Interactions médicamenteuses : données liées et applications

Sébastien Cossin

université de BORDEAUX

Année 2016 Thèse n°3187

Thèse pour obtenir le grade de
DOCTEUR EN MEDECINE DE l'UNIVERSITÉ VICTOR SEGALEN

Spécialité **Santé publique et médecine sociale**

Présenté par

Sébastien COSSIN

Né le 20 Mars 1986 à Paris

INTERACTIONS MÉDICAMENTEUSES, DONNÉES LIÉES ET APPLICATIONS

Directeur de thèse :
M. Vianney JOUHET, Docteur en médecine

Soutenue le 30 novembre 2016 devant le jury composé de :

M. Roger SALAMON	Professeur	Président du jury
Mme. Catherine DUCLOS	Professeur	Rapporteur
M. Antoine PARIENTE	Professeur	Juge
M. Frantz THIESSARD	Maître de conférence	Juge
Mme. Béatrice SAINT-SALVI	Docteur	Juge
M. Vianney JOUHET	Docteur	Juge

Résumé

Introduction Les interactions médicamenteuses sont l'une des principales causes d'effets indésirables médicamenteux évitables. En France, le thesaurus des interactions médicamenteuses, édité par l'agence national de sécurité du médicament (ANSM), est la principale référence. Son format PDF est cependant une limite majeure à son utilisation. L'objectif principal de ce travail était de proposer un format ouvert du thesaurus et d'illustrer son intérêt au travers de deux exemples concrets : comparer le thesaurus avec d'autres sources internationales sur les interactions et utiliser son contenu pour mesurer la fréquence des interactions sur des données de délivrance.

Méthodes Nous avons développé un programme pour transformer le format PDF du thesaurus en triplets RDF, un standard du web sémantique pour l'échange de données. Les molécules décrites dans le thesaurus ont été liées à trois autres sources de données : l'Unified Medical Language System (UMLS), DrugBank (DB) et le répertoire des médicaments de l'ANSM. Le lien vers DB a permis son intégration dans une base de données internationale (merged-PDDI) sur les interactions et sa comparaison avec d'autres sources en mesurant leur chevauchement. Un programme a été développé pour mesurer la fréquence des interactions sur des données de délivrance médicamenteuse.

Résultats Dans sa version de janvier 2016, le thesaurus contient 1 102 molécules et 168 classes thérapeutiques. Il décrit 1 414 interactions entre classes ou molécules représentant 52 870 interactions entre molécules. Parmi ces dernières, 5 651 interactions (13%) ont été retrouvées dans la base merged-PDDI ce qui correspond à un chevauchement faible avec les autres sources. Nous montrons la possibilité d'utiliser ce contenu ouvert pour mesurer la fréquence des interactions sur des données de délivrance. Nous dénombrons 17 112 contre-indications parmi 7 millions de délivrances entre juin et août 2013.

Conclusion Nous proposons dans ce travail un format ouvert du thesaurus des interactions de l'ANSM qui utilise les standards du web sémantique. Ce format facilite son utilisation pour la recherche et permet le développement de nouvelles applications.

Mots Clés Interaction Médicamenteuse ; Web Sémantique ; Interopérabilité ; Pharmacovigilance

Abstract

Introduction One significant cause of adverse drug reactions is drug-drug interaction (DDI). In France, the thesaurus of DDI edited by ANSM (french national drug safety institute) is the main referential. However, its PDF format impedes its utilisation. The objective of this work was to propose an open format of the thesaurus and illustrate its usefulness for research through two examples : to examine the overlap between the thesaurus and other sources and to estimate the prevalence of potential DDIs based on drug deliveries.

Methods We developped a program that transforms the PDF format to RDF triples, a web semantic standard for data exchange. We linked its content to three sources : the Unified Medical Language System (UMLS), DrugBank (DB) and 'le répertoire des médicaments', a french drug database edited by ANSM. The link to DB enables its integration into an international DDI-database (merged-PDDI) and the measure of overlap with other sources. A program was created to estimate the prevalence of potential DDIs based on the thesaurus open format and drug deliveries data.

Results In its version of january 2016, the french thesaurus contains 1 102 drugs and 168 drug classes. It describes 1 414 interactions between classes or molecules resulting in 52 870 DDI. Among the latters, 5 651 (13%) were found in the merged-PDDI dataset corresponding to little overlap with other sources. We identified 17 112 contraindicated drug dispensations among 7 millions drug deliveries between june and august 2013.

Conclusion We propose a RDF format of the french thesaurus of DDI linked to UMLS, DB and the 'repertoire des médicaments'. This format can facilitate research on DDI and the development of new applications.

Keywords Drug Interaction ; Semantic Web ; Interoperability ; Pharmacovigilance

Remerciements

Je voudrais remercier tout d'abord mon directeur de thèse, Dr Jouhet, pour avoir accepté de m'accompagner sur ce sujet, pour son soutien, son enthousiasme et ses précieuses corrections.

Mes remerciements vont ensuite à tous les membres du jury :

au Professeur Duclos, rapporteur de ce travail. Merci d'avoir accepté d'évaluer ce travail et de l'avoir soutenu.

au Docteur Saint-Salvi. Je suis honoré par votre présence, je ne pouvais imaginer un meilleur expert du thesaurus des interactions dans ce jury. Merci pour vos remarques et vos conseils qui m'ont permis de corriger d'importantes erreurs dans ce document.

Au Docteur Thiessard. Merci de m'avoir fait découvrir l'informatique médicale et de m'avoir donné la possibilité de travailler avec vous.

Au Professeur Pariente, je vous remercie de l'intérêt que vous portez à mon travail et pour vos conseils.

Au Professeur Salamon, merci pour le soutien que vous apportez aux internes de santé publique, j'ai beaucoup appris à vos côtés. Je vous remercie de l'honneur que vous me faîtes en acceptant de présider ce jury.

J'adresse toute ma gratitude à Nora pour son aide, ses relectures et son soutien inestimable au quotidien.

A mes parents et à toute ma famille pour m'avoir fourni l'environnement nécessaire, pour avoir cru en moi et m'avoir soutenu.

A mes amis qui, même éloignés, sont toujours présents, et particulièrement mes amis d'enfance.

A l'équipe ERIAS, pour m'avoir fait confiance et pour m'avoir formé à l'informatique médicale.

A tous mes collègues de travail durant cet internat qui m'ont tous accueilli et soutenu dans ma formation. Ce fut un plaisir de travailler avec vous.

A l'équipe AquitHealth. Je n'aurais jamais choisi ce sujet si je n'étais pas venu avec vous au hackathon à Strasbourg.

A la communauté du logiciel libre pour la quantité et la qualité des outils que vous développez et partagez. Vive la liberté d'exécuter, copier, étudier, modifier, améliorer et distribuer.

A Mr Chipper pour sa sagesse et sa compagnie lors de mes promenades.

Table des Matières

Glossaire

— **CIS** Code Identifiant de Spécialité
Code numérique de 8 chiffres identifiant une spécialité pharmaceutique.
— **Spécialité pharmaceutique**
Médicament fabriqué industriellement identifié par une dénomination.
— **CIP** Code numérique Identifiant une Présentation
Code numérique identifiant une présentation d'une spécialité pharmaceutique.
— **SA** Substance active
Composant d'une spécialité pharmaceutique reconnu comme possédant les propriétés thérapeutiques.
— **FT** Fraction thérapeutique
Partie d'une substance active qui porte l'activité pharmacologique. Ne concerne pas les enregistrements des médicaments homéopathiques.
— **LAP/LAD** Logiciel d'Aide à la Prescription/Délivrance
Logiciel utilisé par les professionnels de santé à l'hôpital ou en ville pour les aider dans leur prescription/délivrance. Ils intègrent une base de données sur les médicaments.
— **AMM** Autorisation de Mise sur le Marché
Pour être commercialisée, une spécialité pharmaceutique doit obtenir préalablement une autorisation de mise sur le marché.
— **RCP** Résumé des Caractéristiques du Produit
Annexe de l'AMM contenant les informations destinées aux professionnels de santé, notamment les indications thérapeutiques, les contre-indications, les modalités d'utilisation et les effets indésirables du médicament.
— **SGBD** Système de Gestion de Base de Données
Logiciel destiné à stocker et à partager des informations dans une base de données, en garantissant la qualité, la pérennité et la confidentialité des informations, tout en cachant la complexité des opérations.
— **W3C** World Wide Web Consortium
Organisme de standardisation du web.
— **NLM** National Library of Medicine
— **API** Application Programming Interface
Ensemble de fonctions permettant à une application de fournir des services à une autre application

Introduction

2.1 Les interactions médicamenteuses

Les effets indésirables provoqués par les médicaments sont une cause importante de surcoût et de mortalité [1]. Les interactions médicamenteuses (IM) sont l'une des principales causes évitables[2].

Un comité d'experts américain [3] définit en 2015 une interaction médicamenteuse comme « une altération cliniquement pertinente de l'effet d'un médicament survenue à cause de la co-administration d'un autre médicament ». Cette altération peut provoquer la survenue d'un évènement indésirable ou modifier l'effet thérapeutique du médicament. Une interaction médicamenteuse cliniquement pertinente est « une interaction qui entraîne une toxicité ou une perte d'efficacité thérapeutique qui justifie l'attention d'un professionnel de santé ».

Une interaction médicamenteuse potentielle est définie comme « la co-prescription de deux médicaments connus pour interagir et exposant le patient à un risque de survenue d'une interaction médicamenteuse »[3]. Une interaction médicamenteuse est donc un évènement qui entraîne un effet indésirable et une interaction médicamenteuse potentielle correspond au risque de développer cet évènement. Dans la suite de ce travail, le terme interaction médicamenteuse sera employé à la fois pour désigner l'évènement et le risque de l'évènement.

Les interactions médicamenteuses sont classifiées en deux grands types[4] :

— Pharmacocinétiques

Les interactions pharmacocinétiques conduisent à une modification de la concentration sanguine du médicament "cible". Une augmentation de la concentration majore les effets et peuvent entraîner des évènements indésirables. Une diminution de la concentration provoque une perte d'efficacité thérapeutique. L'interaction peut survenir à plusieurs étapes : l'absorption, la distribution, le métabolisme ou l'élimination du médicament. A chacune d'elles, différents mécanismes sont en jeu. Par exemple, la résorption peut être altérée suite à une complexation avec un autre médicament, à une modification du pH gastrique, à un retard d'évacuation gastrique, à une accélération du transit intestinal...[4]

Les interactions pharmacocinétiques vont avoir une traduction clinique pour les médicaments à marge thérapeutique étroite. Ces derniers sont définis par le comité d'experts [3] comme « un médicament dont une augmentation inférieure à 100% de l'aire sous la courbe des concentrations plasmatiques conduit à des effets indésirables graves ou dont une diminution de 50% de l'aire sous la courbe des concentrations

plasmatiques conduit à une perte d'efficacité ».

— Pharmacodynamiques

Les interactions pharmacodynamiques résultent d'une interférence au niveau du site d'action du médicament. Les mécanismes peuvent être directs (même site d'action) ou indirects et conduire à une perte d'efficacité d'un médicament, une augmentation des effets thérapeutiques et/ou des effets indésirables. Par exemple, l'association d'un antagoniste et d'un agoniste ou l'association de deux agonistes sont des interactions pharmacodynamiques. Contrairement aux interactions pharmacocinétiques, les interactions pharmacodynamiques concernent des classes pharmacologiques.

En pratique, les interactions médicamenteuses peuvent être issues de plusieurs mécanismes différents.

Dans quelques situations, les interactions médicamenteuses sont volontairement recherchées. C'est par exemple le cas de la naloxone, un antagoniste des récepteurs de la morphine, administrée lors d'un surdosage en morphine.

De nombreux éléments contextuels peuvent modifier le risque de développer une interaction médicamenteuse. Ils peuvent être liés au médicament comme la posologie, la durée d'administration et la voie d'administration ou liés aux caractéristiques physiopathologiques du patient comme l'âge (sujet âgé), le sexe, les pathologies (insuffisance rénale ou hépatique sévère) et la pharmacogénomique (métaboliseur lent ou rapide)[3].

La conduite à tenir clinique devant un risque d'interaction médicamenteuse peut être de [5] :

— Surveiller le patient au plan clinique et biologique
— Eviter l'association des deux molécules
— Espacer dans le temps la prise des médicaments
— Ajuster la dose

Les personnes polypathologiques et les personnes âgées sont particulièrement à risque car la multiplicité des affections augmente le nombre de médicaments et le terrain fragile rend plus dramatique la survenue d'un effet indésirable[6]. Le risque d'interactions médicamenteuses est une préoccupation des professionnels de santé lors de la prescription et de la dispensation.

Le risque d'interaction médicamenteuse entre deux molécules peut être estimé à partir d'études randomisées, d'études de cohorte, d'observations cliniques, de la détection de signal dans une base de données ou d'extrapolations d'études in vitro[7]. Différents organismes publics et privés, français et étranger, réalisent un travail de synthèse des informations sur les interactions médicamenteuses.

2.2 Un problème d'Open Data

En France, le thesaurus des interactions médicamenteuses de l'ANSM est la principale référence à consulter. D'après le site de l'ANSM, « ce thesaurus apporte aux professionnels

de santé une information de référence, à la fois fiable et pragmatique » et « il doit être utilisé comme un guide pharmaco-thérapeutique d'aide à la prescription ». Cependant, son format PDF est un frein à la recherche d'information par les professionnels de santé. Il n'est pas possible d'interroger automatiquement ce fichier pour savoir si deux molécules sont concernées par une interaction, ni si deux médicaments prescrits interagissent. Aussi, son format actuel restreint son utilisation dans le cadre de la recherche. Des travaux récents [8] [9] ont intégré plusieurs sources de données sur les interactions médicamenteuses mais se sont limitées aux sources ouvertes et accessibles. Pour être accessible, le contenu du thesaurus nécessite d'être présenté dans un format exploitable par une machine. Son accessibilité est un problème d'Open Data.

L'Open Data désigne l'ouverture et le partage de données par leur mise en ligne dans des formats ouverts, en autorisant la réutilisation libre et gratuite par toute personne. Dans le rapport de la commission Open Data en santé remis à Marisol Touraine le 9 juillet 2014 [10], le thesaurus des interactions médicamenteuses de l'ANSM figure dans la liste des bases dont l'ouverture est souhaitée dans « les plus brefs délais » et « gratuitement, en format ouvert à tous et réutilisable ». Les principaux enjeux de l'Open Data sont de développer la démocratie sanitaire, de soutenir l'innovation et la recherche. Les modalités de mise à disposition des données doivent respecter le principe de la plus grande liberté de réutilisation, en évitant le plus possible les contraintes techniques, financières, juridiques ou autre. Pour éviter les contraintes techniques, l'ouverture des données nécessite de respecter certains standards.

