ICD10      cui      NDF-RT
## 1       1 I73.9 C0021775 N0000001694
## 2       1 I73.9 C0085096 N0000003422
## 3       1 I73.9 C0085617        <NA>
## 4       2 I74.4 C0340579        <NA>
## 5       2 I74.4 C0564750        <NA>
## 6       3 I74.4 C0340579        <NA>
```

Do patients have drug-disease contraindications ?

NDF-RT with *find_relations()* can help answer this question :

```
contraindications <- find_relations(data.x=drug_set,
                                    data.y=disease_set,
                                    data_indices = "patient",
                                    data_elements.x = "NDF-RT",
                                    data_elements.y = "NDF-RT",
                                    target=NDFRT_CI,
                                    target_elements = c("ndf_drug","ndf_diag"),
                                    progress="none")

nb_contraindications <- sum(contraindications != "No known relations")
```

We identified 1 patient having at least one drug-disease contraindication, according to NDF-RT.

```
contraindications[contraindications != "No known relations"][1]
```

```
## $`658`
## # A tibble: 1 x 6
##   ndf_drug    cui_drug label_drug ndf_diag    cui_diag label_diag
##   <chr>       <chr>    <chr>      <chr>       <chr>    <chr>
## 1 N0000020412 C0014710 ERGOTAMINE N0000003422 C0085096 Peripheral Vascula~
```

Do patients have drug interaction ?

DIKB can help answer this question :

```
interactions <- find_relations(data.x=drug_set,
                               data_indices = "patient",
                               data_elements.x = "ATC",
                               target=DIKB,
                               target_elements = c("atc1","atc2"),
                               progress="none")

nb_interact<- sum(interactions != "No known relations")
```

We identified 585 patients who have at least one drug interaction, according to DIKB. Here is an example :

```
interactions[interactions != "No known relations"][1]
```

```
## $`1`
##      drug2   drug1      object precipitant contraindication ddiPkMechanism
## 6  DB01118 DB05039 INDACATEROL  AMIODARONE            FALSE           <NA>
## 9  DB01118 DB00758 CLOPIDOGREL  AMIODARONE            FALSE           <NA>
## 10 DB05039 DB01118  AMIODARONE INDACATEROL            FALSE           <NA>
##    effectConcept label precaution severity  uri     source
## 6           <NA>  <NA>      FALSE     <NA> <NA>   Drugbank
## 9           <NA>  <NA>      FALSE     <NA> <NA> NLM-Corpus
## 10          <NA>  <NA>      FALSE     <NA> <NA>   Drugbank
##       evidenceStatement    atc1    atc2
## 6                  <NA> R03AC18 C01BD01
## 9  Specific_Interaction B01AC04 C01BD01
## 10                 <NA> C01BD01 R03AC18
```

Do patients have drug indicated for their disease or health status ?

DID can help answer this question :

```
indications <- find_relations(data.x=drug_set, data.y=disease_set,
                              data_indices = "patient",
                              data_elements.x= "ATC",
                              data_elements.y ="ICD10",
                              target=DID,
                              target_elements=c("atc","icd10"),
                              progress="none")

nb_indication <- sum(indications != "No known relations")
```

We identified 93 patients having at least one relation of indication between their drugs and their diseases, according to DID.

References

Ayvaz, Serkan, John Horn, Oktie Hassanzadeh, Qian Zhu, Johann Stan, Nicholas P. Tatonetti, Santiago Vilar, et al. 2015. "Toward a Complete Dataset of Drug–drug Interaction Information from Publicly Available Sources." *Journal of Biomedical Informatics* 55 (June): 206–17. doi:10.1016/j.jbi.2015.04.006.

Callahan, Alison, José Cruz-Toledo, Peter Ansell, and Michel Dumontier. 2013. "Bio2RDF Release 2: Improved Coverage, Interoperability and Provenance of Life Science Linked Data." In *The Semantic Web:*

Semantics and Big Data, 200–212. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg. doi:10.1007/978-3-642-38288-8_14.

Ferreira, João D, Daniela Paolotti, Francisco M Couto, and Mário J Silva. 2013. "On the Usefulness of Ontologies in Epidemiology Research and Practice." *Journal of Epidemiology and Community Health* 67 (5): 385–88. doi:10.1136/jech-2012-201142.

Lehmann, Jens, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, et al. 2015. "DBpedia - A Large-Scale, Multilingual Knowledge Base Extracted from Wikipedia." *Undefined*. /paper/DBpedia-A-large-scale%2C-multilingual-knowledge-base-Lehmann-Isele/4fa0d9c4c3d17458085ee255b7a4b7c325d59e32.

Ong, Edison, Zuoshuang Xiang, Bin Zhao, Yue Liu, Yu Lin, Jie Zheng, Chris Mungall, Mélanie Courtot, Alan Ruttenberg, and Yongqun He. 2017. "Ontobee: A Linked Ontology Data Server to Support Ontology Term Dereferencing, Linkage, Query and Integration." *Nucleic Acids Research* 45 (D1): D347–D352. doi:10.1093/nar/gkw918.

Pathak, Jyotishman, Richard C. Kiefer, and Christopher G. Chute. 2013. "Using Linked Data for Mining Drug-Drug Interactions in Electronic Health Records." *Studies in Health Technology and Informatics* 192: 682–86.

Salvadores, Manuel, Paul R. Alexander, Mark A. Musen, and Natalya F. Noy. 2013. "BioPortal as a Dataset of Linked Biomedical Ontologies and Terminologies in RDF." *Semantic Web* 4 (3): 277–84.

Sharp, Mark E. 2017. "Toward a Comprehensive Drug Ontology: Extraction of Drug-Indication Relations from Diverse Information Sources." *Journal of Biomedical Semantics* 8 (1). doi:10.1186/s13326-016-0110-0.

Whetzel, Patricia L., Natalya F. Noy, Nigam H. Shah, Paul R. Alexander, Csongor Nyulas, Tania Tudorache, and Mark A. Musen. 2011. "BioPortal: Enhanced Functionality via New Web Services from the National Center for Biomedical Ontology to Access and Use Ontologies in Software Applications." *Nucleic Acids Research* 39 (Web Server issue): W541–545. doi:10.1093/nar/gkr469.

Note d'application

Soumise à la revue Bioinformatics[3], cette note d'application sous la forme d'un article de deux pages intitulé « queryMed: Semantic Web functions for linking pharmacological and medical knowledge to data », est une description du package queryMed. Les principales fonctions du package y sont présentées, ainsi qu'une application type sur données réelles.

3. Le site de la revue Bioinformatics : https://academic.oup.com/bioinformatics

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Databases and ontologies

queryMed: Semantic Web functions for linking pharmacological and medical knowledge to data

Y. Rivault [1,2], O. Dameron [1]* and N. Le Meur [2]*

[1]Univ Rennes, CNRS, Inria, IRISA - UMR 6074, F-35000 Rennes, France.
[2]Univ Rennes, EHESP, REPERES (Pharmacology and health services research) - EA 7449, F-35000 Rennes, France.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: In public health research and more precisely in the reuse of electronic health data, selecting patients, identifying specific events and interpreting results typically requires biomedical knowledge. The queryMed R package aims to facilitate the integration of medical and pharmacological knowledge from Linked Data into the R statistical environment. We show how it allowed us to identify all the drugs prescribed for chronical limb ischemia (CLI) and also to detect one contraindicated prescription for one patient by linking a medical database of 1003 CLI patients to ontologies.
Availability: queryMed is readily usable for medical data mappings and enrichment. Sources, R vignettes and test data are available on GitHub (https://github.com/yannrivault/queryMed) and are archived on Zenodo (https://doi.org/10.5281/zenodo.1323481).
Contact: nolwenn.lemeur@ehesp.fr, olivier.dameron@univ-rennes1.fr

1 Introduction

Medico-administrative databases are useful for public health research, especially in pharmaco-epidemiology. Patient health status and care prescriptions enable identifying severe drug interactions (Nobili et al., 2009) or adverse drug reactions (Durrieu et al., 2014), and studying their determinant on a large scale population. Medical knowledge is needed for each exploration step: before the analysis to select patients, data, and to identify specific events, or after the analysis to interpret the results. The most relevant data in such patient databases are heterogeneous and typically codified according to various medical nomenclatures. This data codification is instrumental for semantic interoperability and knowledge-based data exploration (Pathak et al., 2013). However, exploiting it typically required ad-hoc processing and has long remained non-trivial.