L'organisme de standardisation du web, le World Wide Wibe Consortium (W3C), propose un système de notation d'une à cinq étoiles (figure 2.1) pour des données publiées sur le web :

- Une étoile si les données sont publiées sous licence libre (Open License - OL), lisibles par un être humain mais non accessibles à une machine à cause du format (par exemple PDF) ou de la structure du fichier.
- Deux étoiles si les données sont accessibles à une machine (Machine Readble - RE) mais dans un format propriétaire [1] (par exemple un fichier Excel)
- Trois étoiles si les données sont publiées dans un format ouvert (Open Format - OF)
- Quatre étoiles si les données sont publiées dans un format recommandé par le W3C (RDF) pour le web de données et si elles sont identifiées par des identifiants uniques sur le web (URI)
- Cinq étoiles si les données sont liées à d'autres ressources sur le web (Linked Data - LOD)

1. format lisible par un logiciel particulier nécessitant l'achat d'une licence

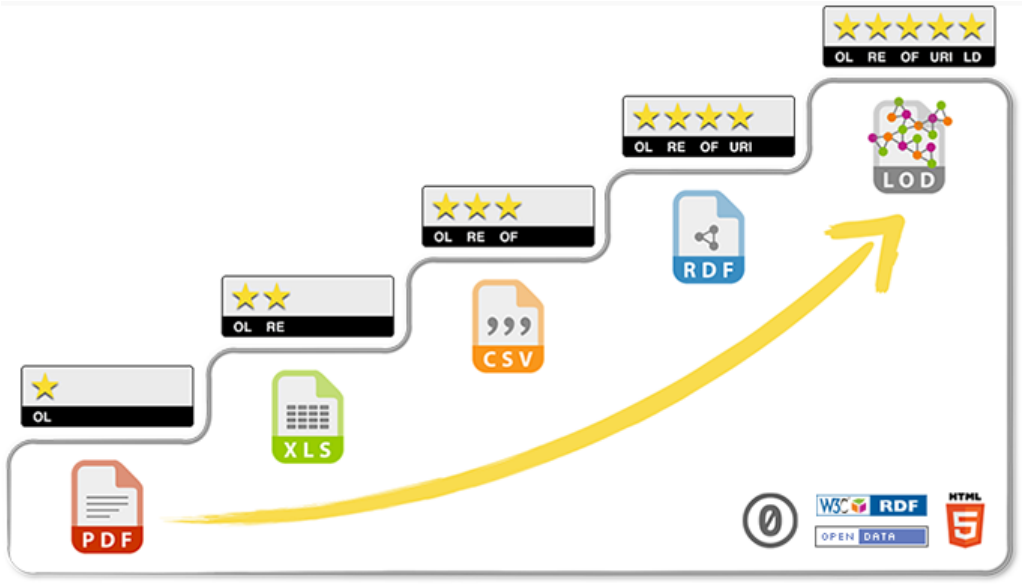

FIGURE 2.1 – Principe des cinq étoiles d'une donnée publiée sur le web. OL : Open License, RE : Machine Readble, OF : Open Format, URI : Unique Ressource Identifier, LOD : Linked Data. Source : http://5stardata.info/en/

2.3 Objectifs

L'objectif de ce travail est de proposer un format ouvert cinq étoiles du thesaurus et d'illustrer son intérêt pour la recherche au travers de deux exemples concrets (objectifs secondaires).

Les données du thesaurus seront extraites et son contenu sera lié à des ressources sur le web sémantique. Le lien réalisé vers d'autres sources de données internationales sur les interactions médicamenteuses permettra de comparer son contenu avec ces sources : ce sera le premier exemple d'application.

Le deuxième exemple sera de mesurer la fréquence des interactions sur des données de délivrance. Il est intéressant de réaliser cette analyse à plusieurs titres :

— La veille sanitaire

La fréquence des interactions sur les ordonnances délivrées, notamment celle des contre-indications, reflète le risque de développer un évènement indésirable grave par interaction médicamenteuse. Il est intéressant d'identifier les principales interactions et de créer un indicateur pour suivre dans le temps l'évolution du risque d'interaction médicamenteuse dans la population. D'après une étude menée par l'assurance maladie [11], les résultats permettent alors d'engager avec les professionnels de santé une réflexion sur la prévention de ce risque.

— La comparaison des sources

Le désaccord entre deux sources sur les interactions est d'autant plus problématique

que celui concerne une contre-indication et que les deux molécules sont fréquemment prescrites ensembles. Il est ainsi possible de trier les désaccords entre deux sources selon leur fréquence d'occurrence sur des données de délivrance.

— Le nombre d'alertes générées
De nombreuses études [12] ont montré qu'un excès d'alertes ou d'informations envoyées aux cliniciens généraient une surcharge cognitive élevée rendant difficile la filtration des informations pertinentes en un temps raisonnable. Il est intéressant de savoir si le thesaurus est enclin à générer de nombreuses alertes.

Le prochain chapitre présente les technologies du web sémantique et les principales sources de données biomédicales accessibles sur ce web. Le chapitre "matériel" présente les données et les outils informatiques utilisés dans ce travail. Le chapitre "méthodes" décrit comment les données du thesaurus ont été extraites, transformées, liées puis comparées à d'autres sources sur les interactions et utilisées pour analyser des délivrances médicamenteuses.

Contexte

3.1 Le web sémantique

A l'heure actuelle, les données biomédicales sont produites en grande quantité, à une vitesse toujours plus grande et sont accessibles dans divers formats. Les données exprimant des connaissances sur un même concept biomédical sont rarement liées sur le web. Les données en silo, c'est-à-dire isolées les unes des autres, posent un problème majeur pour la recherche d'information [13]. La quantité d'information produite étant trop grande pour être traitée par des êtres humains, des outils informatiques sont développés. Intégrer les données, gérer leur provenance et extraire les informations avec leur degré de certitude puis synthétiser les connaissances et les retourner à un utilisateur sont des défis actuels en informatique médicale.

Les traitements réalisables par une machine sont limités par la façon dont les données sont stockées et représentées. Les technologies du web sémantique facilitent la représentation, la publication, les liens entre les données et la recherche d'information [13]. Au lieu de naviguer à travers des pages webs, le web sémantique propose de naviguer à travers les données [14]. D'un web de données en silo où l'information est inaccessible aux machines, le web sémantique, appelé parfois web 3.0, est constitué de liens sémantiques entre plusieurs sources de données distribuées sur le web. Le web sémantique est une encyclopédie de données à l'échelle mondiale, enrichissable par quiconque respectant les règles de publication et accessible aux machines.

L'intégration des données du web est confrontée aux problèmes d'interopérabilité technique et sémantique. Le problème d'interopérabilité technique résulte de la présence de données dans divers formats (PDF, Texte...) et encodées de différentes façons (ISO8559-1,UTF8-UTF16...). L'expression de différentes manières d'un même concept pose le problème d'interopérabilité sémantique : la signification des informations contenues dans les données ne doit pas être ambiguë. Les technologies du web sémantique offrent des solutions pour résoudre ces problèmes d'interopérabilité au travers de standards et d'outils. Elles incluent un modèle de données (RDF), des syntaxes pour sérialiser les données (XML, Turtle), un langage de requête (SPARQL) et des schémas (RDFS, OWL) pour décrire les métadonnées et raisonner sur les données[15]. Ces technologies sont décrites dans les sections suivantes.

3.1.1 RDF

Resource Description Framework (RDF) est un modèle graphe de données. Son unité de base est un triplet constitué de :

— un sujet, il représente la ressource à décrire
— un prédicat (ou relation ou propriété), il décrit une propriété du sujet
— un objet, il représente une valeur (texte, numérique, date...) ou une autre ressource

Par exemple, « le paracétamol est une molécule » est une affirmation en langage naturel qui peut être décrite en RDF : le paracétamol est le sujet, le prédicat est la relation "est un" et l'objet est le concept de molécule. Pour identifier les ressources et les propriétés, un identifiant unique est utilisé : l'URI (Unique Ressource Identifier ou identifiant uniforme de ressource). Par exemple, l'encyclopédie DrugBank identifie le paracétamol par l'URI *http://bio2rdf.org/drugbank:DB00316* et DBpedia, la version structurée de Wikipédia, utilise l'URI *http://dbpedia.org/resource/Paracetamol*. Un triplet RDF peut affirmer que ces deux URI décrivent le même concept et réalise un lien entre DrugBank et DBpédia. Une machine est alors en mesure de réunir les assertions sur le paracétamol de ces deux sources de données sur le web. Les triplets RDF liant des sources de données différentes sont les fondements du web de données liées.

Les URL (Unique Ressource Locator) identifient un document sur le web (page web, fichier audio ...) et constituent un sous-ensemble des URI qui identifient aussi des objets du monde réel (une personne, une ville...) ou un concept abstrait (molécule). Les URI servent à identifier ou référencer toute chose réelle ou non. Bien que les URI contiennent le préfixe http:// comme les URL, il n'est pas garanti qu'une page web soit retournée si l'URI est entré dans la barre d'adresse d'un navigateur. Il est cependant considéré comme une bonne pratique de retourner une description d'un URI quand un navigateur cherche à y accéder. Ce procédé s'appelle le déréférencement de l'URI. Des exemples de déréférencement peuvent être observés en entrant les URI précédents dans un navigateur.

Les propriétés (ou relations) sont aussi identifiées par des URI. Contrairement aux liens hypertextes du web classique, les liens du web de données ont un sens. Les propriétés peuvent être des liens internes pour décrire les données ou des liens externes pour se lier à d'autres sources. Une source de données sur le web sera d'autant plus utile et utilisable qu'elle sera connectée à d'autres sources.

3.1.2 Syntaxes

RDF n'est pas un format de données. Il est nécessaire de définir un format ou syntaxe pour publier des données RDF. Le W3C reconnaît plusieurs syntaxes pour stocker l'information d'un triplet RDF en mémoire : RDF/XML, RDFa, N3, Turtle... La syntaxe Turtle a l'avantage d'être facilement lisible par un être humain. Les trois triplets RDF suivants, au format Turtle, sont donnés à titre d'exemple :

```
@prefix db: <http://bio2rdf.org/drugbank:> .
@prefix dv: <http://bio2rdf.org/drugbank_vocabulary:> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
```