Semantic Web technologies provide the underlying framework for interoperability throughout representation, integration and investigation methods. Based on these technologies, Linked Data promote the publication of structured and interlinked knowledge bases. They can be queried on remote servers, with SPARQL (a Semantic Web technology), or alternatively on REST APIs. For example, BioPortal, one of the principal biomedical ontology repositories, provides methods and tools to access biomedical knowledge through SPARQL endpoints and REST APIs (Whetzel et al., 2011). However, the growing number of knowledge bases, the heterogeneity of their schema representation, and the lack of their conceptual description, make the reuse of knowledge bases –and especially SPARQL query design– a challenge (Jain et al., 2010). Some important works contributed to reduce these bottlenecks. In particular, in medical domain, the Concept Unique Identifier (CUI) from the Unified Medical Language System has been widely used in ontologies to provide mappings between medical terms from different nomenclatures. In addition, the Drug Indication Database (DID) (Sharp, 2017) and the Drug Interaction Knowledge Base (DIKB) (Ayvaz et al., 2015) have gathered knowledge sources of drug indications and interactions from more than a dozen of Linked Data sources. Despite these works, knowledge integration in a patient database is still a laborious process for a non-expert. Ferreira et al. (2012) highlighted the need of tools that facilitate this process, and thus the dissemination of knowledge bases reuse and Semantic Web approaches in epidemiology.

Because data exploration is strongly linked with data analysis, several tools (Van Hage et al., 2013; Willighagen, 2014; Kurbatova et al., 2015) have made available the Linked Data knowledge bases via the Semantic Web technologies in the R statistical environment (R Core Team, 2017). However, there is still a technical barrier to overcome for helping the users to deal with SPARQL queries over medical and pharmacological

ontologies. We present queryMed, an R library of functions that aims to facilitate the access and linkage of medical and pharmacological knowledge to patient databases and thereby knowledge-based exploration.

2 Methods

The main low-level function, `sparql(query, endpoint url)` sends queries to SPARQL endpoints. Because this function requires an expertise in SPARQL query design, biomedical ontologies and appropriate SPARQL endpoints, higher level functions hide predefined SPARQL queries. queryMed offers embedded SPARQL queries suitable for Bio2rdf(Callahan et al., 2013), DBpedia (Lehmann et al., 2015) and Ontobee(Ong et al., 2017) servers. They focus on drugs and diseases and retrieve definitions, abstracts, labels, comments, synonyms, mappings between nomenclatures or particular annotations like contraindications between drugs and diagnosis.

Alternatively, `search(terms, ontologies)` uses the search REST APIs from BioPortal and SIFR BioPortal (Jonquet et al., 2016). This function also requires an expertise in BioPortal repository and its ontologies but is the cornerstone for user-friendly CUI based mapping function. Moreover, `mapping_cui(codes,source,target)` enables nomenclatures linkage, and so expand the potential sources of useful knowledge (Figure 1). Additionally, queryMed provides curated versions of DID and DIKB. Using the retrieved knowledge, `find_relations(...)` function searches for complex semantic relations that involve two medical terms, like drug indications, drug interactions or drug-disease contraindications.

3 Application

The guidelines on follow-up after angioplasty recommend that patients with critical limb ischemia (CLI) should continue a CLI treatment (at least one drug). Drug prescriptions 15 days prior surgery to 31 days after surgery were extracted from the French national system of health data (Medico-administrative data). Using queryMed function calls and indication drug knowledge from DID, we identified 72 CLI patients out 1003 operated in 2015 that had potentially no prescription listed as indicated for CLI after angioplasty. These patients may either had a flawed care trajectory, or had a drug delivery in hospital, or DID might not be exhaustive for CLI drugs.

According to NDF-RT ontology, 91 drugs are contraindicated with CLI. In pharmaco-vigilance context, we searched for patients with at least one prescription contraindicated with CLI. Using a combination of function calls from queryMed, only 1 patient had a prescription for a vasoconstrictor, highly contraindicated with angioplasty (Figure 1 and queryMed vignette for details).

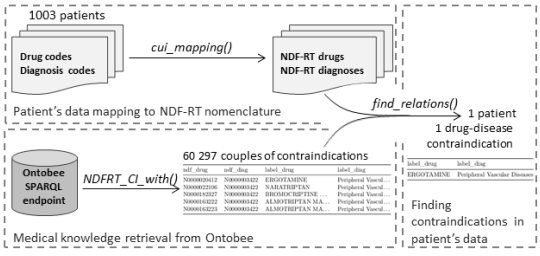

Figure 1. queryMed for a pharmaco-vigilance application on patient's data.

4 Conclusion and discussion

queryMed provides R functions to facilitate linking external knowledge to health data in order to support knowledge-based data analysis. In the context of CLI disease, it permitted answering healthcare recommendations follow up questions and pharmaco-vigilance questions. One limitation of queryMed is that it strongly depends on the quality and sustainability of the medical knowledge published on the Linked Data. Secondly, if the package hides technical complexities such as SPARQL queries or REST APIs, for some function it may lack of flexibility in settings. Future development will focus on increasing flexibility setting, which could also help overcome the evolutions from the Linked Data.

Acknowledgements and Fundings

This work was funded by the French Health Products Agency. The study protocol was approved by the National Institute of Health Data review board (#201) and by the French data protection authority (#1968571). The authors declare no conflict of interest.

References

Serkan Ayvaz et al. Toward a complete dataset of drug-drug interaction information from publicly available sources. *Journal of Biomedical Informatics*, 55:206–217, June 2015. ISSN 15320464. doi: 10.1016/j.jbi.2015.04.006.

Alison Callahan et al. Bio2rdf Release 2: Improved Coverage, Interoperability and Provenance of Life Science Linked Data. In *The Semantic Web: Semantics and Big Data*, Lecture Notes in Computer Science, pages 200–212. Springer, Berlin, Heidelberg, May 2013. ISBN 978-3-642-38287-1 978-3-642-38288-8. doi: 10.1007/978-3-642-38288-8_14.

G. Durrieu et al. Use of administrative hospital database to identify adverse drug reactions in a Pediatric University Hospital. *European Journal of Clinical Pharmacology*, 70(12):1519–1526, December 2014. ISSN 1432-1041. doi: 10.1007/s00228-014-1763-1.

João D Ferreira et al. On the usefulness of ontologies in epidemiology research and practice. *Journal of Epidemiology and Community Health*, 67(5):385–388, nov 2012. doi: 10.1136/jech-2012-201142.

Prateek Jain et al. Linked Data is Merely More Data. *Papers from the AAAI Spring Symposium: Linked Data Meets Artificial Intelligence*, pages 82–86, January 2010.

Clement Jonquet et al. SIFR BioPortal : Un portail ouvert et générique d'ontologies et de terminologies biomédicales françaises au service de l'annotation sémantique. In *JFIM: Journées Francophones d'Informatique Médicale*, e-health pour tous, Genève, Switzerland, June 2016.

Natalja Kurbatova et al. *ontoCAT: Ontology traversal and search*, 2015. R package version 1.29.0.

Jens Lehmann et al. Dbpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6:167–195, 2015. doi: 10.3233/SW-140134.

A. Nobili et al. Potentially severe drug interactions in elderly outpatients: results of an observational study of an administrative prescription database. *Journal of Clinical Pharmacy and Therapeutics*, 34(4):377–386, August 2009. ISSN 1365-2710. doi: 10.1111/j.1365-2710.2009.01021.x.

Edison Ong et al. Ontobee: A linked ontology data server to support ontology term dereferencing, linkage, query and integration. *Nucleic Acids Research*, 45(D1): D347–D352, January 2017. ISSN 1362-4962. doi: 10.1093/nar/gkw918.

Jyotishman Pathak et al. Using linked data for mining drug-drug interactions in electronic health records. *Studies in Health Technology and Informatics*, 192: 682–686, 2013. ISSN 0926-9630. doi: 10.3233/978-1-61499-289-9-682.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

Mark E Sharp. Toward a comprehensive drug ontology: extraction of drug-indication relations from diverse information sources. *Journal of Biomedical Semantics*, 8 (1), December 2017. ISSN 2041-1480. doi: 10.1186/s13326-016-0110-0.