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

db:DB00316 rdf:type dv:Small-molecule ;
                    rdfs:label "Paracetamol"@fr ;
                    rdfs:label "Acetaminophen"@en .
```

Les préfixes, déclarés avant les triplets, améliorent la lisibilité des déclarations. "db:DB00316" est équivalent à *http://bio2rdf.org/drugbank:DB00316*. Les propriétés "rdf:type" et "rdfs:label" sont définies par le W3C comme leur préfixe l'indique. Le point termine la déclaration d'un triplet comme il termine une phrase en langage naturel. Le point virgule permet de déclarer un nouveau triplet en conservant le sujet. La traduction en langage naturel de ces triplets en syntaxe Turtle peut être la suivante : « la ressource DrugBank dont l'identifiant est DB00316 appartient à la classe "Small-molecule", son libellé en français est "paracetamol", son libellé en anglais est "acetaminophen".» DrugBank fournit la description suivante de sa classe "Small-molecule" : « médicaments qui n'ont pas une origine biologique et qui sont synthétisés ». Ces triplets peuvent être représentés graphiquement (figure 3.1).

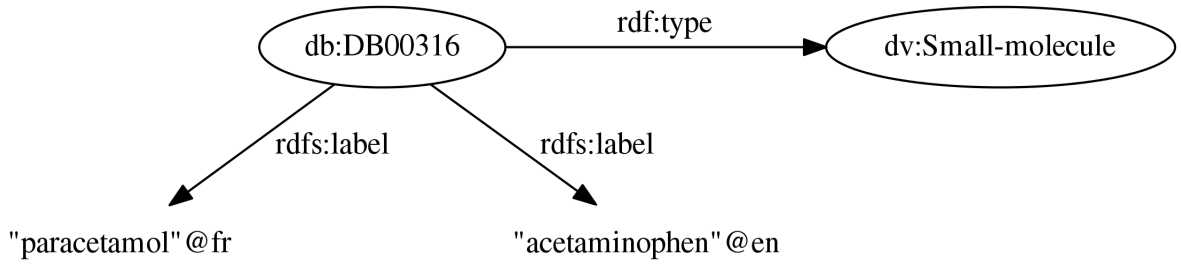

FIGURE 3.1 – Représentation graphique de trois triplets RDF

3.1.3 Triplestore et SPARQL

Les triplets RDF forment un graphe qu'il est possible de stocker et d'interroger. Un triplestore est un système de gestion de bases de données (SGBD) spécialement conçu pour les données RDF. Il permet de stocker de façon efficiente les données RDF pour les interroger et les récupérer avec le langage de requête SPARQL. Il existe de nombreux triplestores dont certains sous licence libre. L'interface d'un triplestore qui permet d'envoyer des requêtes SPARQL est appelé un SPARQL endpoint. De nombreux sites Open Data sur le web fournisse un SPARQL endpoint pour interroger leur base de données. `http://dbpedia.org/sparql` et `http://drugbank.bio2rdf.org/sparql` sont des endpoint pour accéder à DBpedia et Drugbank respectivement. L'une des forces du langage SPARQL est d'offrir la possibilité d'interroger plusieurs SPARQL endpoint simultanément lors d'une même requête. Contrairement aux entrepôts de données où toutes les données doivent être stockées au même endroit

avant d'être exploitées, le web sémantique fournit une solution d'intégration distribuée [13]. Elle permet de récupérer des informations distribuées sur le web sans avoir besoin de télécharger l'intégralité des bases de données. Il est par exemple possible de récupérer toutes les connaissances sur le paracétamol sur le web de données.

3.1.4 Schémas et représentation des connaissances

Les données contiennent souvent une structure qu'il est utile de décrire formellement. RDFS (RDF Schema) est un langage de représentation des connaissances. Il définit par exemple la notion de "Classe" (rdfs:Class) qui permet d'expliquer à une machine que le concept de molécule est une classe et que le paracétamol est un exemple concret ou instance de cette classe. Une classe peut être vue comme un ensemble d'objets possédant des propriétés communes. RDFS possède des extensions pour exprimer des relations plus complexes entre les concepts. Le langage OWL est un exemple d'extension de RDFS. Il permet de définir des relations basées sur la logique de description (quantificateur universel, restriction de cardinalité, disjonction...). Son expressivité permet de créer des ontologies. Une ontologie est une spécification explicite d'une conceptualisation [16]. Elle permet de représenter les connaissances de façon formelle, c'est-à-dire compréhensibles par une machine. Une ontologie est composée principalement de concepts, des instances de concepts, de relations et d'axiomes. La machine peut utiliser une ontologie pour inférer, à l'aide de programmes appelés "raisonneur", de nouvelles connaissances non décrites explicitement dans les données. Par exemple, un raisonneur peut déduire que le paracétamol est une molécule même si cette affirmation n'est pas présente dans les triplets RDF. Grâce à cette technologie, il n'est pas nécessaire de stocker en mémoire toutes les informations dont on dispose car il est possible de les générer par raisonnement.

3.1.5 Web sémantique de données pharmacologiques

De nombreuses sources de données sur les médicaments existent sur le web : résultats d'essais cliniques, effets des molécules sur l'expression génique, propriétés pharmacologiques des molécules, ventes de médicament... LODD (Linking Open Drug Data) est un groupe de travail [1] du W3C qui a rendu disponible plusieurs bases de données biomédicales dans un format ouvert et lié (figure 3.2).

1. task force en anglais

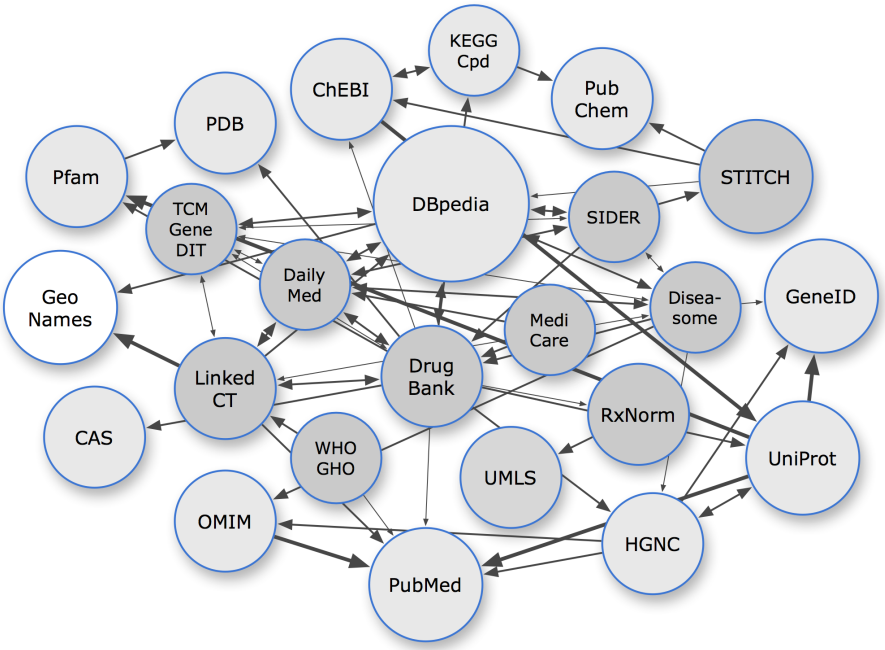

FIGURE 3.2 – Bases de données publiées par LODD (en gris foncé), reliées à d'autres sources biomédicales (en gris clair) et à des sources généralistes (en blanc). Une flèche directe d'une source A vers une source B indique que la source A contient des triplets liés à des ressources de B. L'épaisseur des flèches est proportionnelle au nombre de liens. Source : Samwald et al. [17]

Les données biomédicales liées sont utilisées par les chercheurs et les industriels. Par exemple, le projet Open Pharmacological Concepts Triple Store (Open PHACTS)[18] réunit de nombreuses sources de données au format RDF (DrugBank, GeneOntology, ChemSpider...) pour répondre à des questions complexes et découvrir de nouvelles molécules.

Il est recommandé de lier les données vers des ressources répandues et interconnectées sur le web. Nos recherches sur le web ont identifié trois principales sources pour relier les données du thesaurus : l'UMLS, DrugBank et le répertoire du médicament de l'ANSM. Ces sources sont très connectées (figure 3.3) et sont présentées dans les sections suivantes.

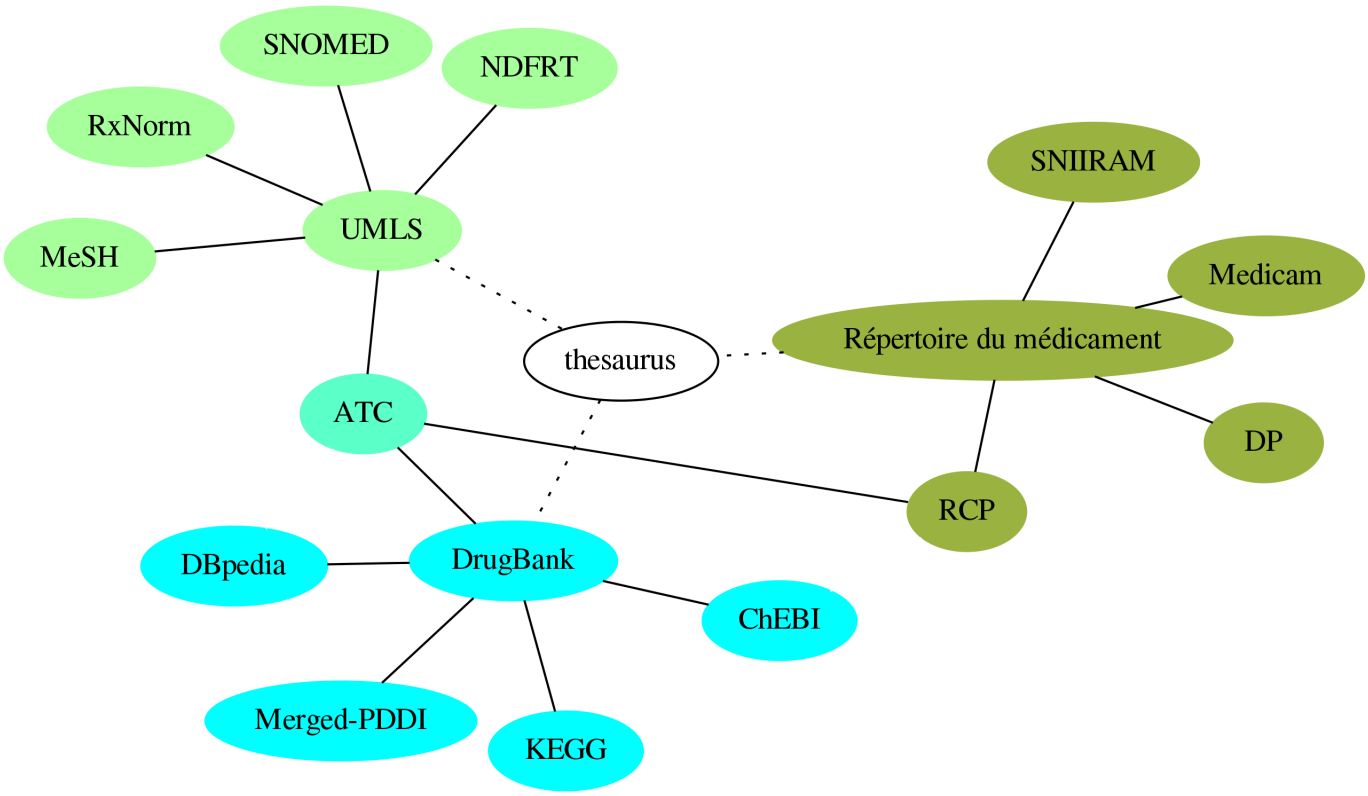

FIGURE 3.3 – Trois sources ont été identifiées pour relier les données du thesaurus sur le web : l'UMLS, DrugBank et le répertoire des médicaments de l'ANSM. En pointillé les liens à réaliser. En trait plein, des liens déjà existants

3.2 Sources de données sur les molécules et les médicaments

3.2.1 UMLS

L'UMLS (Unified Medical Language System) est l'une des plus grandes ressources biomédicales sur le web. Elle intègre plusieurs terminologies et classifications dans le domaine biomédical [19] dont la CIM-10 utilisée pour le codage des pathologies, LOINC pour la biologie, MeSH pour le référencement des articles biomédicaux, RxNorm pour coder les médicaments et la plus grande terminologie médicale au monde : la SNOMED_CT. Chaque code d'une classification, appelé atome, est rattaché à un concept UMLS (CUI : Concept Unique Identifier) et dès lors lié aux atomes des autres terminologies rattachés au même concept. Un concept UMLS est représenté par un libellé préféré choisi parmi les libellés des différentes terminologies. Par exemple, le libellé préféré du concept C0027051 est "myocardial infarction" et de multiples terminologies y font référence via d'autres libellés comme "heart attack", "heart infarction", "cardiac infarction" en anglais et "infarkt myokardu", "infarctus du myocarde" en tchèque et en français respectivement. L'UMLS permet ainsi de rendre interopérables des systèmes informatiques utilisant des terminologies différentes. Il est aussi utilisé pour la recherche d'information, la fouille de données et la recherche bioinformatique. Il a été conçu et est toujours maintenu par la National Library of Medicine (NLM), institution gouvernementale des Etats-Unis. Dans sa dernière version disponible au moment de l'écriture de ce travail (2016AA), il contient plus de 3,25 millions de concepts et environ 13 millions de termes uniques provenant de 190 terminologies[20].

Bien que le metathesaurus soit multilingue, la langue anglaise est largement prédominante avec 128 terminologies présentes alors que seulement 9 sont en langue française. Il contient plusieurs terminologies décrivant des médicaments dont RxNorm.

3.2.2 RxNorm

L'utilisation de différentes terminologies sur les médicaments a conduit la NLM à construire RxNorm pour permettre l'interopérabilité des systèmes d'information[21]. De façon similaire au metathesaurus, RxNorm est construit à partir des terminologies qu'elle contient. Elle définit des concepts RxNorm, appelés RXCUI, et rattache les codes des terminologies sur les médicaments à ses concepts. Comme RxNorm est intégré dans l'UMLS, chaque concept RxNorm (RXCUI) est rattaché à un concept UMLS (CUI). Dans la version 2016AA, RxNorm contient 14 terminologies dont l'ATC (la classification Anatomique, Thérapeutique et Chimique), le MeSH, la partie médicament de la SNOMED_CT et NDF-RT (National Drug File Reference Terminology). Contrairement à l'UMLS, les libellés préférés ne sont pas issus des terminologies qu'elle contient. RxNorm produit des libellés normalisés revus par des experts

humains [21] pour représenter ses concepts. La NLM donne l'exemple suivant : "Naproxen Tab 250 MG", "Naproxen 250 tablet (product)", "Naproxen@250 mg@ORAL@TABLET", "Naproxen 250 MILLIGRAM In 1 TABLET ORAL TABLET" et "NAPROXEN 250MG TAB,UD [VA Product]" sont tous des synonymes, issus de différentes sources, du même concept. Ils sont tous rattachés à un concept RxNorm(RXCUI) dont l'identifiant unique 198013 est le numéro utilisé par les systèmes d'information pour communiquer. Ce concept a pour libellé préféré "Naproxen 250 MG Oral Tablet" qui est un libellé normalisé par un expert de la NLM. Le libellé normalisé d'un médicament est produit en combinant ses ingrédients, leur dosage et la forme galénique.

Pour chaque nouveau médicament, les concepts suivant : ingrédient, ingrédient + dosage, ingrédient + voie d'administration sont créés s'ils n'existent pas encore. Dans l'exemple précédent, ils correspondent aux concepts "Naproxen"(7258), "Naproxen 250 MG"(316326) et "Naproxen Oral Tablet"(373005). RxNorm est donc à la fois une solution d'intégration de plusieurs terminologies sur les médicaments et une terminologie en elle-même, appelée SAB=RXNORM.

Celle-ci décrit des relations entre les concepts : "a_pour_ingredient"(has_ingredient), "a_pour_nom_commercial"(has_tradename_of)... Ces relations sont très utiles lors de la prescription ou la dispensation pour rechercher par exemple les différents dosages d'un ingrédient ou les médicaments génériques.

RxNorm bénéficie aussi des relations des terminologies qu'elle contient. Par exemple, la terminologie NDFRT indique que l'ingrédient amoxicilline peut traiter la maladie de Lyme. RxNorm est disponible sous forme de triplets RDF. Des exemples de triplets RDF décrivant au format Turtle le concept 316326 (Naproxen 250 MG) est présenté ci-dessous. Des commentaires en langage naturel figurant après // ont été ajoutés.

```
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix owl:  <http://www.w3.org/2002/07/owl#> .
@prefix rdfs:  <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix umls: <http://bioportal.bioontology.org/ontologies/umls/> .
@prefix RxNorm: <http://purl.bioontology.org/ontology/RXNORM/>