Willem Robert Van Hage et al. *SPARQL: SPARQL client*, 2013. R package version 1.16.

Patricia L. Whetzel et al. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Research*, 39(Web Server issue):W541–545, July 2011. ISSN 1362-4962. doi: 10.1093/nar/gkr469.

Egon Willighagen. Accessing biological data in R with semantic web technologies. Technical Report e185v3, PeerJ Inc., March 2014.

Analyse de trajectoires de soins

En plus de la présentation de la section 5.1 lors de la 16ème conférence d'intelligence artificielle en médecine, un article (Rivault et al., 2017b) associé de cinq page a été publié dans les *Lectures Notes in Computer Science* (LNCS) [4]

4. LNCS sur le site de Springer : https://www.springer.com/gp/computer-science/lncs

A Similarity Measure Based on Care Trajectories as Sequences of Sets

Yann Rivault[1,3,4], Nolwenn Le Meur[1,4], Olivier Dameron[2,3,4]

[1] EHESP Rennes, Sorbonne Paris Cité, EA 7449 REPERES, Recherche en Pharmaco-Epidémiologie et Recours aux Soins, France
[2] Université de Rennes 1, 35000 Rennes, France
[3] IRISA équipe Dyliss, 35042 Rennes
[4] PEPS, Pharmacoepidemiology for health products safety

{Yann.Rivault, Nolwenn.LeMeur}@ehesp.fr
Olivier.Dameron@univ-rennes1.fr

Abstract. Comparing care trajectories helps improve health services. Medico-administrative databases are useful for automatically reconstructing the patients' history of care. Care trajectories can be compared by determining their overlapping parts. This comparison relies on both semantically-rich representation formalism for care trajectories and an adequate similarity measure. The longest common subsequence (LCS) approach could have been appropriate if representing complex care trajectories as simple sequences was expressive enough. Furthermore, by failing to take into account similarities between different but semantically close medical events, the LCS overestimates differences. We propose a generalization of the LCS to a more expressive representation of care trajectories as sequences of sets. A set represents a medical episode composed by one or several medical events, such as diagnosis, drug prescription or medical procedures. Moreover, we propose to take events' semantic similarity into account for comparing medical episodes. To assess our approach, we applied the method on a care trajectories' sample from patients who underwent a surgical act among three kinds of acts. The formalism reduced calculation time, and introducing semantic similarity made the three groups more homogeneous.

Keywords: Care trajectories · LCS-based similarity · Semantic similarity

1 Introduction

Medico-administrative databases are valuable data source for health research notably because of their large population coverage, as well as of their longitudinal properties [1]. In France, the French national health insurance inter-regime information system (SNIIRAM) records the reimbursements of health care covered by the main insurance funds for workers. These data include ambulatory care data and hospital discharge summaries issued from the French hospital discharge information systems (PMSI). Although their primary goals are essentially financial and managerial, these

databases make possible to explore and analyse patients' care trajectories [2]. Understanding and analysing these data are crucial for efficient healthcare planning and fair allocation of health care resources [3]. Moreover, care trajectory analysis can be an asset for epidemiology studies by providing statistical indicators to understand and explain care seeking behaviours [4]. Care trajectories' comparison, which is part of their analysis, relies (i) on their representation and (ii) on an adequate comparison method. If we consider medico-administrative data for composing care trajectories, such as diagnoses, medical procedures or drugs prescriptions, an intuitive way of representing it is to write it down as a sequence. However, such formalism could be too simplistic considering the complexity of a care trajectory. The main complexity could be that the temporal order between events is not always known or even meaningful, especially when they occur simultaneously or in a really short time range. The second challenge is to handle the complexity of the large alphabet of trajectories' components. They are medical concepts that belong to detailed taxonomies, and taking into account semantic similarities [5] between these codes could render the method more robust to small variations [6]. The goal of this article is to introduce a representation of care trajectories that better accounts for administrative data complexity and an associated similarity measure for comparing them.

2 Materials and Methods

2.1 Representing trajectories as sequences of sets

To take into account the uncertainty or simultaneity ignored with the simple sequence formalism, we proposed to group the events as unordered sets of events. Trajectories can then be seen as sequences of sets composed of simultaneous or related events. Similarity measures between sequences [7] could then be generalized to this formalism. To determine the overlapping part between two trajectories, we generalized the principle of the longest common subsequence (LCS) to this formalism.

2.2 Comparing Sequences of Sets

Longest Common Subsequence for Sequences of Sets

Definition 1: sequence of sets

A sequence of sets is a non-empty sequence composed of sets of elements.

Definition 2: size of a sequence of sets

Let $X = (x_1, x_2, \dots, x_m)$ be a sequence of sets. The size of a sequence of set is:

$$|X| = \sum_{i=1}^{m} |x_i| \quad (1)$$

Definition 3: subsequence of a sequence of sets

Given two sequences of sets $X = (x_1, x_2, \dots, x_m)$ and $Y = (y_1, y_2, \dots, y_n)$, with $m \leq n$, X is a subsequence of Y if it exists the indexes $1 \leq j_1 < j_2 < \cdots < j_m \leq n$ such as $x_i \subseteq y_{j_i}$ is true for all $i = 1,2, \dots, m$.

Definition 4: longest common subsequence of two sequences of sets
Given two sequences of sets $X = (x_1, x_2, \dots, x_m)$ and $Y = (y_1, y_2, \dots, y_n)$, Z is a LCS of X and Y if $|Z| \geq |Z'|$, for all other common subsequence Z' of X and Y.

Trajectories Structural Similarity. To get a similarity measure between sequences, it is intuitive and popular to normalize the size of the LCS by maximal size of the compared sequences.

Definition 5: similarity between trajectories
We define a similarity measure between X and Y sequences of sets as:

$$sim(X,Y) = \frac{|LCS(X,Y)|}{max(|X|,|Y|)} \quad (2)$$

Considering Semantic Similarity between Events. In order to take into account similarities between elements of two trajectories, we proposed a modification of the $|LCS(\mathrm{X},\mathrm{Y})|$ calculation. Instead of using intersection between sets to determine the common part of two trajectories, we used a similarity measure between sets, which requires a similarity between elements of these sets. These elements are issued from taxonomies, and several semantic similarities based on their hierarchical structures are conceivable [5]. This similarity allows us to compute similarity measure between two sets of concepts.

Definition 6: similarity measure between two sets of concepts
Given two sets of concepts $X = (x_1, x_2, \dots, x_m)$ and $Y = (y_1, y_2, \dots, y_n)$, we note C the set of all matchings between elements of X and Y, and $sem(,)$ a semantic similarity between concepts. The similarity measure between X and Y is defined as:

$$sim(X,Y) = \max_{c \in C} \sum_{(x,y) \in c} sem(x,y) \quad (3)$$

Due to this modification, the similarity measure is not anymore based on the longest common subsequence but on what we could call the longest similar subsequence. As for solving many algorithmic text problems, such as sequence alignment, both similarities can be computed using a dynamic programming algorithm [9].

2.3 Experimentations

We performed a retrospective analysis using the permanent sample of the SNIIRAM database (EGB). The EGB is a representative cross-sectional sample of the population covered by National Health Insurance. First, we extracted the hospital stay and reimbursed drug prescription information of the patients who underwent an ambulatory care surgery in 2012. All available medical codes, i.e. principal diagnoses, related diagnoses, associated diagnoses, clinical acts and drugs delivered in pharmacies were extracted from three months before to three months after the hospital stay to reconstruct the care trajectories. Next, to experiment the use of our method in performing cluster analysis, we selected three sub-groups of ambulatory surgeries, namely angioplasties, eye surgeries and breast surgeries, which constitute a sample of 287 patients. Elements of the care trajectories were drugs, diagnoses and medical acts, represented respectively by the Anatomical Therapeutic Chemical Classification System (ATC),

the International Statistical Classification of Diseases and Related Health Problems – 10th revision (ICD-10), and the Common Classification of Medical Procedures (CCAM). To compute semantic similarities between elements from a same classification, we used the Wu and Palmer's similarity [10] which is based on their hierarchical structure. We then computed equation (2) between each pair of the 287 trajectories, with and then without taking into account the semantic similarities. We performed a cluster analysis using the R software (version 3.1.1), with an ascending hierarchical classification and a Ward linkage. Three classes were identifed based on the highest drop of inertia between classes. Then we focused on intra-class and inter-class similarity. Because we worked on three sub-groups of patients, we had an *a priori* knowledge of the class a patient belongs to. We computed the ratio between the sum of patient's similarities with its own group and the sum of patient's similarities with the other groups, for both kind of similarities and for each patient.

3 Results

Before introducing semantic similarities in the method, the running time was twice faster when considering sequences of sets than sequences of atomic elements. It was no longer the case with their introduction, because it is more complicated to compute semantic similarities between sets of concepts than between atomic concepts.

Before evaluating the relevance of using semantic similarity, we ensured that a cluster analysis based on these similarities led to three distinct clusters associated to the three initial sub-groups. Only three patients were not classified in the correct cluster. Further analysis revealed that all three shared frequent comorbidities with patients from the other groups.

Because including semantic similarity to the care trajectory similarity measure could only increase the final similarity values, we focused on the ratios between intra-class and inter-class similarity. A statistical comparison has shown that they were significantly higher with the semantic similarities' introduction (Wilcoxon signed-rank test, p=0.002). Overall, with this enrichment, the similarity of a patient with the patients of its own group has thus more increased than its similarity with the patients of the other groups, which is desirable in a cluster analysis.

4 Discussion and Conclusion

Thanks to its expressivity, the formalism of sequence of sets is appropriate to represent care trajectories based on medico-administrative data, e.g. clinical acts, diagnosis and drug codes. We have proposed a method to compare two care trajectories written as sequences of sets, which relies on a generalization of the LCS problem, in order to identify the homologous parts of two patients' care trajectories. And to make this method less strict, more robust to small variations, we tried to take into account the possible similarity between the care trajectories' components.

Our next objective will be to study the potential of the method in a clinical context, specifically the hospital stays for an angioplasty, to see if the method could be useful

for predicting post hospital stay outcomes (e.g. rehospitalisation and adverse events), explaining disease conditions severity or a mode of taking in charge the patients (e.g. ambulatory or inpatient care), and discovering trends in the care consumptions.

We envision enriching the method by taking into account other kinds of similarity between events, such as delay between events or events' durations. It is also our objective to apply this method to patient and guideline comparison, to know the homologous part between a care trajectory and a guideline.

Funding

Doctoral fellowship funded by PEPS Research consortium, supported by Agence Nationale de Sécurité des Médicaments et produits de santé (ANSM).

References

1. Tuppin, P., de Roquefeuil, L., Weill, A., Ricordeau, P., Merlière, Y.: French national health insurance information system and the permanent beneficiaries sample. Rev Epidemiol Sante Publique. 58, 286–290 (2010).
2. Moulis, G., Lapeyre-Mestre, M., Palmaro, A., Pugnet, G., Montastruc, J.-L., Sailler, L.: French health insurance databases: What interest for medical research? Rev Med Interne. 36, 411–417 (2015).
3. Jay, N., Nuemi, G., Gadreau, M., Quantin, C.: A data mining approach for grouping and analyzing trajectories of care using claim data: the example of breast cancer. BMC Med Inform Decis Mak. 13, 130 (2013).
4. Le Meur, N., Gao, F., Bayat, S.: Mining care trajectories using health administrative information systems: the use of state sequence analysis to assess disparities in prenatal care consumption. BMC Health Serv Res. 15, 200 (2015).
5. Pesquita, C., Faria, D., Falcão, A.O., Lord, P., Couto, F.M.: Semantic Similarity in Biomedical Ontologies. PLoS Computational Biology. 5, e1000443 (2009).
6. Girardi, D., Wartner, S., Halmerbauer, G., Ehrenmüller, M., Kosorus, H., Dreiseitl, S.: Using concept hierarchies to improve calculation of patient similarity. Journal of Biomedical Informatics. 63, 66–73 (2016).
7. Studer, M., Ritschard, G., Ritschard, G., Ritschard, G.: A comparative review of sequence dissimilarity measures. LIVES Working Papers. 2014, 1–47 (2014).
8. Hirschberg, D.S.: A linear space algorithm for computing maximal common subsequences. Communications of the ACM. 18, 341–343 (1975).
9. Bellman, R.: The theory of dynamic programming. Bull. Amer. Math. Soc. 60, 503–515 (1954).
10. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. Presented at the Proceedings of the 32nd annual meeting on Association for Computational Linguistics June 27 (1994).

Bibliographie

Adeyemi, S., Demir, E., and Chaussalet, T. Towards an evidence-based decision making healthcare system management: Modelling patient pathways to improve clinical outcomes. *Decision Support Systems*, 55(1):117–125, April 2013. ISSN 0167-9236. doi: 10.1016/j.dss.2012.12.039. (Cité en page 14.)

Agrawal, R. and Srikant, R. Fast Algorithms for Mining Association Rules in Large Databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc. ISBN 978-1-55860-153-6. (Cité en pages 88, 90 et 100.)

Agrawal, R. and Srikant, R. Mining Sequential Patterns. In *Proceedings of the Eleventh International Conference on Data Engineering*, ICDE '95, pages 3–14, Washington, DC, USA, 1995. IEEE Computer Society. ISBN 978-0-8186-6910-1. (Cité en page 30.)

Agrawal, R., Imieliński, T., and Swami, A. Mining Association Rules Between Sets of Items in Large Databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, SIGMOD '93, pages 207–216, New York, NY, USA, 1993. ACM. ISBN 978-0-89791-592-2. doi: 10.1145/170035.170072. (Cité en pages 30 et 31.)

Ainsworth, J. and Buchan, I. COCPIT: a tool for integrated care pathway variance analysis. *Studies in Health Technology and Informatics*, 180:995–999, 2012. ISSN 0926-9630. (Cité en page 27.)

Anjum, A., Bloodsworth, P., Branson, A., Hauer, T., McClatchey, R., Munir, K., Rogulin, D., and Shamdasani, J. The Requirements for Ontologies in Medical Data Integration: A Case Study, 2007. (Cité en page 2.)

Ayvaz, S., Horn, J., Hassanzadeh, O., Zhu, Q., Stan, J., Tatonetti, N. P., Vilar, S., Brochhausen, M., Samwald, M., Rastegar-Mojarad, M., Dumontier, M., and Boyce, R. D. Toward a complete dataset of drug-drug interaction information from publicly available sources. *Journal of Biomedical Informatics*, 55:206–217, June 2015. ISSN 1532-0480. doi: 10.1016/j.jbi.2015.04.006. (Cité en page 65.)

Bard, J. B. L. and Rhee, S. Y. Ontologies in biology: design, applications and future challenges. *Nature Reviews. Genetics*, 5(3):213–222, March 2004. ISSN 1471-0056. doi: 10.1038/nrg1295. (Cité en page 15.)

Bayardo, R. J., Agrawal, R., and Gunopulos, D. Constraint-Based Rule Mining in Large, Dense Databases. *Data Mining and Knowledge Discovery*, 4(2):217–240, July 2000. ISSN 1573-756X. doi: 10.1023/A:1009895914772. (Cité en pages 89, 91 et 96.)

Bellman, R. The theory of dynamic programming. *Bulletin of the American Mathematical Society*, 60(6):503–515, 1954. ISSN 0002-9904, 1936-881X. doi: 10.1090/S0002-9904-1954-09848-8. (Cité en page 77.)

Bettembourg, C., Dameron, O., Bretaudeau, A., and Legeai, F. AskOmics : Intégration et interrogation de réseaux de régulation génomique et post-génomique. page 7, June 2015. (Cité en page 65.)

Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–270, January 2004. ISSN 1362-4962. doi: 10.1093/nar/gkh061. (Cité en page 23.)

Callahan, A., Cruz-Toledo, J., Ansell, P., and Dumontier, M. Bio2rdf Release 2: Improved Coverage, Interoperability and Provenance of Life Science Linked Data. In *The Semantic Web: Semantics and Big Data*, Lecture Notes in Computer Science, pages 200–212. Springer, Berlin, Heidelberg, May 2013. ISBN 978-3-642-38287-1 978-3-642-38288-8. doi: 10.1007/978-3-642-38288-8_14. (Cité en pages 23 et 66.)

Campbell, H., Hotchkiss, R., Bradshaw, N., and Porteous, M. Integrated care pathways. *BMJ*, 316(7125):133–137, January 1998. ISSN 0959-8138, 1468-5833. doi: 10.1136/bmj.316.7125.133. (Cité en page 14.)