RxNorm:316326 a owl:Class ;        // RxNorm:316326 est une classe OWL
        skos:prefLabel "Naproxen 250 MG"@eng ; // son libellé préféré
        RxNorm:has_ingredient RxNorm:7258> ; // contient l'ingrédient Naproxen
        RxNorm:has_tradename RxNorm:564237> ; // a pour nom commercial Naprosyn
        RxNorm:RXN_STRENGTH "250 MG"^^xsd:string ; // son dosage
        RxNorm:RXAUI "1478753"^^xsd:string ; // identifiant de l'atome dans RxNorm
        RxNorm:RXCUI "316326"^^xsd:string ; // identifiant RxNorm
        umls:cui "C0987997"^^xsd:string . // identifiant UMLS
```

La terminologie RXNORM définit plusieurs concepts qu'il est intéressant d'étudier car ceux-ci seront retrouvés par la suite dans le thesaurus des interactions et dans les données sur les médicaments :

— Ingrédient
  Il s'agit d'un composé qui donne à un médicament sa propriété clinique. Ils sont généralement exprimés en dénomination commune internationale (DCI) en France ou en dénomination commune américaine (USAN : United States Adopted Name). Par exemple, l'ésoméprazole est un ingrédient.

— Ingrédient précis [2]
  La forme spécifique d'un ingrédient. La plupart sont des sels ou des formes isomères. Par exemple, l'ésoméprazole sodique est l'ingrédient précis de l'ésoméprazole.

L'ingrédient est relié à l'ingrédient précis par la relation "RxNorm:has_form". La relation inverse est "RxNorm:form_of".

3.2.3 DrugBank

DrugBank [22] est un projet scientifique canadien financé par Genome Canada et une base de données regroupant de nombreuses informations biochimiques et pharmacologiques sur les molécules et les médicaments. Il vise à fournir des informations complètes aux chercheurs, chimistes, pharmaciens, médecins et au grand public [23]. Sa première version est parue en 2006 et était limitée aux médicaments approuvées par la FDA (Food and Drug Administration) avec leur cible thérapeutique. Dans sa version actuelle (5.0), il contient des informations biochimiques et pharmacologiques sur 8 206 molécules : formule chimique, poids et structure moléculaire, absorption, biodisponibilité, métabolisme, cible thérapeutique, demi-vie, clairance, code ATC, interactions pharmacodynamiques, toxicité ... Il affiche aussi des liens vers d'autres bases de données (DBpedia, KEGG Drug, ChEBI, RxList, Drugs.com ...). Drugbank est l'une des base de données les plus consultées avec plus de 8 millions de visites par an [23]. Son contenu est librement accessible en ligne http://www.drugbank.ca.

3.2.4 Répertoire des spécialités pharmaceutiques

En France, chaque présentation d'une spécialité pharmaceutique est identifiée par un code dit "code CIP" [24]. Ce code figure sur la boîte d'un médicament (figure 3.4), il permet d'identifier chaque médicament dans la base de données nationale de l'assurance maladie et correspond à son numéro d'autorisation de mise sur le marché (AMM).

Une spécialité pharmaceutique identifie un médicament par son nom commercial, son titulaire (laboratoire pharmaceutique), son dosage et sa forme pharmaceutique (comprimé,

2. Precise ingrédient en anglais

gélule, sirop...). Chaque spécialité pharmaceutique est identifiée par un code, appelé "code CIS", et est associée à un résumé des caractéristiques du produit (RCP).

Par exemple, "MOTILIUM 10 mg, comprimé pelliculé" (figure 3.4) est une spécialité pharmaceutique qui a pour code CIS 63679194, sa vente est autorisée dans une plaquette d'aluminium PVC de 40 comprimés (code CIP : 3400932341122) ou 30 comprimés (code CIP : 3400933688295).

FIGURE 3.4 – "MOTILIUM 10 mg, comprimé pelliculé" est le nom de la spécialité pharmaceutique de ce médicament. Sa substance active est la dompéridone. Sa présentation sous forme de 40 comprimés dans une plaquette d'aluminium PVC est identifiée par le code à 13 chiffres : 3400932341122. Son titulaire est le laboratoire Janssen.

La spécialité pharmaceutique générique "DOMPERIDONE ALMUS 10 mg, comprimé pelliculé" de ce médicament est identifiée par un code CIS différent (63630635). Ainsi, il existe par exemple deux RCP différents pour ce médicament et son générique même si les contenus de ces RCP sont très proches.

Une spécialité pharmaceutique peut contenir un ou plusieurs principes actifs. Un principe actif peut être présent sous différentes formes ou formulations. Le RCP d'un médicament contenant de l'ésoméprazole indique par exemple : « chaque gélule contient 20 mg d'ésoméprazole (sous forme d'ésoméprazole magnésique dihydraté) ». D'autres formes d'ésoméprazole existent : "ésoméprazole sodique", "ésoméprazole magnésique" et "ésoméprazole magnésique trihydraté". On retrouve ces autres formes dans les médicaments génériques. A ce sujet, l'article L.5121-1 du Code de la santé publique [3] affirme que « les différents sels, esters, éthers, isomères, mélanges d'isomères, complexes ou dérivés d'un principe actif sont regardés comme ayant la même composition qualitative en principe actif, sauf s'ils présentent des propriétés sensiblement différentes au regard de la sécurité ou de l'efficacité ». Conceptuellement, les

3. https://www.legifrance.gouv.fr/affichTexteArticle.do?idArticle=JORFARTI000001857638&cidTexte=LEGITEXT000006055532&categorieLien=id

formes d'un principe actif correspondent au concept de "precise ingredient" de RxNorm et le principe actif au concept de "ingredient".

L'ANSM définit la substance active (SA) comme la partie d'un médicament possédant les propriétés thérapeutiques, par opposition aux excipients, et la fraction thérapeutique (FT) comme la partie de la substance active qui porte l'activité pharmacologique. Un exemple de relation entre spécialité pharmaceutique, présentations, fractions thérapeutiques et substances actives est donné sur la figure 3.5. Ici, la substance active est un sel (amoxicilline sodique) et la fraction thérapeutique son ingrédient (amoxicilline).

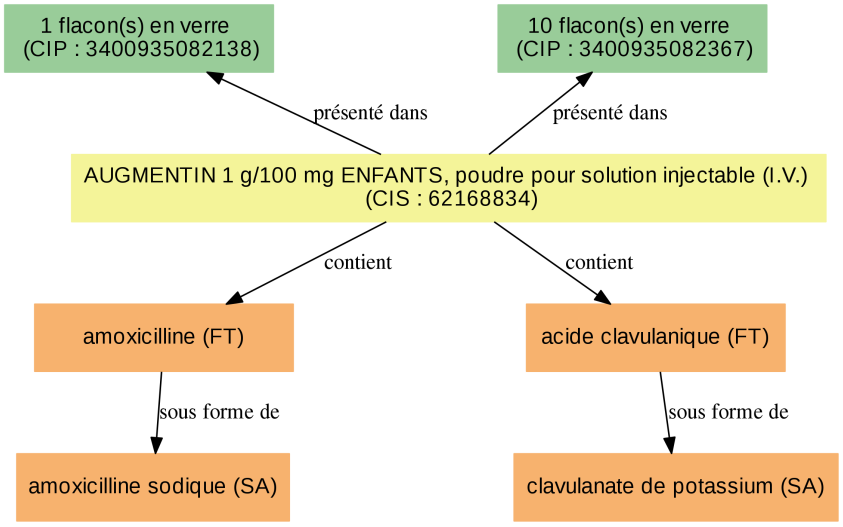

FIGURE 3.5 – Exemple de relations entre une spécialité pharmaceutique identifiée par un code CIS, ses présentations identifiées par un code CIP et ses fractions thérapeutiques (FT) sous forme de substance active (SA).

Il existe deux sources de données Open Data qui décrivent les substances des médicaments en vente en France : le répertoire des médicaments de l'ANSM [4] et la la base de données publique des médicaments [5]. Cette dernière est mise en œuvre par l'ANSM, en liaison avec la Haute Autorité de Santé et l'Union nationale des caisses d'assurance maladie (Uncam). Elle contient « des données et documents de référence sur les médicaments commercialisés ou ayant été commercialisés durant les trois dernières années en France ».

La base de données publique des médicaments est un sous-ensemble du répertoire des médicaments de l'ANSM qui est plus exhaustif, elle contient « toutes les spécialités ayant une autorisation en cours de validité, ainsi que pour les spécialités suspendues, retirées ou abrogées depuis la mise en ligne du répertoire ». Ces deux bases de données ne contiennent aucune donnée sur les ventes de médicaments en France.

4. http://ansm.sante.fr/Services/Repertoire-des-medicaments
5. http://base-donnees-publique.medicaments.gouv.fr

3.2.5 SNIIRAM

Le Système National d'Information Inter-Régime de l'Assurance Maladie (SNIIRAM) est la plus importante des bases publiques de données de santé en France, voire dans le monde [25]. Gérée par la Caisse nationale de l'assurance maladie (Cnamts), elle tire initialement ses données des feuilles de soins soit environ 1,2 milliard d'enregistrements par an. Les points forts de cette base sont l'exhaustivité de la population et la présence de données sur les consommations de soins en ville et à l'hôpital avec un chaînage (suivi longitudinal). Elle enregistre tous les médicaments achetés en pharmacie s'ils sont remboursés. Ces derniers sont identifiés par un code CIP dans la base. Son principal défaut est sa difficulté d'accès pour les chercheurs.

3.2.6 DP

Le Dossier Pharmaceutique (DP) recense, pour chaque bénéficiaire de l'assurance maladie qui le souhaite, tous les produits de santé (médicaments et dispositifs médicaux) délivrés au cours des quatre derniers mois, qu'ils soient prescrits par le médecin ou conseillés par le pharmacien (21 ans pour les vaccins, 3 ans pour les médicaments biologiques). En 2016, 34 millions de dossiers pharmaceutiques sont actifs, 99,8% des pharmacies d'officine l'utilisent ainsi que 9% des pharmacies à usage intérieur dans les établissements de santé [6]. Sa mise en œuvre a été confiée au Conseil National de l'Ordre des Pharmaciens (CNOP) par la loi du 30 janvier 2007 relative à l'organisation de certaines professions de santé.

3.2.7 Medic'AM

Medic'AM est une base de données ouverte sur les remboursements de l'assurance maladie. Elle est hébergée sur le site data.gouv.fr [7] qui est une plateforme d'Open Data « mettant à disposition de tous des données publiques pour permettre une réutilisation de ces données au-delà de leur utilisation première administrative ». Toutes les données sont extraites du SNIIRAM. La base contient pour chaque code CIP la base de remboursement, le montant remboursé et le nombre de boîtes délivrées par mois. Elle ne contient aucune donnée sur les co-prescriptions.

3.2.8 ATC

La classification ATC (Anatomique, Thérapeutique et Chimique) a été établie par l'Organisation Mondiale de la Santé (OMS). Elle est recommandée et utilisée internationalement dans les études pharmacologiques. C'est une classification hiérarchique monoaxiale composée

6. http://www.ordre.pharmacien.fr/Le-Dossier-Pharmaceutique/Le-DP-raconte-par-Isabelle-Adenot
7. https://www.data.gouv.fr/fr/datasets/medicaments-rembourses-par-lassurance-maladie/

de 14 axes identifiés par une lettre majuscule correspondant à des groupes anatomiques. La hiérarchie comprend cinq niveaux :

— 1er niveau : groupes anatomiques
  A : Alimentary tract and metabolism
— 2ème niveau : sous-groupe thérapeutique
  A10 : Antidiabétiques
— 3ème niveau : sous-groupe pharmacologique
  A10B : Antidiabétiques oraux
— 4ème niveau : groupes chimiques
  A10BA : Biguanides
— 5ème niveau : nom de la molécule en DCI
  A10BA0 : Metformine

Un code ATC est constitué de 1 à 7 charactères alphanumériques selon le niveau.

La classification est utilisée pour classer des médicaments. Comme certaines molécules sont présentes seulement en combinaison avec d'autres, elles n'ont pas d'identifiant individuel au 5ème niveau. C'est le cas de l'acide clavulanique par exemple. Comme une molécule peut avoir plusieurs indications, elle peut avoir plusieurs codes ATC. Par exemple, la bromocriptine est utilisée dans le traitement de la maladie de Parkinson (N04BC01) et des hyperprolactinémies physiologiques ou pathologiques (G02CB01). Bien qu'il s'agisse de la même molécule, il n'existe pas de relation entre ces deux codes dans la hiérarchie. Comme les indications varient selon les autorisations, une molécule peut avoir des codes ATC différents selon les pays.

3.3 Sources de données sur les interactions

3.3.1 Thesaurus de l'ANSM

D'après le site de l'ANSM : « L'ANSM met à la disposition des professionnels de santé l'ensemble des interactions médicamenteuses identifiées par le Groupe de Travail ad hoc et regroupées dans un thesaurus. Ce thesaurus apporte aux professionnels de santé une information de référence, à la fois fiable et pragmatique, avec des libellés volontairement simples utilisant des mots clés. Il doit être utilisé comme un guide pharmaco-thérapeutique d'aide à la prescription ». Les membres du groupe de travail sur les interactions médicamenteuses, les ordres du jour et les comptes-rendus de réunion sont consultables sur le site de l'ANSM [8]. La décision d'inclure une nouvelle interaction ou de modifier un niveau de contrainte est une décision collégiable après revue des données disponibles.

En France, les éditeurs de logiciels d'aide à la prescription (LAP) et de logiciels d'aide à la dispensation (LAD) certifiés par la HAS doivent utiliser une base de données sur les

8. http://ansm.sante.fr/L-ANSM2/Groupes-de-travail/Groupes-de-travail-d-expertise/Groupes-de-travail-d-expertise/Groupe-de-travail-Interactions-medicamenteuses

médicaments agréées par la HAS [26]. Toutes les bases de données agréées à ce jour (Vidal, Thériaque, Thésorimed, Base Claude Bernard, Clickadoc) disent utiliser le thesaurus des interactions médicamenteuses édité par l'ANSM dans leurs réponses au questionnaire d'évaluation. Ce référentiel national est donc probablement largement utilisé par les professionnels de santé pour la prescription et pour la dispensation des médicaments via leur LAP ou LAD.

D'après les informations sur le thesaurus, « pour être retenue, une interaction doit avoir une traduction clinique significative, décrite ou potentiellement grave, c'est-à-dire susceptible de provoquer ou majorer des effets indésirables, ou d'entraîner, par réduction de l'activité, une moindre efficacité des traitements ».

Les interactions sont décrites selon quatre niveaux de contrainte :

— Contre-indication
  La contre-indication revêt un caractère absolu.
— Association déconseillée
  L'association doit être le plus souvent évitée sauf après un examen approfondi du rapport bénéfice/risque.
— Précaution d'emploi
  L'association est possible dès lors que sont respectées des recommandations simples pour éviter la survenue de l'interaction.
— A prendre en compte
  Le risque d'interaction médicamenteuse existe. Il s'agit le plus souvent d'une addition d'effets indésirables. Aucune recommandation ne peut être proposée.

Le thesaurus est mis à jour deux fois par an. D'après les informations générales du thesaurus, « l'information exhaustive sur les interactions médicamenteuses d'une spécialité pharmaceutique donnée repose sur la consultation du thesaurus ainsi que le résumé des caractéristiques du produit de cette spécialité ».

3.3.2 RCP

Le résumé des caractéristiques du produit (RCP) est la partie de l'AMM (l'annexe 1) destinée aux professionnels de santé. Ils sont accessibles en ligne au format html ou PDF lorsque l'AMM est issue d'une procédure centralisée européenne. Le format d'un RCP est semi-structuré : plusieurs balises identifient les différentes sections (indications thérapeutiques, posologie et mode d'administration ...) dont le contenu est en texte libre. Une section du RCP concerne les interactions médicamenteuses.

Les RCP sont hébergés sur le site de la base de données publiques des médicaments. Par exemple, le RCP de la spécialité "MOTILIUM 10 mg, comprimé pelliculé" est accessible à cette URL : http://base-donnees-publique.medicaments.gouv.fr/affichageDoc.php?specid=63679194&typedoc=R. On remarque la présence du code CIS (63679194) de la spécialité pharmaceutique dans la barre d'adresse. L'extraction automatique des informations

contenues dans les RCP est un sujet de recherche. Rubrichi et al. [27] ont obtenu de très bons résultats pour l'extraction des interactions médicamenteuses sur des RCP italiens qui suivent le même format.

3.3.3 Sources internationales

A l'échelle internationale, il n'existe pas de consensus ni une source de référence sur les interactions médicateuses[8][9][28]. Deux travaux récents [8][9] ont cherché à regrouper les interactions de plusieurs sources. L'idée est que l'existence d'une interaction est d'autant plus probable que celle-ci est décrite dans plusieurs sources différentes.

Les sources peuvent être publiques ou propriétaires. Elles peuvent être dans un format structuré comme une base de données ou non structuré en texte libre. Ayvaz et al.[8] ont mené une revue bibliographique des ressources disponibles sur les interactions médicamenteuses puis regroupé ces sources dans une même base : Merged-PDDI. Ils ont identifié 14 sources publiquement accessibles qu'ils ont classé en 3 catégories :

- Sources pour applications cliniques
  - CredibleMeds [9]
    L'organisation se décrit comme « un organisme de recherche indépendant qui a pour mission d'améliorer les outils thérapeutiques et de diminuer les évènements indésirables des médicaments ». Elle fournit une liste de médicaments allongeant le QT ou à risque de torsades de pointes.
  - National Drug File - Reference Terminology (NDF-RT)
    En plus de fournir des informations sur les médicaments vendus aux Etats-Unis, NDF-RT contenait des informations sur les interactions médicamenteuses. L'éditeur, le département des anciens combattants des Etats-Unis (Veterans Affairs - VA), a cessé de maintenir sa liste en novembre 2014[29]. La terminologie est présente dans RxNorm.
  - ONC
    Le bureau de la coordination nationale pour les technologies de l'information en santé (ONC) a publié une liste portant sur les molécules contre-indiquées en 2012 : ONC High Priority[30] et une autre sur les interactions moins graves cliniquement en 2013 : ONC Non-interruptive [12].
  - OSCAR
    Une liste d'interaction médicamenteuse créée par consensus d'expert en 1997 [31] intégrée dans le logiciel du même nom.
- Sources développées pour la détection automatique d'interactions
  Des experts humains ont manuellement étiquetté les interactions dans différents documents en texte libre (notice du produit...) pour évaluer des méthodes automatiques

9. https://crediblemeds.org/

d'extraction. Quatre sources ont été identifiées dans cette catégorie.

- Sources développées pour la pharmacovigilance ou pour des applications bioinformatiques.
  - KEGG (Kyoto Encyclopedia of Genes and Genomes)
    Cette source contient une liste d'interactions médicamenteuses générée par traitement automatique de la langue sur les notices des médicaments commercialisés au Japon [32].
  - TWOSIDES
    Il s'agit d'une liste d'interactions produite par fouille de données sur des effets indésirables spontanément rapportés [33]. TWOSIDES fournit un niveau de certitude sur l'existence de l'interaction.
  - DrugBank
    L'encyclopédie fournit une liste avec une brève description de l'interaction. Par exemple, DrugBank affirme que la simvastatine interagit avec le vérapamil avec la description suivante : "The serum concentration of Simvastatin can be increased when it is combined with Verapamil".
  - DIKB (Drug Interaction Knowledge Base)
    DIKB est un système de représentation des connaissances sur les molécules, leurs métabolites et leurs récepteurs pour prédire les interactions médicamenteuses [34].
  - SemMedDB
    SemMedDB [35] est une base de données de triplets RDF générée par traitement automatique de la langue sur les articles de la base MEDLINE. Ayvaz et al. ont extrait les triplets où le sujet et l'objet sont un médicament et la relation est "interagit avec".

La base Merged-PDDI utilise les identifiants DrugBank car l'encyclopédie contient de nombreux liens vers d'autres sources ce qui facilite grandement les alignements à réaliser entre toutes ces sources. Cette base est sous licence libre. Un service web en ligne permet de rechercher des interactions dans cette base : https://www.dikb.org/Merged-PDDI/.

D'autres sources internationales existent mais ne sont pas librement accessibles. Quelques exemples sont :

- Multum
  Multum est une source commerciale propriétaire accessible gratuitement en ligne [10]. Elle est utilisée comme référence dans plusieurs articles [30][28]. Elle fournit une description de l'interaction et trois niveaux de sévérité : mineur, modéré et majeur.
- Référentiel des Pays-Bas
  Aux Pays-Bas, il existe un référentiel national [7] comme en France. Il est maintenu par un groupe de travail composé de 22 experts. Les informations sur les interactions

10. https://www.drugs.com/

sont directement intégrées dans une base de données nationale sur les médicaments appelée G-Standaard [11].

— Le Stockley
Le Stockley est un référentiel utilisé au Royaume-Uni.

3.4 Méthodes d'alignement

Comme le thesaurus des interactions ne contient pas d'autre information que des libellés pour représenter des molécules, les possibilités de liaison sont limitées à ces derniers. Dans un livre [36], Dusetzina et al. expliquent les méthodes et les difficultés de lier des données par des termes. Une phase de normalisation est souvent nécessaire pour retirer les accents, les tirets, les parenthèses et mettre en minuscule. Il existe deux grands types d'algorithme pour lier des données : déterministe et probabiliste.

La méthode déterministe consiste à rechercher un alignement parfait ou partiel entre deux chaînes de caractères : "abciximab" et "abciximab (c7E3 Fab)" est un exemple d'alignement partiel où la première chaîne de caractères est incluse dans la deuxième. La méthode déterministe peut utiliser des techniques de correction de fautes d'orthographe. Il en existe deux principales. La première se base sur la phonétique, l'algorithme Soundex est le plus largement connu des algorithmes phonétiques [12]. Par exemple, "acide salicylique" et "acid salicylic" se prononcent de la même façon d'après l'algorithme Soundex. La deuxième se base sur une mesure de similarité entre deux chaînes de caractères. Une mesure bien connue est la distance de Levenshtein qui est égale au nombre minimal de caractères qu'il faut supprimer, insérer ou remplacer pour passer d'une chaîne à l'autre. Par exemple, la distance de Levenshtein entre racecadotril et racecadodril est de 1 car il n'existe qu'un remplacement de lettre à réaliser entre les deux chaînes.

La méthode probabiliste repose sur des algorithmes qui permettent d'ordonner les alignements selon leur probabilité. Il est par exemple plus probable que "liothyronine sodique" soit lié à "liothyronine" plutôt qu'à "catioresine sulfo sodique" car "liothyronine" est une unité lexicale rare tandis que "sodique" est beaucoup plus fréquente. Cette mesure de rareté peut se calculer avec l'IDF (Inverse Document Frequency), une mesure bien connue des moteurs de recherche. Elle est calculée pour chaque unité lexicale dans un terme. Elle est maximale si l'unité lexicale apparaît une seule fois et décroît proportionnellement à sa fréquence d'apparition.

11. https://www.z-index.nl/english
12. https://fr.wikipedia.org/wiki/Soundex

Matériel

4.1 Données sur les interactions

4.1.1 Thesaurus des interactions de l'ANSM

Le référentiel national est publié deux fois par an. Dans sa dernière version (janvier 2016) au moment de l'écriture de ce travail, le fichier PDF contenait 226 pages. Il est accompagné de deux autres fichiers : l'index des substances qui liste les classes thérapeutiques de chaque substance le cas échéant et indique si une substance possède des interactions en propre ; l'index des classes thérapeutiques qui liste les molécules de chaque classe thérapeutique. Les classes "thérapeutiques" définies dans le thesaurus sont des classes pharmacologiques propres au thesaurus qui regroupent un ensemble de molécules qui possèdent des mécanismes d'interaction similaires.

Le manuel d'utilisation sur le site explique l'organisation du fichier : « l'interaction est définie par un couple de protagonistes 'a + b' qui peuvent être :

— une substance active, désignée par sa dénomination commune internationale (DCI)

— une classe thérapeutique, elle-même faisant l'objet d'interactions "de classe"

Le premier protagoniste de l'interaction ('a') apparaît en grisé dans le thesaurus. Les protagonistes 'b' sont ensuite déclinés, précédés d'un signe '+'.»

Un exemple est donné sur la figure 4.1 entre les immunosuppresseurs et les inducteurs enzymatiques. On remarque que la liste des molécules figure pour le protagoniste 'a' mais pas pour le protagoniste 'b'. Comme les couples d'interaction sont présents en miroir, il faut se rendre à la page du thesaurus où la classe des inducteurs enzymatiques est le protagoniste 'a' pour voir la liste des molécules de cette classe ou consulter le fichier des classes thérapeutiques.

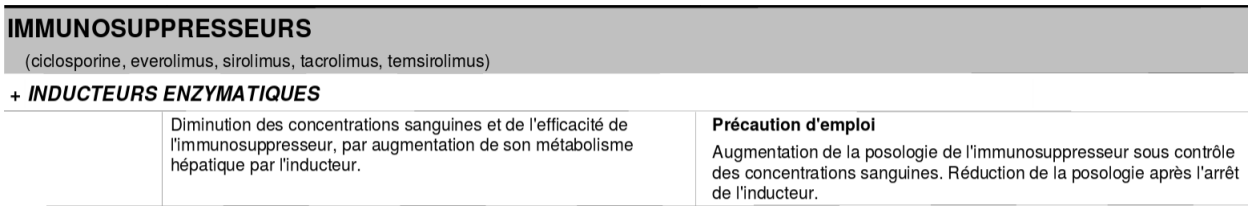

FIGURE 4.1 – Exemple d'une interaction entre deux classes thérapeutiques dans le thesaurus de l'ANSM

La relation d'interaction est dite symétrique : si un protagoniste 'a' interagit avec le protagoniste 'b' alors 'b' interagit avec 'a'.

La figure 4.2 montre les liens directs ou indirects d'interaction entre deux molécules. On dira que le lien est direct si la molécule est un protagoniste, indirect sinon.

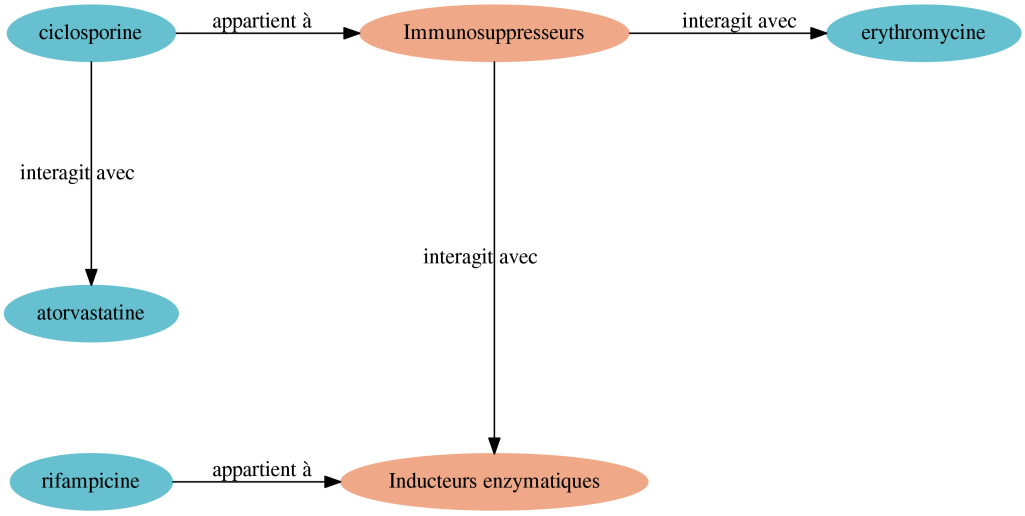

FIGURE 4.2 – Représentation de quelques interactions avec la ciclosporine d'après le thesaurus de l'ANSM. Le lien d'interaction entre deux molécules peut être direct, comme pour le couple ciclosporine-atorvastatine, ou indirect comme avec les couples ciclosporine-rifampicine et ciclosporine-érythromycine.

Ce format de stockage des données rend difficile la recherche d'information pour un utilisateur. Par exemple, la ciclosporine appartient à quatre classes dans le thesaurus et possède aussi des interactions en propre. Il faut donc consulter la liste des interactions de chacun de ces protagonistes pour connaître les interactions de la ciclosporine. Il n'existe pas d'ancre (bookmark) qui permettrait de naviguer facilement dans le document. Au total, la ciclosporine est en interaction directe ou indirecte avec 256 molécules. Ce dénombrement est très chronophage à réaliser manuellement.

La recherche d'information peut aussi être trompeuse pour un utilisateur non alerté qui rechercherait une interaction de classe thérapeutique. Par exemple, à la page 79 du thesaurus, il est noté que les diurétiques de l'anse (bumétanide, furosémide, pirétanide, torasémide) interagissent avec seulement deux molécules et deux classes thérapeutiques. Or, toutes les molécules de cette famille appartiennent aussi à la famille des diurétiques. On peut donc inférer que la famille des diurétiques de l'anse est une sous-famille des diurétiques et qu'elle hérite de ses interactions. Comme la figure 4.3 le montre, il est possible d'inférer (relations en pointillé) que les diurétiques de l'anse interagissent avec les AINS, information non présente dans le thesaurus. Il existe des relations de subsomption entre les classes thérapeutiques du thesaurus intéressantes à identifier.

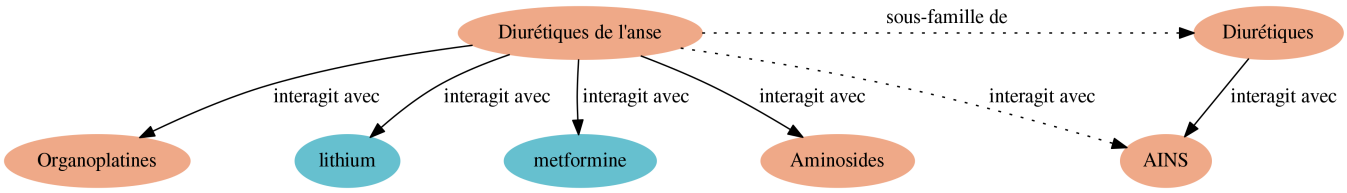

FIGURE 4.3 – En trait plein figurent les relations présentes dans le thesaurus, en pointillé les relations qu'il est possible d'inférer. Comme toutes les molécules des diurétiques de l'anse appartiennent aux diurétiques, il est possible d'inférer que la première est sous-famille de la deuxième et que les diurétiques de l'anse interagissent aussi avec les AINS

4.1.2 Merged-PDDI

Ayvaz et al. [8] ont rendu disponible leur travail multi-sources sur les interactions médicamenteuses sur la plateforme GitHub [1]. Ils fournissent un fichier CSV de leur travail d'intégration téléchargeable [2]. Chaque ligne du fichier contient notamment un couple de molécules identifiées par des identifiants DrugBank et la source. Trois lignes sont présentées dans le tableau 4.1 à titre d'exemple. Le fichier contient 27 457 couples uniques à risque d'interaction et 12 sources différentes (figure 4.4). Les sources KEGG et TWOSIDES n'apparaissent pas dans ce fichier car la méthode d'intégration est moins fiable que pour les autres sources [8]. Les sources ONC high-priority, ONC non-interruptive, OSCAR, CredibleMeds, DIKB et DrugBank ont été intégrées entièrement tandis que pour les autres sources l'intégration est incomplète.

id1	id2	source	molécule 1	molécule 2
DB01175	DB00338	DIKB	escitalopram	omeprazole
DB01175	DB01238	DIKB	escitalopram	aripiprazole
DB01175	DB00501	DIKB	escitalopram	cimetidine

Tableau 4.1 – Ayvaz et al. fournissent un fichier de leur travail d'intégration contenant le couple des molécules à risque identifiées par un code DrugBank et la source. Trois lignes sont données à titre d'exemple. id : identifiant DrugBank.

1. https://github.com/dbmi-pitt/public-PDDI-analysis
2. http://purl.org/net/drug-interaction-knowledge-base/PDDI-data-merged-non-conservative

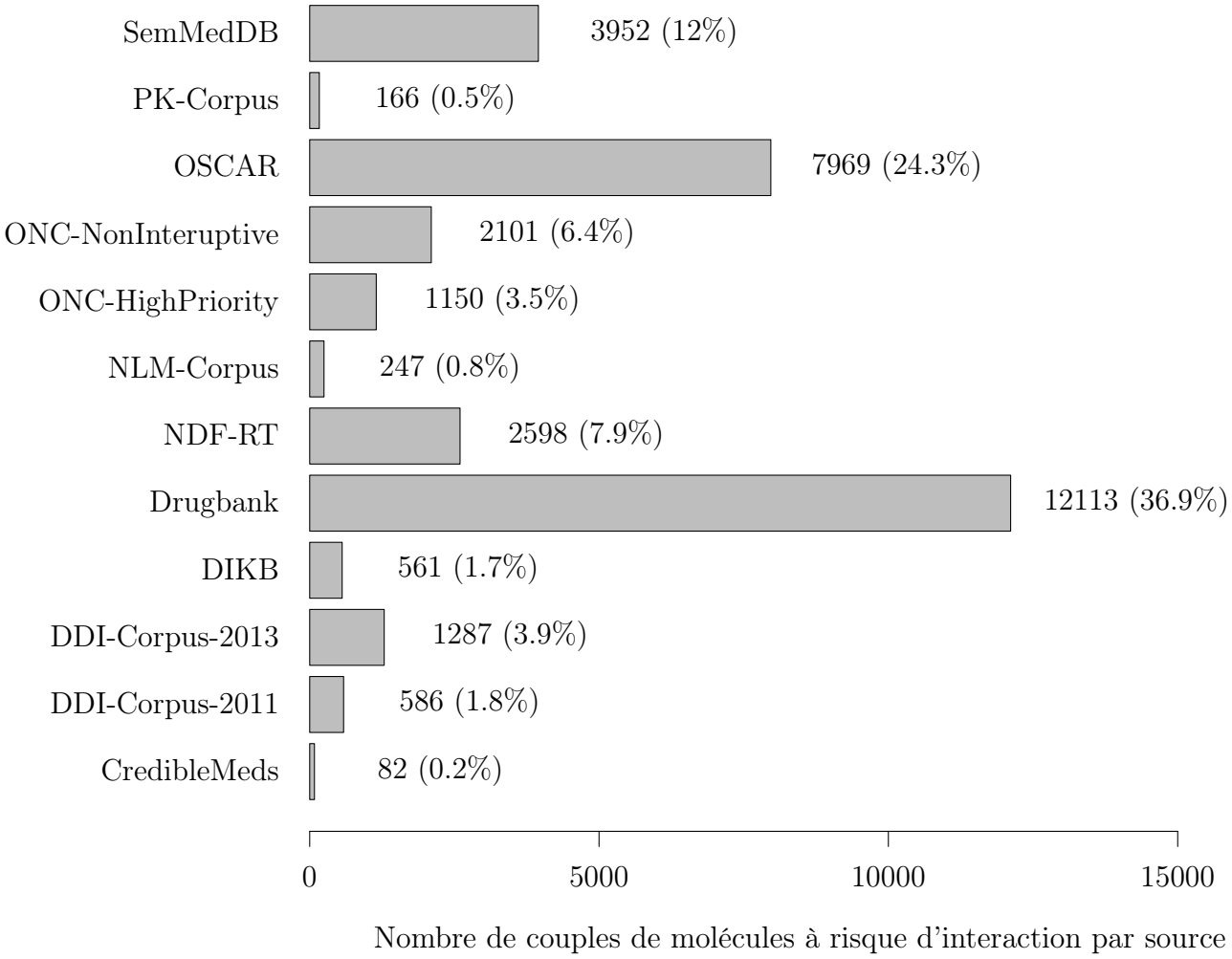

FIGURE 4.4 – Nombre de couples à risque d'interaction par source dans la base Merged-PDDI

4.2 Données sur les médicaments

4.2.1 UMLS

Après acceptation de la licence d'utilisation, l'utilisateur peut accéder gratuitement au contenu de l'UMLS via un navigateur web ou télécharger les fichiers pour manipuler le meta-thesaurus localement. Cette dernière option nécessite de disposer de 30 gigaoctets d'espace puis le chargement dans une base de données pour réaliser des requêtes. Une interface de programmation (API) est mise à disposition des développeurs pour accéder à cette ressource

via le langage de programmation JAVA [3]. Elle permet de réaliser des recherches automatiquement, par exemple pour rechercher des concepts UMLS à partir d'une liste de termes. Cette API a été utilisée pour les alignements vers l'UMLS et le navigateur web a été utilisé pour effectuer des recherches manuelles.

4.2.2 DrugBank

Les informations de la base de données DrugBank peuvent être accessibles en ligne avec un navigateur web [4] sans inscription. Après inscription gratuite en ligne, l'intégralité de la base de données peut être téléchargée au format XML. Le site fournit aussi d'autres fichiers comme un fichier CSV (all_drugbank_vocabulary.csv) contenant le vocabulaire utilisé par Drugbank pour faciliter les alignements vers d'autres sources. Le fichier vocabulaire de DrugBank contient 8221 identifiants de molécule et 21 710 libellés préférés ou synonymes uniques.

4.2.3 Répertoire du médicament

Dans sa volonté de « mettre à disposition du grand public et des professionnels de santé les informations officielles sur le médicament » l'ANSM propose trois fichiers au format CSV à télécharger sur son site [5].

— le fichier des présentations
  Il contient, pour chaque code CIS, les codes CIP (à 7 et 13 chiffres), le libellé de la présentation (ex : plaquette thermoformée PVC aluminium de 30 comprimés), la date de commercialisation, le statut de la présentation (active ou abrogée) et l'état de la commercialisation (déclaration de commercialisation, déclaration d'arrêt de commercialisation...).

— le fichier des spécialités
  Il contient, pour chaque code CIS, la dénomination commerciale du médicament, sa forme galénique (ex : comprimé), sa voie d'administration, la date et le type de procédure de l'AMM (nationale, centralisée ...).

— le fichier des compositions
  Il contient, pour chaque code CIS, les substances et leur dosage.

Cette base de données contient 118 532 codes CIP13 tandis que la base de données publique du médicament, un sous-ensemble, contient 19 441 codes CIP13.

Les substances des spécialités pharmaceutiques sont décrites dans le fichier composition qui contient 4684 codes et 5804 libellés différents. La colonne "nature" de la table composition indique si la substance est une substance active (SA) ou une fraction thérapeutique (FT).

3. https://documentation.uts.nlm.nih.gov/soap/home.html
4. http://www.drugbank.ca/
5. Toutes les variables n'ont pas été décrites ici, pour plus d'information veuillez consulter le site

Plusieurs libellés peuvent avoir le même code. Par exemple, "amoxicilline", "amoxicilline anhydre" et "amoxicilline base" sont identifiés par le code 5248. Les codes n'ont pas de signification.

4.2.4 Extrait du dossier pharmaceutique

Au Hacking Health Camp de Strasbourg [6], l'ordre des pharmaciens a fourni en Open Data une extraction anonymisée contenant toutes les délivrances enregistrées dans huit départements français durant les mois de juin, juillet et août 2013. Cet extrait représentait environ 11% de l'activité nationale sur cette période.

Les fichiers fournis étaient au format CSV et contenaient notamment :

— **date** La date de dispensation

— **num** Le numéro de dispensation
Il est unique à chaque délivrance. Le chaînage des numéros n'était pas fourni, il était impossible de rattacher plusieurs délivrances à un même patient.

— **CIP** Les codes CIP7 des produits délivrés pour chaque délivrance

— **Age** L'âge du patient en années au moment de la prescription.

Par la suite, le mot "délivrance" sera employé pour désigner tous les médicaments possédant le même numéro de dispensation. Le tableau 4.2 montre un exemple de délivrance.

date	num	age	CIP
01/06/2013	239972491	72	3999903
01/06/2013	239972491	72	2174023
01/06/2013	239972491	72	3474419
01/06/2013	239972491	72	3149240
01/06/2013	239972491	72	3990084

Tableau 4.2 – Exemple de délivrance médicamenteuse dans les données fournies par l'ordre des pharmaciens.

Ces données ont été utilisées pour notre deuxième exemple. Les interactions médicamenteuses seront recherchées dans ces données.

4.3 Outils informatiques

R est un langage de programmation et un logiciel libre de traitement des données et d'analyse statistique[37]. Il incorpore des fonctionnalités de base qui sont extensibles par l'ajout d'extensions, appelés paquets (packages). Développés par des développeurs/utilisateurs pour

6. 2ème édition du Hacking Health Camp à Strasbourg. 19 au 22 mars 2015

résoudre des problèmes spécifiques, ces paquets sont mis librement à disposition de la communauté des utilisateurs. Dans le cadre de ce travail, les paquets suivants ont été utilisés : SPARQL[38], stringr[39], tikzDevice[40], devtools[41] et roxygen2[42]. La programmation en R version 3.2.3, ainsi que la rédaction de ce travail en LaTeX, a été réalisée dans l'environnement de développement RStudio[43] version 0.99.491.