Carbon, S., Dietze, H., Lewis, S. E., Mungall, C. J., Munoz-Torres, M. C., Basu, S., Chisholm, R. L., Dodson, R. J., Fey, P., Thomas, P. D., Mi, H., Muruganujan, A., Huang, X., Poudel, S., Hu, J. C., Aleksander, S. A., McIntosh, B. K., Renfro, D. P., Siegele, D. A., Antonazzo, G., Attrill, H., Brown, N. H., Marygold, S. J., Mc-Quilton, P., Ponting, L., Millburn, G. H., Rey, A. J., Stefancsik, R., Tweedie, S., Falls, K., Schroeder, A. J., Courtot, M., Osumi-Sutherland, D., Parkinson, H., Roncaglia, P., Lovering, R. C., Foulger, R. E., Huntley, R. P., Denny, P., Campbell, N. H., Kramarz, B., Patel, S., Buxton, J. L., Umrao, Z., Deng, A. T., Alrohaif, H., Mitchell, K., Ratnaraj, F., Omer, W., Rodríguez-López, M., and Consortium, T. G. O. Expansion of the gene ontology knowledgebase and resources: The gene ontology consortium. *Nucleic acids research*, 45(D1):D331–D338, January 2017. ISSN 0305-1048. doi: 10.1093/nar/gkw1108. (Cité en page 107.)

Century, Institute of Medicine (US) Committee on Assuring the Health of the Public in the 21st. *Understanding Population Health and Its Determinants*. National Academies Press (US), 2002. (Cité en page 6.)

Chandanan, A. K. and Shukla, M. K. Removal of Duplicate Rules for Association Rule Mining from Multilevel Dataset. *Procedia Computer Science*, 45:143–149, January 2015. ISSN 1877-0509. doi: 10.1016/j.procs.2015.03.106. (Cité en page 89.)

Chen, P., Verma, R. M., Meininger, J. C., and Chan, W. Semantic Analysis of Association Rules. In *FLAIRS Conference*, 2008. (Cité en pages 35 et 99.)

Collen, M. F. The Development of Medical Databases. In Collen, M. F., editor, *Computer Medical Databases: The First Six Decades (1950–2010)*, Health Informatics, pages 33–55. Springer London, London, 2012. ISBN 978-0-85729-962-8. doi: 10.1007/978-0-85729-962-8_2. (Cité en page 47.)

Concaro, S., Sacchi, L., Cerra, C., Fratino, P., and Bellazzi, R. Mining Healthcare Data with Temporal Association Rules: Improvements and Assessment for a Practical Use. In Combi, C., Shahar, Y., and Abu-Hanna, A., editors, *Artificial Intelligence in Medicine*, Lecture Notes in Computer Science, pages 16–25. Springer Berlin Heidelberg, 2009. ISBN 978-3-642-02976-9. (Cité en page 31.)

Daien, V., Korobelnik, J.-F., Delcourt, C., Cougnard-Gregoire, A., Delyfer, M. N., Bron, A. M., Carrière, I., Villain, M., Daures, J. P., Lacombe, S., Mariet, A. S., Quantin, C., and Creuzot-Garcher, C. French Medical-Administrative Database for Epidemiology and Safety in Ophthalmology (EPISAFE): The EPISAFE Collaboration Program in Cataract Surgery. *Ophthalmic Research*, 58(2):67–73, 2017. ISSN 0030-3747, 1423-0259. doi: 10.1159/000456721. (Cité en page 7.)

Damerau, F. J. A Technique for Computer Detection and Correction of Spelling Errors. *Commun. ACM*, 7(3):171–176, March 1964. ISSN 0001-0782. doi: 10.1145/363958.363994. (Cité en page 28.)

Dauxais, Y., Guyet, T., Gross-Amblard, D., and Happe, A. Discriminant Chronicles Mining. In ten Teije, A., Popow, C., Holmes, J. H., and Sacchi, L., editors, *Artificial Intelligence in Medicine*, Lecture Notes in Computer Science, pages 234–244. Springer International Publishing, 2017. ISBN 978-3-319-59758-4. (Cité en page 33.)

Defossez, G., Rollet, A., Dameron, O., and Ingrand, P. Temporal representation of care trajectories of cancer patients using data from a regional information system: an application in breast cancer. *BMC Medical Informatics and Decision Making*, 14(1):24, April 2014. ISSN 1472-6947. doi: 10.1186/1472-6947-14-24. (Cité en page 14.)

Deneckere, S., Euwema, M., Van Herck, P., Lodewijckx, C., Panella, M., Sermeus, W., and Vanhaecht, K. Care pathways lead to better teamwork: results of a systematic review. *Social Science & Medicine (1982)*, 75(2):264–268, July 2012. ISSN 1873-5347. doi: 10.1016/j.socscimed.2012.02.060. (Cité en page 14.)

Dowle, M. and Srinivasan, A. *data.table: Extension of 'data.frame'*. 2018. (Cité en page 64.)

Egho, E., Jay, N., Raïssi, C., Nuemi, G., Quantin, C., and Napoli, A. An Approach for Mining Care Trajectories for Chronic Diseases. In Peek, N., Marín Morales, R., and Peleg, M., editors, *Artificial Intelligence in Medicine*, Lecture Notes in Computer Science, pages 258–267. Springer Berlin Heidelberg, 2013a. ISBN 978-3-642-38326-7. (Cité en page 29.)

Egho, E., Raïssi, C., Ienco, D., Jay, N., Napoli, A., Poncelet, P., Quantin, C., and Teisseire, M. Healthcare Trajectory Mining by Combining Multidimensional Component and Itemsets. In Appice, A., Ceci, M., Loglisci, C., Manco, G., Masciari, E., and Ras, Z. W., editors, *New Frontiers in Mining Complex Patterns*, Lecture Notes in Computer Science, pages 109–123. Springer Berlin Heidelberg, 2013b. ISBN 978-3-642-37382-4. (Cité en page 35.)

Egho, E., Raïssi, C., Calders, T., Jay, N., and Napoli, A. On measuring similarity for sequences of itemsets. *Data Mining and Knowledge Discovery*, 29(3):732–764, May 2015. ISSN 1573-756X. doi: 10.1007/s10618-014-0362-1. (Cité en page 74.)

Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., and Borsboom, D. qgraph: Network Visualizations of Relationships in Psychometric Data. *Journal of Statistical Software*, 48(4):1–18, 2012. (Cité en page 82.)

Ferreira, J. D., Paolotti, D., Couto, F. M., and Silva, M. J. On the usefulness of ontologies in epidemiology research and practice. *Journal of Epidemiology and Community Health*, 67(5):385–388, May 2013. ISSN 0143-005X. doi: 10.1136/jech-2012-201142. (Cité en pages 3, 25, 26 et 65.)

Fetter, R. B., Shin, Y., Freeman, J. L., Averill, R. F., and Thompson, J. D. Case mix definition by diagnosis-related groups. *Medical Care*, 18(2 Suppl):iii, 1–53, February 1980. ISSN 0025-7079. (Cité en page 9.)

Fourquet, F., Demont, F., Lecuyer, A. I., Rogers, M. A., and Bloc, D. H. PMSI et surveillance des infections nosocomiales : théorie et faisabilité. *Médecine et Maladies Infectieuses*, 33(2):110–113, February 2003. ISSN 0399-077X. doi: 10.1016/S0399-077X(02)00005-7. (Cité en page 13.)

Geng, L. and Hamilton, H. J. Choosing the Right Lens: Finding What is Interesting in Data Mining. In Kacprzyk, J., Guillet, F. J., and Hamilton, H. J., editors, *Quality Measures in Data Mining*, volume 43, pages 3–24. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007. ISBN 978-3-540-44911-9 978-3-540-44918-8. doi: 10.1007/978-3-540-44918-8_1. (Cité en pages 32 et 88.)

Georgescu, I. and Hartmann, F. G. H. Sources of financial pressure and up coding behavior in French public hospitals. *Health Policy (Amsterdam, Netherlands)*, 110 (2-3):156–163, May 2013. ISSN 1872-6054. doi: 10.1016/j.healthpol.2013.02.003. (Cité en page 13.)