Java™ est un langage de programmation informatique. La programmation en Java version 1.8.0_101 a été réalisée dans l'environnement de développement Eclipse version 4.5.1.

Lucene™ est une bibliothèque JAVA utilisée pour indexer et rechercher des documents textuels. Elle est développée par la fondation Apache, communauté décentralisée de développeurs, auteur de nombreux projets open source. La version 4.7.0 a été utilisée.

Tika™ est une boîte à outils JAVA développée par la fondation Apache qui permet d'extraire le contenu et les métadonnées de différents types de format (PDF, PPT, DOC...). Tika™1.13 a été utilisée pour transformer le format PDF vers le format texte.

Jena™ est une bibliothèque Java pour le développement d'applications du web sémantique. Elle intègre un triplestore (TDB) et un SPARQL endpoint (Fuseki). Elle est développée par la fondation Apache. Sa version 2.4.0 a été utilisée pour manipuler les triplets RDF.

Protégé™ est un logiciel conçu pour le développement d'ontologies. Il a été créé à l'université de Stanford. Développé en Java, il est gratuit et open source.

HermiT HermiT est un raisonneur OWL [44]. Il est intégré à Protégé™mais peut aussi être utilisé de manière autonome. HermiT 1.3.8 a été utilisé.

GraphViz™ est un ensemble d'outils open source pour la création de graphes. De nombreux graphiques ont été réalisés avec sa version 2.36.0.

Git™ est un logiciel de gestion de versions. Sa version 1.9.1 a été utilisée. GitHub est un site hébergeant des projets informatiques qui utilisent Git.

Avakas est un cluster du Mésocentre de Calcul Intensif Aquitain (MCIA) hébergé dans les locaux de la Direction Informatique de l'Université de Bordeaux [45]. Tout personnel affecté à une structure partenaire du mésocentre peut demander un accès au supercalculateur. Le cluster dispose de 264 nœuds de calcul composés de 2x6 cœurs (Intel Xeon x5675 @3,06 GHz) par nœud et 48 Go de mémoire RAM ; soit un total de 3168 cœurs.

Méthodes

5.1 Extraction des données du PDF

Le fichier PDF du thesaurus des interactions médicamenteuses de l'ANSM a été transformé au format texte par l'outil Tika™. Un script R a ensuite transformé ce fichier texte en un fichier structuré au format CSV qui à son tour a été transformé en triplets RDF. Pour vérifier l'exactitude des données transformées, l'index des substances et l'index des classes thérapeutiques, deux autres fichiers PDF fournis par l'ANSM, ont été utilisés.

Une petite ontologie OWL a été créée avec Protégé™. L'objectif de cette ontologie était d'identifier automatiquement les relations de subsomption entre les classes thérapeutiques et d'inférer les relations d'interaction. Nous souhaitions par exemple inférer que la famille des diurétiques contient la famille des diurétiques de l'anse et que la famille des diurétiques de l'anse interagit avec les anti-inflammatoires non stéroïdiens (figure 4.3). Cette ontologie n'est pas indispensable à la création d'un format ouvert du thesaurus mais permet d'expliquer à une machine les raisonnements qu'un être humain réalise sur ces données. Les détails de son implémentation sont décrits en annexe.

5.1.1 Structure des classes du thesaurus

Relation subClassOf

Les logiques de description, utilisées par le langage de connaissance OWL, emploient la théorie des ensembles. Dans notre cas, une classe thérapeutique peut être représentée comme un ensemble de molécules. Une classe thérapeutique A est un sous ensemble de B ($A \subset B$, figure 5.1) si toutes les molécules de A sont présentes dans B. Un raisonneur utilise cette théorie pour déduire des relations hiérarchiques ('rdfs:subClassOf') entre les classes thérapeutiques du thesaurus. La subsomption est une notion équivalente à la relation "contient" en logique ensembliste.

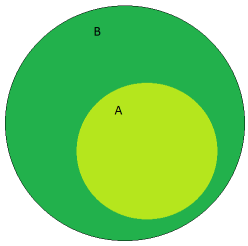

FIGURE 5.1 – $A \subset B$

Exception subClassOf

La relation 'rdfs:subClassOf' est établie entre deux classes A et B si et seulement si l'ensemble des molécules de la classe A est compris dans l'ensemble de la classe B. Il est intéressant d'étudier les cas où cette relation n'est pas établie à cause d'une (ou quelques) molécule(s) de la classe A. En effet, cette molécule a possiblement des propriétés que les autres molécules de sa classe n'ont pas car elle est la seule à ne pas appartenir à la classe B. Ceci permet aussi de vérifier que cette molécule n'a pas été oubliée lors de la classification. On peut dire dans ces cas que A est 'rdfs:subClassOf' de B à l'exception de cette (ou quelques) molécule(s).

La différence ensembliste de A et B ($A \setminus B$) est l'ensemble des éléments de A qui n'appartiennent pas à B. Dans la relation 'rdfs:subClassOf', cet ensemble est vide : $A \setminus B = \emptyset$. Les couples de classes recherchés sont ceux qui ont une différence ensembliste très petite par rapport à l'intersection des deux classes (figure 5.2).

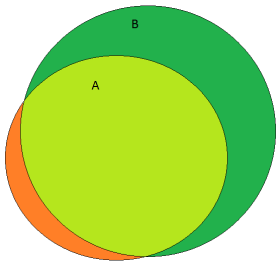

FIGURE 5.2 – La différence ensembliste de A et B ($A \setminus B$), en orange, est petite par rapport à l'intersection des deux classes ($A \cap B$) en jaune

Pour chaque couple de classes possible, un score de subClassOf est calculé d'après la formule 5.1.

$$\text{Score subClassOf}\,(A, B) = \frac{|A \setminus B|}{|A \cap B|} \tag{5.1}$$

Le score est égal à la cardinalité de la différence ensembliste divisée par la cardinalité de l'intersection des deux classes. Ce score est calculé uniquement si l'intersection des deux

classes n'est pas l'ensemble vide ($A \cap B = \emptyset$), c'est-à-dire si les deux classes partagent au moins une molécule. Le score est égal à 0 si A est subClassOf de B et il est d'autant plus petit que le nombre de molécules en commun est grand et que le nombre de molécules de différence est petit. Le langage R a été utilisé pour réaliser ces calculs.

Bien qu'il eût été possible d'utiliser cette formule pour connaître les relations 'rdfs:subClassOf' de la section précédente, la solution du raisonneur est plus élégante car elle utilise la représentation des connaissances. La machine est capable d'inférer par elle-même les relations 'rdfs:subClassOf' sans qu'un être humain n'aît à lui transmettre ces informations.

5.2 Alignement des molécules du thesaurus

5.2.1 vers l'UMLS

L'objectif était de rattacher chaque libellé de molécule du thesaurus à un concept UMLS. L'API fournie par la NLM a été utilisée pour rechercher un concept à partir d'un terme. Le libellé était recherché parmi les 3,25 millions de termes que l'UMLS contient. Un programme JAVA a été développé pour interroger le serveur de la NLM via l'API et a été configuré pour recevoir un seul concept, le plus spécifique trouvé. Si le libellé du thesaurus correspondait exactement au libellé d'un terme de l'UMLS alors le programme assignait automatiquement l'identifiant du concept à la molécule du thesaurus. Dans les autres cas où les libellés ne correspondaient pas exactement ou le terme n'était pas trouvé, une recherche manuelle était réalisée. Un alignement parfait entre deux sources de données correspond à la relation "owl:sameAs" en RDF. Comme RxNorm est contenu dans l'UMLS, l'API est capable de renvoyer le concept RxNorm, s'il existe, à partir d'un concept UMLS. Il a ainsi été possible de savoir combien de molécules du thesaurus étaient présentes dans RxNorm.

5.2.2 vers Drugbank

A notre connaissance, il n'existe pas de source officielle faisant le lien entre DrugBank et l'UMLS. L'objectif était de rattacher chaque libellé de molécule du thesaurus à un identifiant DrugBank.

Afin de ne pas être limité aux termes du thesaurus, l'étape précédente a été utilisée pour accomplir cette tâche. Tous les termes en anglais et en français rattachés au même concept UMLS d'une molécule du thesaurus ont été récupérés avec l'API fournie par la NLM. Ces termes sont des synonymes car ils sont liés au même concept. Notre vocabulaire "source" contenait tous les termes du thesaurus des interactions avec les libellés préférés et les synonymes des concepts UMLS. Cet ensemble de termes a été utilisé pour rechercher un identifiant dans DrugBank. Chaque identifiant DrugBank contient un libellé préféré et aucun ou plusieurs synonymes (vocabulaire "cible").

Pour prendre en compte les possibles variations lexicales, les termes ont été dans un premier temps normalisés et séparés en unités lexicales élémentaires (tokénisation). L'étape de normalisation a consisté à mettre les termes en minuscule, à retirer les accents et les caractères spéciaux comme les parenthèses. Si un terme comportait plusieurs mots, la tokénisation permettait de les séparer. Les termes étaient ensuite indexés dans un index Lucene™pour faciliter leur recherche. Le tableau 5.1 montre un exemple de ce procédé.

Les alignement vers DrugBank ont été réalisés dans l'ordre suivant :

— Alignement parfait sur les libellés préférés

Si le libellé préféré de DrugBank était identifique au libellé du thesaurus ou au libellé préféré du concept UMLS, alors l'alignement était automatique et aucune vérification manuelle n'était réalisée. L'hypothèse faite était qu'il n'existait pas de termes homonymes parmi ces libellés.

— Alignement parfait sur les syonymes

Si un syonyme de DrugBank était identique à un synonyme de l'UMLS alors l'alignement était semi-automatique et une validation manuelle était réalisée. Comme les synonymes contenaient parfois des sigles, par exemple 'ATP', des homonymes étaient possibles et une vérification manuelle était donc nécessaire.

— Alignement imparfait

Un alignement était proposé si deux libellés, préférés ou synonymes, partageaient une unité lexicale commune. Par exemple, le terme "abciximab (c 7e3b fab)" issu du thesaurus et "abciximab" de Drugbank partagent l'unité lexiale "abciximab". Lucene™renvoyait une valeur numérique d'autant plus élevée que l'unité lexicale commune était rare dans les autres termes, ce qui donnait un indice de la probabilité de l'alignement. Une validation manuelle était ensuite réalisée.

Lorsque le terme n'était pas trouvé, une mesure de similarité était calculée entre deux chaînes de charactères. La fonction utilisée, search_fuzzy de Lucene™, utilise la distance de Levenshtein. Un alignement était proposé si Lucene™trouvait une unité lexicale ou un terme commun distant d'un caractère. Par exemple, le terme "esclicarbazepine" du thesaurus est distant d'un caractère (ajout d'un c) de l'unité lexicale "eslicarbazepine" dans le terme "eslicarbazepine acetate" de Drugbank.

— Recherche manuelle

Enfin, si aucune des étapes précédentes n'était fructueuse, une recherche manuelle était réalisée.

Les validations manuelles consistaient à dire si les deux termes représentaient le même concept ("owl:sameAs"), s'il existait une relation de forme : "RxNorm:has_form" ou "RxNorm:is_form" entre les deux termes, s'il s'agissait d'une autre relation ou s'il n'existait aucun lien.

Terme	Terme normalisé	Terme tokenizé
Sodium (chlorure de)	sodium chlorure de	sodium/chlorure/de

Tableau 5.1 – Exemple de normalisation et tokénisation avant indexation du terme dans Lucene™

5.2.3 vers le répertoire du médicament

L'objectif était de rattacher les molécules du thesaurus des interactions de l'ANSM vers les substances décrites dans la table composition du répertoire du médicament de l'ANSM. Ces dernières peuvent correspondre à une fraction thérapeutique (FT) ou à une substance active (SA). Les méthodes d'alignement décrites plus haut pour DrugBank ont été réutilisées ici. Le vocabulaire "source" était l'ensemble des termes du thesaurus et le vocabulaire "cible" l'ensemble des substances décrites dans le répertoire des médicaments.

Comme cette étape est indépendante des précédentes et qu'elle sera utilisée pour la recherche d'interactions sur des données de délivrance sur la période de juin à août 2013, le vocabulaire du thesaurus des interactions de juillet 2013 a été utilisé.

5.3 Comparaison avec Merged-PDDI

Une fois l'alignement réalisé avec DrugBank, les interactions du thesaurus pouvaient être décrites par un couple d'identifiants DrugBank. Pour réaliser la comparaison, seuls les identifiants DrugBank communs au thesaurus et à la base de données Merged-PDDI ont été conservés. Ceci permettait de prendre en compte des différences évidentes liées à des molécules non commercialisées en France ou aux Etats-Unis.

Le chevauchement entre une source 'A' par rapport à une source 'B', comme défini par Ayvaz. et al.[8], est le rapport entre le nombre d'interaction en commun et le nombre d'interaction décrit dans la source 'A'.

5.4 Analyse des délivrances médicamenteuses

L'alignement réalisé avec le répertoire du médicament permettait d'inférer que deux spécialités pharmaceutiques interagissent car l'interaction de leur(s) substance(s) était décrite dans le thesaurus.

Une bonne modélisation des connaissances avec une ontologie aurait permis de réaliser cette inférence automatiquement par une machine. Cependant, cette modélisation est compliquée à développer et pourraît être l'objet d'un travail de recherche spécifique. Aussi, comme l'objectif était d'analyser un grand nombre de délivrances et non une seule, il a été nécessaire de développer un programme spécifique pour cette tâche.

L'objectif était de rechercher les interactions médicamenteuses sur des données de délivrance sur la période juin à août 2013 à partir des informations du thesaurus des interactions de l'ANSM de juillet 2013.

Une spécialité pharmaceutique était considérée à risque d'interaction si celle-ci n'était pas un médicament homéopathique ou un médicament administré localement. Les médicaments homéopathiques ne contiennent aucun principe actif et les médicaments d'application locale, sauf exceptions citées plus bas, ne présentent pas de risque d'interaction.

Les critères d'exclusion des spécialités pour l'analyse étaient les suivants :

— format incorrect du code CIP7
  Le code doit comporter strictement 7 chiffres.
— code CIP7 inconnu du répertoire des médicaments
  Le dossier pharmaceutique contient des médicaments et d'autres produits de santé. Ces derniers ne figurent pas dans le répertoire des médicaments de l'ANSM. Certains codes CIP7 sont donc inconnus.
— médicament homéopathique
— médicament administré localement
  D'après les informations générales du thesaurus, les voies locales ne sont pas concernées par les interactions des voies systémiques sauf dans les cas suivants :
  — bêta-bloquants en collyre
  — pilocarpine en collyre
  — miconazole gingival
  — éconazole toutes formes
  — spermicides vaginaux
  — antiseptiques iodés et mercuriels

La recherche d'interaction médicamenteuse sur une liste de spécialités pharmaceutiques peut s'effectuer de deux façons :

1. Extraction des substances de chaque médicament et dénombrement des couples de substances. Sur ces couples de substances sont recherchés des interactions.
2. Dénombrement des couples de médicaments puis recherche des interactions entre les substances de chaque couple.

La différence entre ces deux méthodes est liée au fait qu'un médicament peut contenir deux substances actives qui sont à risque d'interagir d'après le thesaurus. Par exemple, le DUOPLAVIN contient 75 mg de clopidogrel et 75mg d'acide acétylsalicylique. La première méthode comptera une interaction car la prise de deux antiagrégants plaquettaires est à risque d'interaction (majoration du risque hémorragique par addition des activités antiagrégantes plaquettaires). La deuxième méthode n'en comptera aucune. Les deux méthodes donneront le même résultat si ces deux substances sont prescrites dans deux médicaments différents, par exemple KARDEGIC 75 et PLAVIX 75. La deuxième méthode a été utilisée car c'est celle utilisée par les logiciels d'aide à la prescription. Il serait en effet difficilement compréhensible

par un utilisateur de recevoir une alerte en présence d'un seul médicament. Le nombre de médicaments générant à eux seuls une alerte avec la première méthode a été dénombré.

5.4.1 Implémentation

Pour des raisons d'efficience algorithmique, chaque délivrance ne doit pas être analysée individuellement. Un même couple de médicaments peut apparaître sur plusieurs délivrances ; il n'est pas efficient de répéter l'analyse. Nous souhaitions aussi éviter de répéter toute l'analyse en cas de modification de la source d'interaction. Un index inversé a été créé pour répondre à ces problématiques. Cette structure de données est très utilisée par les moteurs de recherche pour trouver en quelques millisecondes des termes présents dans plusieurs millions de documents sur le web. Dans notre cas, les documents correspondaient aux délivrances et les termes aux médicaments. Par exemple, à partir des délivrances 1 et 2 suivantes :