Girardi, D., Dirnberger, J., and Giretzlehner, M. An ontology-based clinical data warehouse for scientific research. *Safety in Health*, 1(1):6, July 2015. ISSN 2056-5917. doi: 10.1186/2056-5917-1-6. (Cité en page 25.)

Girardi, D., Wartner, S., Halmerbauer, G., Ehrenmüller, M., Kosorus, H., and Dreiseitl, S. Using concept hierarchies to improve calculation of patient similarity. *Journal of Biomedical Informatics*, 63:66–73, October 2016. ISSN 1532-0464. doi: 10.1016/j.jbi.2016.07.021. (Cité en pages 34 et 79.)

Goldberg, M., Quantin, C., Guégen, A., and Zins, M. Bases de données médico-administratives et épidémiologie : intérêts et limites. May 2008. ISSN 2107-0903. (Cité en page 13.)

Grammatico-Guillon, L., Baron, S., Gaborit, C., Rusch, E., and Astagneau, P. Quality assessment of hospital discharge database for routine surveillance of hip and knee arthroplasty-related infections. *Infection Control and Hospital Epidemiology*, 35(6):646–651, June 2014. ISSN 1559-6834. doi: 10.1086/676423. (Cité en page 13.)

Grosjean, J., Merabti, T., Dahamna, B., Kergourlay, I., Thirion, B., Soualmia, L. F., and Darmoni, S. J. Health multi-terminology portal: a semantic added-value for patient safety. *Studies in Health Technology and Informatics*, 166:129–138, 2011. ISSN 0926-9630. (Cité en page 22.)

Grothendieck, G. *sqldf: Manipulate R Data Frames Using SQL*. 2017. (Cité en page 47.)

Gruber, T. R. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *Int. J. Hum.-Comput. Stud.*, 43(5-6):907–928, December 1995. ISSN 1071-5819. doi: 10.1006/ijhc.1995.1081. (Cité en page 15.)

Guillet, F. and Hamilton, H. J., editors. *Quality Measures in Data Mining*. Studies in Computational Intelligence. Springer-Verlag, Berlin Heidelberg, 2007. ISBN 978-3-540-44911-9. (Cité en page 31.)

Guyet, T., Happe, A., and Dauxais, Y. Declarative Sequential Pattern Mining of Care Pathways. volume 24, pages 1161 – 266, June 2017. doi: 10.1007/978-3-319-59758-4_29. (Cité en page 108.)

Hage, W. R. V. and others. *SPARQL: SPARQL client*. 2013. (Cité en page 66.)

Hahsler, M., Gruen, B., and Hornik, K. arules – A Computational Environment for Mining Association Rules and Frequent Item Sets. *Journal of Statistical Software*, 14(15):1–25, October 2005. ISSN 1548-7660. (Cité en pages 90 et 100.)

Hamming, R. W. Error detecting and error correcting codes. *The Bell System Technical Journal*, 29(2):147–160, April 1950. ISSN 0005-8580. doi: 10.1002/j.1538-7305.1950.tb00463.x. (Cité en page 28.)

Hirschberg, D. S. A linear space algorithm for computing maximal common subsequences. *Communications of the ACM*, 18(6):341–343, June 1975. ISSN 00010782. doi: 10.1145/360825.360861. (Cité en pages 28 et 75.)

Horrocks, I. OWL: A Description Logic Based Ontology Language. In van Beek, P., editor, *Principles and Practice of Constraint Programming - CP 2005*, Lecture Notes in Computer Science, pages 5–8. Springer Berlin Heidelberg, 2005. ISBN 978-3-540-32050-0. (Cité en page 18.)

Hylek, E. M., Heiman, H., Skates, S. J., Sheehan, M. A., and Singer, D. E. Acetaminophen and other risk factors for excessive warfarin anticoagulation. *JAMA*, 279(9):657–662, March 1998. ISSN 0098-7484. (Cité en page 105.)

Ivanović, M. and Budimac, Z. An overview of ontologies and data resources in medical domains. *Expert Systems with Applications*, 41(11):5158–5166, September 2014. ISSN 0957-4174. doi: 10.1016/j.eswa.2014.02.045. (Cité en page 25.)

Jain, P., Hitzler, P., Yeh, P., Verma, K., and Sheth, A. Linked Data is Merely More Data. *Papers from the AAAI Spring Symposium: Linked Data Meets Artificial Intelligence*, pages 82–86, January 2010. (Cité en pages 3, 26 et 65.)

Janke, D., Skubella, A., and Staab, S. Evaluating SPARQL 1.1 Property Path Support. In *BLINK/NLIWoD3@ISWC*, 2017. (Cité en page 111.)

Jaro, M. A. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, 84(406):414–420, 1989. ISSN 0162-1459. doi: 10.2307/2289924. (Cité en page 28.)

Jay, N. and d'Aquin, M. Linked Data and Online Classifications to Organise Mined Patterns in Patient Data. *AMIA Annual Symposium Proceedings*, 2013:681–690, November 2013. ISSN 1942-597X. (Cité en page 74.)

Jay, N., Nuemi, G., Gadreau, M., and Quantin, C. A data mining approach for grouping and analyzing trajectories of care using claim data: the example of breast cancer. *BMC Medical Informatics and Decision Making*, 13(1):130, November 2013. ISSN 1472-6947. doi: 10.1186/1472-6947-13-130. (Cité en pages 7 et 14.)

Jonquet, C., Shah, N., and Musen, M. Un service Web pour l'annotation sémantique de données biomédicales avec des ontologies. In Fieschi, M., Staccini, P., Bouhaddou, O., and Lovis, C., editors, *13èmes Journées Francophones d'Informatique Médicale, JFIM'09*, volume 17, page 12, Nice, France, France, April 2009. (Cité en pages 22 et 26.)

Jonquet, C., Annane, A., Bouarech, K., Emonet, V., and Melzi, S. SIFR BioPortal : Un portail ouvert et générique d'ontologies et de terminologies biomédicales françaises au service de l'annotation sémantique. In *JFIM: Journées Francophones d'Informatique Médicale*, e-health pour tous, Genève, Switzerland, June 2016. (Cité en page 21.)

Kost, R., Littenberg, B., and Chen, E. S. Exploring Generalized Association Rule Mining for Disease Co-Occurrences. *AMIA Annual Symposium Proceedings*, 2012: 1284–1293, November 2012. ISSN 1942-597X. (Cité en page 35.)

Kurbatova, N. and others. ontoCAT: an R package for ontology traversal and search. *Bioinformatics*, 27(17):2468–2470, September 2011. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btr375. (Cité en page 66.)

Le Meur, N., Gao, F., and Bayat, S. Mining care trajectories using health administrative information systems: the use of state sequence analysis to assess disparities in prenatal care consumption. *BMC Health Services Research*, 15(1):200, May 2015. ISSN 1472-6963. doi: 10.1186/s12913-015-0857-5. (Cité en pages 2, 10, 14, 27 et 29.)

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., Kleef, P. v., Auer, S., and Bizer, C. DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia, 2015. (Cité en pages 23 et 66.)

Létinier, L., Cossin, S., Mansiaux, Y., Arnaud, M., Bezin, J., and Pariente, A. Prévalence et description des situations à risque d'interactions médicamenteuses en France à partir des données de l'Assurance maladie. *Revue d'Épidémiologie et de Santé Publique*, 66:S22, March 2018. ISSN 0398-7620. doi: 10.1016/j.respe.2018.01.044. (Cité en page 10.)

Levenshtein, V. I. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707, February 1966. (Cité en page 28.)

Li, X., Mei, J., Liu, H., Yu, Y., Xie, G., Hu, J., and Wang, F. Analysis of care pathway variation patterns in patient records. *Studies in Health Technology and Informatics*, 210:692–696, 2015. ISSN 0926-9630. (Cité en page 27.)

Lim, S., Lee, K., and Kang, J. Drug drug interaction extraction from the literature using a recursive neural network. *PLOS ONE*, 13(1):e0190926, January 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0190926. (Cité en page 107.)

Mahé, I. Principes d'utilisation du traitement anticoagulant chez le sujet âgé en médecine. *Sang Thrombose Vaisseaux*, 16(7):339–345, September 2004. ISSN 0999-7385. (Cité en page 105.)

Manda, P., Ozkan, S., Wang, H., McCarthy, F., and Bridges, S. M. Cross-Ontology Multi-level Association Rule Mining in the Gene Ontology. *PLoS ONE*, 7(10), October 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0047411. (Cité en page 88.)

Maura, G., Billionnet, C., Coste, J., Weill, A., Neumann, A., and Pariente, A. Nonbleeding Adverse Events with the Use of Direct Oral Anticoagulants: A Sequence Symmetry Analysis. *Drug Safety*, 41(9):881–897, 2018. ISSN 0114-5916. doi: 10.1007/s40264-018-0668-9. (Cité en page 10.)

Navarro, G. A Guided Tour to Approximate String Matching. *ACM Comput. Surv.*, 33(1):31–88, March 2001. ISSN 0360-0300. doi: 10.1145/375360.375365. (Cité en pages 28 et 75.)

Ong, E., Xiang, Z., Zhao, B., Liu, Y., Lin, Y., Zheng, J., Mungall, C., Courtot, M., Ruttenberg, A., and He, Y. Ontobee: A linked ontology data server to support ontology term dereferencing, linkage, query and integration. *Nucleic Acids Research*, 45(D1):D347–D352, January 2017. ISSN 1362-4962. doi: 10.1093/nar/gkw918. (Cité en pages 23 et 66.)

Ornetti, P., Ciappuccini, R., Tavernier, C., and Maillefert, J. F. Interaction between paracetamol and oral anticoagulants. *Rheumatology*, 44(12):1584–1585, December 2005. ISSN 1462-0324. doi: 10.1093/rheumatology/kei102. (Cité en page 105.)

Owens, P. L., Barrett, M. L., Raetzman, S., Maggard-Gibbons, M., and Steiner, C. A. Surgical Site Infections Following Ambulatory Surgery Procedures. *JAMA*, 311(7):709, February 2014. ISSN 0098-7484. doi: 10.1001/jama.2014.4. (Cité en page 46.)

Panella, M., Marchisio, S., and Di Stanislao, F. Reducing clinical variations with clinical pathways: do pathways work? *International Journal for Quality in Health Care*, 15(6):509–521, December 2003. ISSN 1353-4505. doi: 10.1093/intqhc/mzg057. (Cité en page 14.)

Park, H. and Hardiker, N. R. Clinical terminologies : a solution for semantic interoperability. *Journal of Korean Society of Medical Informatics*, 15:1–11, March 2009. ISSN 1225-8903. doi: 10.4258/jksmi.2009.15.1.1. (Cité en page 2.)

Pathak, J., Kiefer, R. C., and Chute, C. G. Applying Linked Data Principles to Represent Patient's Electronic Health Records at Mayo Clinic: A Case Report. In *Proceedings of the 2Nd ACM SIGHIT International Health Informatics Symposium*, IHI '12, pages 455–464, New York, NY, USA, 2012a. ACM. ISBN 978-1-4503-0781-9. doi: 10.1145/2110363.2110415. (Cité en page 25.)

Pathak, J., Kiefer, R. C., and Chute, C. G. Using semantic web technologies for cohort identification from electronic health records for clinical research. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2012:10–19, 2012b. ISSN 2153-4063. (Cité en page 26.)

Pathak, J., Kiefer, R. C., and Chute, C. G. Using linked data for mining drug-drug interactions in electronic health records. *Studies in Health Technology and Informatics*, 192:682–686, 2013. ISSN 0926-9630. (Cité en pages 3 et 26.)

Paul, R., Groza, T., Hunter, J., and Zankl, A. Semantic interestingness measures for discovering association rules in the skeletal dysplasia domain. *Journal of Biomedical Semantics*, 5:8, February 2014. ISSN 2041-1480. doi: 10.1186/2041-1480-5-8. (Cité en page 35.)

Pesquita, C., Faria, D., Falcão, A. O., Lord, P., and Couto, F. M. Semantic Similarity in Biomedical Ontologies. *PLoS Computational Biology*, 5(7):e1000443, July 2009. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000443. (Cité en page 79.)

Pinaire, J., Azé, J., Bringay, S., and Landais, P. Infarctus du myocarde : quelles sont les trajectoires de soins pronostiques du décès à l'hôpital? In *IC*, 2017a. (Cité en page 30.)

Pinaire, J., Azé, J., Bringay, S., and Landais, P. Patient healthcare trajectory. An essential monitoring tool: a systematic review. *Health Information Science and Systems*, 5(1), April 2017b. ISSN 2047-2501. doi: 10.1007/s13755-017-0020-2. (Cité en page 14.)

Piro, R., Nenov, Y., Motik, B., Horrocks, I., Hendler, P., Kimberly, S., and Rossman, M. Semantic Technologies for Data Analysis in Health Care. In Groth, P., Simperl, E., Gray, A., Sabou, M., Krötzsch, M., Lecue, F., Flöck, F., and Gil, Y., editors, *The Semantic Web – ISWC 2016*, Lecture Notes in Computer Science, pages 400–417. Springer International Publishing, 2016. ISBN 978-3-319-46547-0. (Cité en page 25.)

Plantevit, M., Laurent, A., Laurent, D., Teisseire, M., and Choong, Y. W. Mining multidimensional and multilevel sequential patterns. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4:1–37, January 2010. doi: 10.1145/1644873.1644877. (Cité en page 74.)

Pollock, G. Holistic trajectories: a study of combined employment, housing and family careers by using multiple-sequence analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(1):167–183, January 2007. ISSN 1467-985X. doi: 10.1111/j.1467-985X.2006.00450.x. (Cité en page 87.)

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. (Cité en pages 38 et 66.)

Rinne, M., Abdullah, H., Törmä, S., and Nuutila, E. Processing Heterogeneous RDF Events with Standing SPARQL Update Rules. In Meersman, R., Panetto, H., Dillon, T., Rinderle-Ma, S., Dadam, P., Zhou, X., Pearson, S., Ferscha, A., Bergamaschi, S., and Cruz, I. F., editors, *On the Move to Meaningful Internet Systems: OTM 2012*, Lecture Notes in Computer Science, pages 797–806. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-33615-7. (Cité en page 47.)

Ritschard, G., Gabadinho, A., Mueller, N. S., and Studer, M. Mining event histories: A social science perspective. *International Journal of Data Mining, Modelling and Management*, 1(1):68–90, 2008. ISSN 1759-1163. (Cité en page 2.)

Rotter, T., Kinsman, L., James, E., Machotta, A., Gothe, H., Willis, J., Snow, P., and Kugler, J. Clinical pathways: effects on professional practice, patient outcomes, length of stay and hospital costs. *The Cochrane Database of Systematic Reviews*, (3):CD006632, March 2010. ISSN 1469-493X. doi: 10.1002/14651858.CD006632.pub2. (Cité en page 14.)

Roux, J., Grimaud, O., and Leray, E. Multichannel sequence analysis: An innovative method to study patterns of care pathways. Application to multiple sclerosis based

on French Health Insurance data. *Revue d'Épidémiologie et de Santé Publique*, 66: S430–S431, July 2018a. ISSN 0398-7620. doi: 10.1016/j.respe.2018.05.534. (Cité en page 87.)

Roux, J., Grimaud, O., and Leray, E. Use of state sequence analysis for care pathway analysis: The example of multiple sclerosis. *Statistical Methods in Medical Research*, page 0962280218772068, May 2018b. ISSN 0962-2802. doi: 10.1177/0962280218772068. (Cité en pages 2, 10 et 27.)

Salvadores, M., Horridge, M., Alexander, P. R., Fergerson, R. W., Musen, M. A., and Noy, N. F. Using SPARQL to Query Bioportal Ontologies and Metadata. In *Proceedings of the 11th International Conference on The Semantic Web - Volume Part II*, ISWC'12, pages 180–195, Berlin, Heidelberg, 2012. Springer-Verlag. ISBN 978-3-642-35172-3. doi: 10.1007/978-3-642-35173-0_12. (Cité en page 22.)

Samet, A., Guyet, T., and Negrevergne, B. Mining rare sequential patterns with ASP. In *ILP 2017 - 27th International Conference on Inductive Logic Programming*, Orléans, France, September 2017. (Cité en page 108.)

Schrijvers, G., van Hoorn, A., and Huiskes, N. The care pathway: concepts and theories: an introduction. *International Journal of Integrated Care*, 12(Special Edition Integrated Care Pathways), September 2012. ISSN 1568-4156. (Cité en page 14.)