```
délivrance1 = {paracétamol,dompéridone}
délivrance2 = {paracétamol,dompéridone, escitalopram}
```

L'index inversé suivant a été créé :

```
paracétamol = {délivrance1, délivrance2}
dompéridone = {délivrance1, délivrance2}
escitalopram = {délivrance2}
```

Une recherche booléenne "molecule1" ET "molecule2" dans l'index inversé a permis de trouver toutes les délivrances contenant les deux molécules. Par exemple, l'ensemble des délivrances contenant le couple {dompéridone,escitalopram} correspond à l'intersection entre l'ensemble des délivrances contenant la dompéridone et l'ensemble des délivrances contenant l'escitalopram. Dans l'exemple plus haut, il s'agit de la délivrance 2 et on dénombre une seule ordonnance pour ce couple. Contrairement à l'exemple donné, les codes CIP et non les molécules ont été utilisés comme entrées de l'index inversé.

Comme les délivrances sont indépendantes les unes des autres, il est possible de distribuer les analyses sur plusieurs machines. Un pool d'ordonnances peut être analysé par une machine, puis les résultats de plusieurs machines sont fusionnés. Quarante machines du cluster Avakas ont été utilisées pour les analyses.

Une fois l'index inversé créé, le dénombrement du nombre de couples d'interaction a consisté à réaliser une recherche booléenne dans l'index. La liste des couples du thesaurus de juillet 2013 a été utilisée puis la liste américaine ONC high-priority [30] lui a été comparée. Cette liste est utilisée par certains logiciels aux Etats-Unis pour la détection de contre-indications médicamenteuses.

Résultats

6.1 Extraction du thesaurus

6.1.1 Transformation en RDF

Le résultat de la tranformation du texte brut en format structuré pour la molécule abatacept est donné dans le tableau 6.1.