Sharp, M. E. Toward a comprehensive drug ontology: extraction of drug-indication relations from diverse information sources. *Journal of Biomedical Semantics*, 8, January 2017. ISSN 2041-1480. doi: 10.1186/s13326-016-0110-0. (Cité en pages 25 et 65.)

Shaw, G., Xu, Y., and Geva, S. Eliminating redundant association rules in multi-level datasets. In *Faculty of Science and Technology*, pages 313–319, Las Vegas, Nevada, USA, March 2009. CSREA Press. ISBN 978-1-60132-062-9. (Cité en page 89.)

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., Consortium, T. O., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R. H., Shah, N., Whetzel, P. L., and Lewis, S. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11): 1251–1255, November 2007. ISSN 1546-1696. doi: 10.1038/nbt1346. (Cité en pages 21 et 23.)

Srikant, R. and Agrawal, R. Mining sequential patterns: Generalizations and performance improvements. In Apers, P., Bouzeghoub, M., and Gardarin, G., editors, *Advances in Database Technology — EDBT '96*, Lecture Notes in Computer Science, pages 1–17. Springer Berlin Heidelberg, 1996. ISBN 978-3-540-49943-5. (Cité en page 88.)

Studer, M. and Ritschard, G. What matters in differences between life trajectories: a comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179(2):481–511, February 2016. ISSN 1467-985X. doi: 10.1111/rssa.12125. (Cité en pages 27, 75 et 87.)

Tuppin, P., de Roquefeuil, L., Weill, A., Ricordeau, P., and Merlière, Y. French national health insurance information system and the permanent beneficiaries sample. *Revue D'epidemiologie Et De Sante Publique*, 58(4):286–290, August 2010. ISSN 0398-7620. doi: 10.1016/j.respe.2010.04.005. (Cité en page 10.)

Tuppin, P., Rudant, J., Constantinou, P., Gastaldi-Ménager, C., Rachas, A., de Roquefeuil, L., Maura, G., Caillol, H., Tajahmady, A., Coste, J., Gissot, C., Weill, A., and Fagot-Campagna, A. Value of a national administrative database to guide public decisions: From the système national d'information interrégimes de l'Assurance Maladie (SNIIRAM) to the système national des données de santé (SNDS) in France. *Revue D'epidemiologie Et De Sante Publique*, 65 Suppl 4:S149–S167, October 2017. ISSN 0398-7620. doi: 10.1016/j.respe.2017.05.004. (Cité en page 7.)

Whetzel, P. L., Noy, N. F., Shah, N. H., Alexander, P. R., Nyulas, C., Tudorache, T., and Musen, M. A. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Research*, 39(Web Server issue):W541–545, July 2011. ISSN 1362-4962. doi: 10.1093/nar/gkr469. (Cité en page 21.)

Wickham, H., François, R., Henry, L., and Müller, K. *dplyr: A Grammar of Data Manipulation*. 2018. (Cité en page 64.)

Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1 (6):80–83, 1945. ISSN 0099-4987. doi: 10.2307/3001968. (Cité en page 86.)

Williams, R., Buchan, I. E., Prosperi, M., and Ainsworth, J. Using String Metrics to Identify Patient Journeys through Care Pathways. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2014:1208–1217, 2014. ISSN 1942-597X. (Cité en pages 27 et 29.)

Willighagen, E. Accessing biological data in R with semantic web technologies. Technical Report e185v3, PeerJ Inc., March 2014. (Cité en pages 47 et 66.)

Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., and Hassanali, M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research*, 36(Database issue):D901–906, January 2008. ISSN 1362-4962. doi: 10.1093/nar/gkm958. (Cité en page 23.)

Wongsuphasawat, K., Plaisant, C., Taieb-Maimon, M., and Shneiderman, B. Querying Event Sequences by Exact Match or Similarity Search: Design and Empirical Evaluation. *Interacting with computers*, 24(2):55–68, March 2012. ISSN 0953-5438. doi: 10.1016/j.intcom.2012.01.003. (Cité en page 47.)

Wu, Z. and Palmer, M. Verbs semantics and lexical selection. pages 133–138. Association for Computational Linguistics, 1994. doi: 10.3115/981732.981751. (Cité en pages 79 et 80.)

Yamaguchi, A., Kozaki, K., Lenz, K., Wu, H., and Kobayashi, N. An Intelligent SPARQL Query Builder for Exploration of Various Life-science Databases. In *Proceedings of the 3rd International Conference on Intelligent Exploration of Semantic Data - Volume 1279*, IESD'14, pages 83–94, Aachen, Germany, Germany, 2014. CEUR-WS.org. (Cité en page 65.)

Zaki, M. J. SPADE: An Efficient Algorithm for Mining Frequent Sequences. *Machine Learning*, 42(1):31–60, January 2001. ISSN 1573-0565. doi: 10.1023/A:1007652502315. (Cité en page 88.)

Zhao, L., Yuan, S. S., Peng, S., and Wang, L. T. A New Efficient Data Cleansing Method. In Hameurlain, A., Cicchetti, R., and Traunmüller, R., editors, *Database and Expert Systems Applications*, Lecture Notes in Computer Science, pages 484–493. Springer Berlin Heidelberg, 2002. ISBN 978-3-540-46146-3. (Cité en page 28.)

UNIVERSITE BIOLOGIE BRETAGNE SANTE LOIRE

UNIVERSITÉ DE RENNES 1

Titre : Analyse de trajectoires de soins à partir de bases de données médico-administratives : apport d'un enrichissement par des connaissances biomédicales issues du Web des données

Mots clés : Santé publique, Pharmaco-épidémiologie, Bases de données médico-administratives, Trajectoires de soins, Web Sémantique

Résumé : Pour la recherche en santé publique, réutiliser les bases médico-administratives est pertinent et ouvre de nouvelles perspectives. En pharmaco-épidémiologie, ces données permettent d'étudier à grande échelle l'état de santé, les maladies ainsi que la consommation et le recours aux soins d'une population. Le traitement de ces données est cependant limité par des complexités inhérentes à la nature comptable des données. Cette thèse porte sur l'utilisation conjointe de bases de données médicoadministratives et de connaissances biomédicales pour l'étude des trajectoires de soin. Cela recouvre à la fois (1) l'exploration et l'identification des trajectoires de soins pertinentes dans des flux volumineux au moyen de requêtes et (2) l'analyse des trajectoires retenues.

Les technologies du Web Sémantique et les ontologies du Web des données ont permis d'explorer efficacement les données médico-administratives, en identifiant dans des trajectoires de soins des interactions, ou encore des contre-indications. Nous avons également développé le package R *queryMed* afin de rendre plus accessible les ontologies médicales aux chercheurs en santé publique.
Après avoir permis d'identifier les trajectoires intéressantes, les connaissances relatives aux nomenclatures médicales de ces bases de données ont permis d'enrichir des méthodes d'analyse de trajectoires de soins pour mieux prendre en compte leurs complexités. Cela s'est notamment traduit par l'intégration de similarités sémantiques entre concepts médicaux. Les technologies du Web Sémantique ont également été utilisées pour explorer les résultats obtenus.

Title : Care trajectory analysis using medico-administrative data : contribution of a knowledge-based enrichment from the Linked Data

Keywords: Medico-administrative databases, Care trajectories, Semantic Web

Abstract: Reusing healthcare administrative databases for public health research is relevant and opens new perspectives. In pharmaco-epidemiology, it allows to study large scale diseases as well as care consumption for a population. Nevertheless, reusing these information systems that were initially designed for accounting purposes and whose interoperability is limited raises new challenges in terms of representation, integration, exploration and analysis. This thesis deals with the joint use of healthcare administrative databases and biomedical knowledge for the study of patient care trajectories. This includes both (1) exploration and identification through queries of relevant care pathways in voluminous flows, and (2) analysis of retained trajectories.

Semantic Web technologies and biomedical ontologies from the Linked Data allowed to identify care trajectories containing a drug interaction or a potential contraindication between a prescribed drug and the patient's state of health. In addition, we have developed the R *queryMed* package to enable public health researchers to carry out such studies by overcoming the difficulties of using Semantic Web technologies and ontologies. After identifying potentially interesting trajectories, knowledge from biomedical nomenclatures and ontologies has also enriched existing methods of analysing care trajectories to better take into account the complexity of data. This resulted notably in the integration of semantic similarities between medical concepts. Semantic Web technologies have also been used to explore obtained results.