```
ABATACEPT
+ ANTI-TNF ALPHA
Association DECONSEILLEE Majoration de l'immunodépression.
+ VACCINS VIVANTS ATTÉNUÉS
Association DECONSEILLEE ainsi que pendant les 3 mois suivant l'arrêt du traitement.
Risque de maladie vaccinale généralisée, éventuellement mortelle.
```

protagoniste 1	protagoniste 2	niveau	mécanisme	description
abatacept	anti-TNF alpha	AD	Majoration de l'immunodepression.	
abatacept	Vaccins vivants atténués	AD	Risque de maladie vaccinale ...	Ainsi que pendant les 3 mois...

Tableau 6.1 – Transformation des interactions de l'abatacept en format structuré.

Ce format structuré a ensuite été transformé en triplets RDF. Les triplets concernant l'abatacept et les anti-TNF alpha sont représentés sur la figure 6.1. Ces triplets peuvent ensuite être chargés dans un triplestore pour les interroger en SPARQL. Avec l'ontologie, un raisonneur est capable d'inférer les relations d'interaction entre l'abatacept et les molécules appartenant à la classe des anti-TNF alpha.

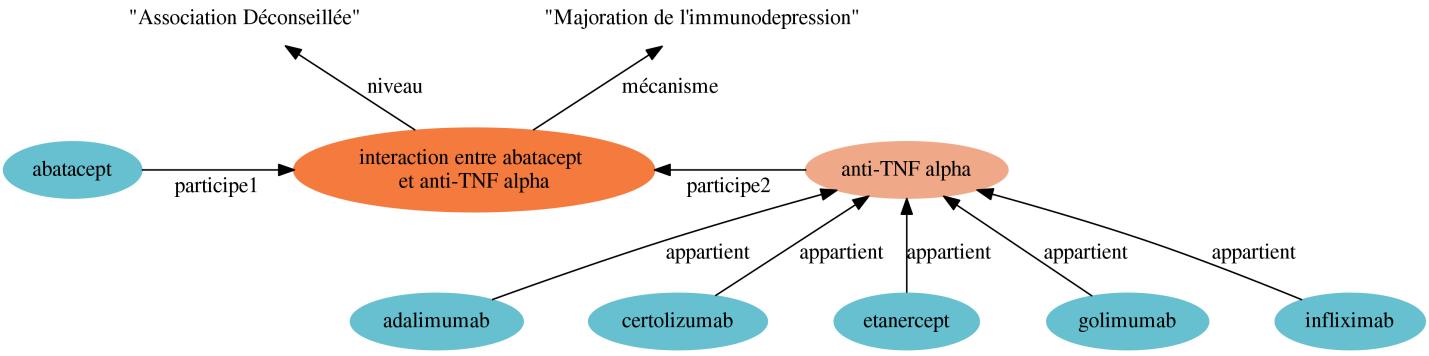

FIGURE 6.1 – Représentation graphique des triplets RDF décrivant l'interaction entre l'abatacept et les anti-TNF alpha

6.1.2 Description de son contenu

Le thesaurus des interactions de l'ANSM dans sa version de janvier 2016 décrit 168 classes thérapeutiques. Pour chaque classe thérapeutique, la liste des molécules la constituant est fournie sauf pour quatre classes :

- Les médicaments administrés par voie orale
- Les médicaments utilisés par voie vaginale
- Les sels de fer par voie injectable
- Les médicaments agissant sur l'hémostase

Il existe aussi des aliments (l'alcool, la théine et le jus de pamplemousse) et des plantes (ricinus communis, la menthe, le cascara, le sené...) qui sont considérés comme des molécules car sources possibles d'interactions. Le thesaurus contient 2772 couples d'interaction, 1358 sont présentés en miroir, 56 sans miroir dont 22 sont des couples du même protagoniste (ex : aminosides - aminosides). Au total, il existe 1414 couples uniques.

Nombre de classes thérapeutiques	168
Nombre de molécules	1102
Nombre de molécules appartenant à une classe thérapeutique	969
Nombre de molécules sans classe thérapeutique	133
Nombre de couples "protagoniste1 - protagoniste2"	1414
Nombre de couples "classe thérapeutique1 - classe thérapeutique2"	228
Nombre de couples "classe thérapeutique - molécule"	564
Nombre de couples "molécule1 - molécule2"	622
Nombre de couples "molécule1 - molécule2" après inférence	52870

Tableau 6.2 – Analyse descriptive du thesaurus des interactions de l'ANSM de janvier 2016.

Dans la version de juin 2015, une faute d'orthographe dans le nom de la substance racécadotril (racécadodril), présente dans le fichier index des substances mais pas dans le thesaurus, montre que ces fichiers n'ont pas été générés à partir de la même source. Les accents ne sont pas toujours retirés du nom des molécules. Dans la version de janvier 2016, on trouve des doublons dans le pdf de l'index des substances : cétuximab et cetuximab.

L'interaction entre deux molécules peut être décrite plusieurs fois dans le thesaurus car issue de protagonistes 'a' et 'b' différents. Un exemple pour le couple "métoprolol - quinidine" est donné dans le tableau 6.3.

Protagoniste 'a' (contient métoprolol)	Protagoniste 'b' (contient quinidine)	niveau
beta-bloquants dans l'insuffisance cardiaque	médicaments susceptibles de donner des TdP	CI
bradycardisants	bradycardisants	PC
bradycardisants	médicaments susceptibles de donner des TdP	PE
beta-bloquants (sauf esmolol et sotalol)	antiarythmiques classe I (sauf lidocaïne)	PE
beta-bloquants dans l'insuffisance cardiaque	antiarythmiques classe I (sauf lidocaïne)	PE

Tableau 6.3 – Au total, cinq protagonistes décrivent l'interaction entre le métoprolol et la quinidine. TdP : torsades de pointes, PC : à prendre en compte, PE : précaution d'emploi, CI : contre-indication

En décomposant par inférence les 1414 couples uniques "protagoniste1 - protagoniste2" présents dans le thesaurus des interactions de l'ANSM 2016, on obtient 52 870 couples uniques "molécule1 - molécule2". Parmi ces 52 870 couples uniques, 10 774 (20%) ont au moins deux origines. Un couple "molécule1 - molécule2" peut être généré par 1 à 5 couples "protagoniste1 - protagoniste2".

Ces couples sont uniques, c'est-à-dire que les couples "molécule2-molécule1" ne sont pas comptés. Pour 792 couples inférés, molécule1 et molécule2 sont identiques. Ceci survient quand les deux protagonistes sont identiques. Ces 792 couples ne répondent pas à la définition d'une interaction médicamenteuse qui requière deux molécules différentes. L'ontologie ne prend pas en compte ces exceptions.

Le niveau de contrainte par couple de protagonistes est donné sur la figure 6.2. Le niveau de contrainte par couple de molécules est donné sur la figure 6.3.

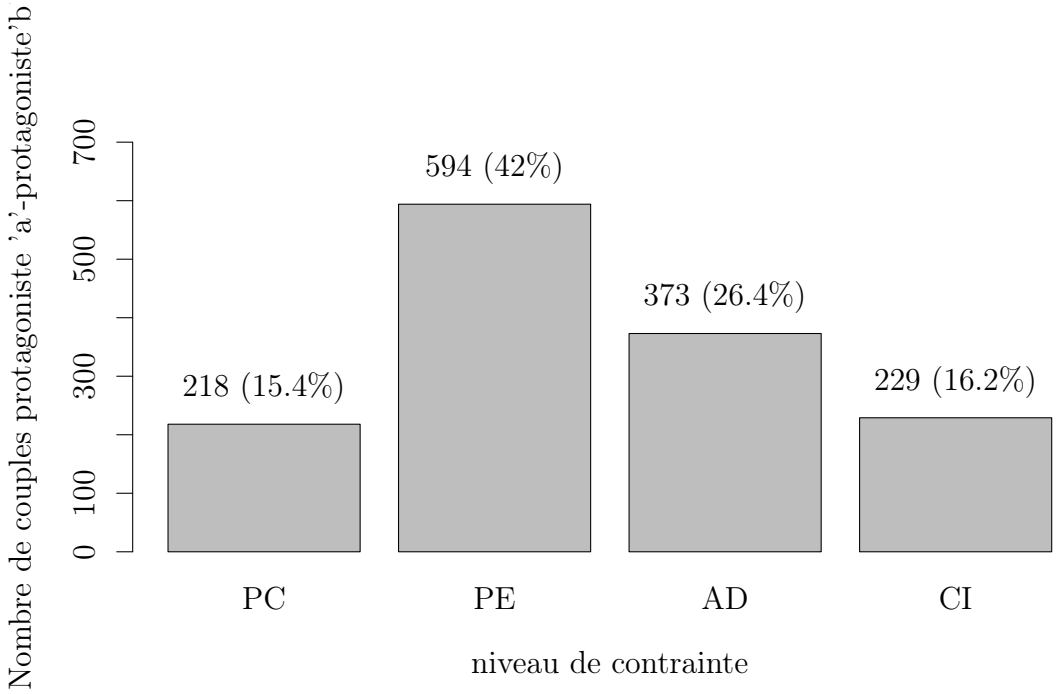

FIGURE 6.2 – Répartition du niveau de contrainte pour les couples uniques de protagonistes dans le thesaurus de l'ANSM de janvier 2016. PC : à prendre en compte, PE : précaution d'emploi, AD : association déconseillée, CI : contre-indication.

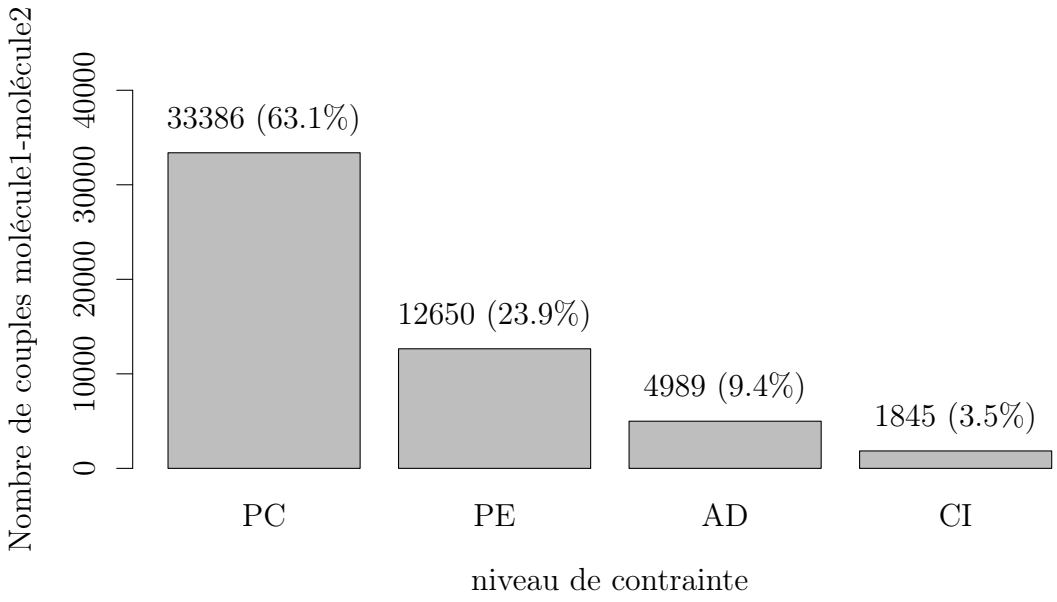

FIGURE 6.3 – Répartition du niveau de contrainte pour les couples uniques de molécules dans le thesaurus de l'ANSM de janvier 2016. PC : à prendre en compte, PE : précaution d'emploi, AD : association déconseillée, CI : contre-indication.

6.1.3 Structure hiérarchique

Parmi les 168 classes thérapeutiques présentes dans le thesaurus, 160 relations de type "subClassOf" ont été créées entre deux classes par le raisonneur HermiT. Cette relation signifie que toutes les molécules d'une famille 'a' sont contenues dans une famille 'b'. Ces relations ne sont pas décrites dans le thesaurus. Des exemples de relations sont donnés sur la figure 6.4.

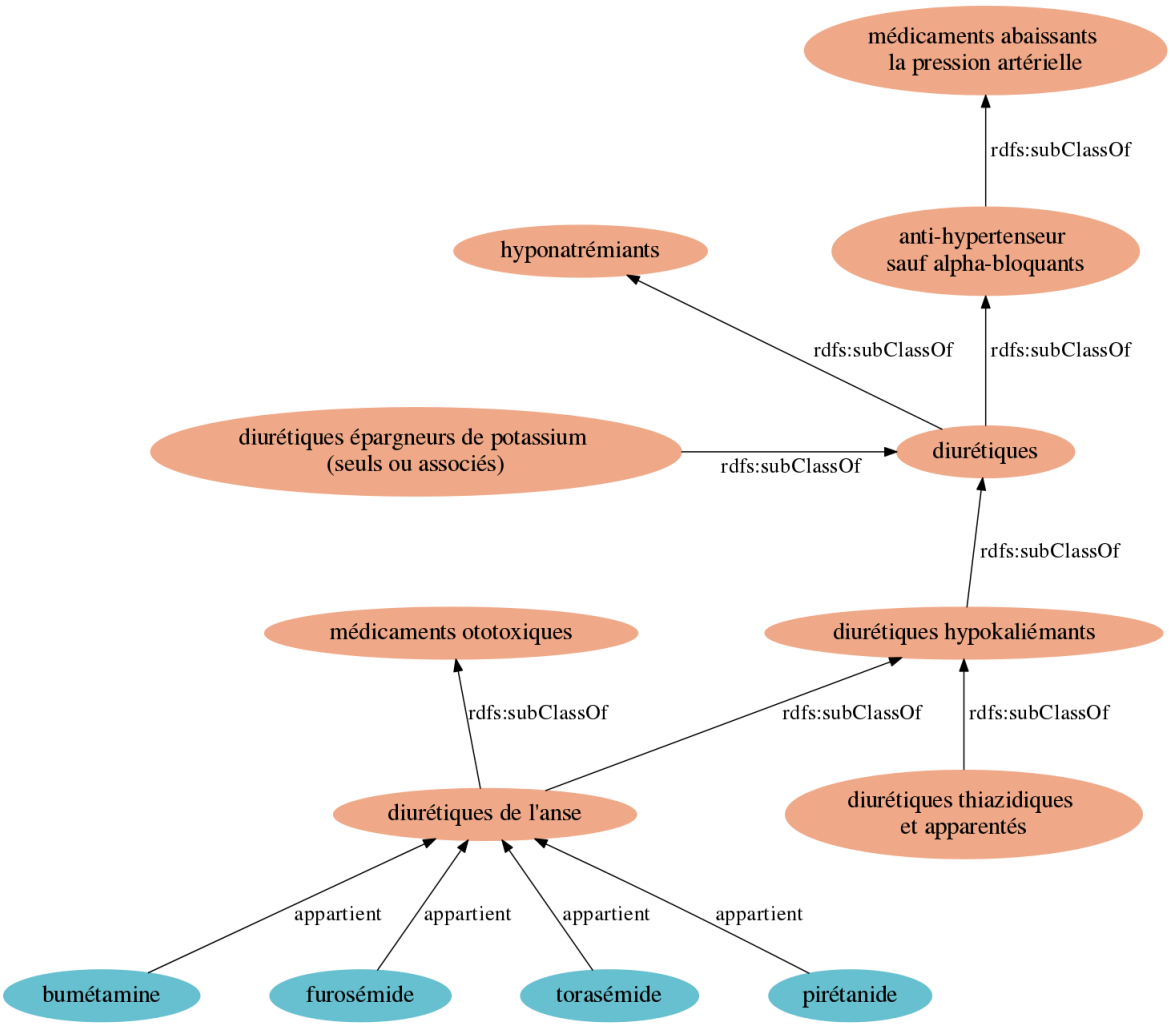

FIGURE 6.4 – Exemples de relations subClassOf inférées par le raisonneur HermiT entre les classes thérapeutiques du thesaurus des interactions de l'ANSM de janvier 2016. Comme tous les médicaments des diurétiques de l'anse (bumétamine, furosémide, torasémide, pirétanide) appartiennent à la classe des médicaments ototoxiques, le raisonneur infère que la première classe est une sous-classe de la seconde.

6.1.4 Score subClassOf

Les relations "subClassOf" ne sont pas créées si une classe A n'est pas totalement incluse dans une classe B. Cette relation est parfois rejetée car une seule molécule de la classe A n'appartient pas à la classe B. Les cinq premiers exemples triés par score subClassOf sont donnés dans le tableau 6.4. En prenant l'exemple de la première ligne, la lecture du tableau est la suivante : toutes les molécules appartenant à la classe des "neuroleptiques susceptibles de donner des torsades de pointes" (15) sont des "médicaments à l'origine d'une hypotension orthostatique" sauf une : l'amisulpride. La classe des neuroleptiques susceptibles de donner

des torsades de pointes n'est donc pas une subClassOf de la classe des médicaments à l'origine d'une hypotension orthostatique.

famille 1	N	famille 2	N_sous	absentes	score
médicaments à l'origine d'une HO	81	neuroleptiques susceptibles de donner des TdP	15	amisulpride	0.07
médicaments sédatifs	188	analgésiques morphiniques agonistes	14	phénopéridine	0.08
morphiniques	24	analgésiques morphiniques agonistes	14	tapentadol	0.08
medicaments à l'origine d'une HO	81	antiparkinsoniens dopaminergiques	13	rotigotine	0.08
médicaments sédatifs	188	analgésiques morphiniques de palier III	11	phénopéridine	0.10

Tableau 6.4 – Toutes les molécules de la famille 2 appartiennent à la famille 1 sauf les molécules absentes. La relation subClassOf n'est donc pas créée à cause de ces différences. score : score de subClassOf. HO : hypotension orthostatique. TdP : torsade de pointes. absentes : molécules absentes

6.2 Alignement

6.2.1 vers l'UMLS

Sur les 1102 molécules du thesaurus, 951 ont été alignées automatiquement (86,3%), 13 semi-automatiquement et 138 ont nécessité une recherche manuelle. Des exemples sont donnés dans le tableau 6.5.

libellé thesaurus	terminologie	libellé synonyme	CUI	libellé préféré	alignement
Sodium (oxybate de)	MSHFRE	Oxybate de sodium	C0037537	Sodium Oxybate	semi-automatique
Acepromet azine	RXNORM	acepromet azine	C0608826	acepromet azine	automatique
Aciclovir	RCD	Aciclovir	C0001367	Acyclovir	automatique
Acide alendronique	-	non trouvé	-	-	manuel
Acide folinique	MSHFRE	Acide folinique	C0023413	Leucovorin	automatique

Tableau 6.5 – Exemples d'alignement vers l'UMLS. Pour chaque libellé du thesaurus, un libellé similaire est recherché parmi tous les termes de l'UMLS. Le libellé du thesaurus est rattaché à un concept UMLS automatiquement s'il correspond exactement au libellé préféré ou synonyme. MSHFRE : Mesh en français, RCD : Read Codes, CUI : identifiant du concept UMLS

La recherche manuelle a conduit à modifier le libellé dans le thesaurus pour permettre un alignement exact. Par exemple, la "caféine" a été modifiée en "Caffeine", le "p a s sodique" fait référence à l'acide aminosalicylique et les acides ont subi une même modification : le "e"

de acide a été retiré et le suffixe "ique" transformé en "ic". Au total, 45 règles manuelles ont permis d'aligner 136 molécules non trouvées automatiquement. La seule molécule non trouvée dans l'UMLS est la "monmectite". Le moteur de recherche Google ne renvoie que 539 résultats pour celle-ci. Elle appartient à la classe des "topiques gastro-intestinaux, antiacides et adsorbants" dans le thesaurus. 75 concepts UMLS (7%) ne sont pas rattachés à un concept RxNorm.

6.2.2 vers DrugBank

Dans un premier temps, les synonymes en anglais des termes du thesaurus sont recherchés grâce aux alignements de l'UMLS réalisés à l'étape précédente. Le nombre de synonymes retrouvés par terme varie de 0 à 51, la médiane est de 8 synonymes. Par exemple, pour la cafédrine, "cafedrine" est le libellé préféré du concept (C0068978) et deux synonymes sont trouvés : "norephendrinetheophylline" et "7-(2-(1-methyl-2-hydroxy-2-phenylethylamino)ethyl) theophylline". Dans un second temps, les termes du thesaurus, les libellés préférés des concepts et leurs synonymes sont utilisés pour trouver une correspondance dans Drugbank. Le fichier vocabulaire de DrugBank contient 8221 identifiants de molécule et 21 710 libellés préférés ou synonymes uniques.

Au total, 860 molécules sont alignées automatiquement vers DrugBank (78%). Des exemples d'alignements automatiques sont donnés dans le tableau 6.6.

Identifiant DrugBank	libellé préféré DrugBank	Identifiant UMLS	libellé thesaurus	libellé préféré UMLS
DB00787	Aciclovir	C0001367	Aciclovir	Acyclovir
DB01284	Cosyntropin	C0010192	Tetracosactide	Cosyntropin
DB04077	Glycerol	C0017861	Glycerol	Glycerin
DB09462	Glycerin	C0017861	Glycerol	Glycerin
DB00650	Leucovorin	C0023413	Acide folinique	Leucovorin

Tableau 6.6 – L'alignement est automatique quand le libellé préféré de Drugbank est identique au libellé du thesaurus (aciclovir par exemple) ou le libellé préféré du concept UMLS (leucovorin par exemple). Un conflit apparaît pour le glycérol.

Un seul conflit est survenu à cette étape : DrugBank considère que le glycérol et la glycérine sont deux molécules distinctes alors que l'UMLS considère qu'il s'agit de la même molécule. Dans ce cas, c'est le terme du thesaurus (glycérol) qui a été privilégié plutôt que glycérine. Treize alignements supplémentaires ont été réalisés entre un synonyme UMLS et le libellé préféré de DrugBank (tableau 6.7).

idDB	synonyme UMLS	libellé préféré DrugBank	libellé thesaurus	libellé préféré UMLS
DB06799	Hexamethylenetetramine	Hexamethylenetetramine	Methenamine	Methenamine
DB00583	L-Carnitine	L-carnitine	Levocarnitine	Levocarnitine
DB01323	St. John's Wort	St. john's wort	Millepertuis	Hypericum perforatum

Tableau 6.7 – Les synonymes des libellés du thesaurus trouvés dans l'UMLS permettent de trouver le libellé préféré utilisé par DrugBank. idDB : identifiant DrugBank

Enfin, 28 alignements ont été réalisés entre un synonyme UMLS et un synonyme de DrugBank. Certains homonymes ont été exclus manuellement, par exemple la lettre "F" est le synonyme de la L-Phenylalanine dans DrugBank tandis qu'elle représente le Fluor dans l'UMLS.

Les alignements imparfaits ont été revus. La majorité concernait un ingrédient et son sel. Le thesaurus peut décrire l'ingrédient et DrugBank son sel comme Fondaparinux et Fondaparinux sodium et inversement comme citrate de magnésium et magnésium. Dans ces cas, les relations sont respectivement "RxNorm:has_form" et "RxNorm:form_of". D'autres alignements imparfaits concernaient des relations de partonomie comme "Pristinamycin IIA" décrit dans DrugBank et la "pristinamycine" décrit dans le thesaurus. La molécule est composée de deux parties : la pristinamycine IA et la pristinamycine IIA.

Au total, 958 substances du thesaurus (86,7%) ont pu être reliées à Drugbank dont 907 sont considérées être exactement les mêmes (owl:sameAs). Parmi les substances ne semblant pas figurer dans l'encyclopédie, on trouve des molécules non commercialisées outre-atlantique comme le fluindione, tous les vaccins, certains éléments chimiques comme l'or et le fluor, certaines plantes : ricinus communis, la menthe, le cascara (le sené figure néanmoins dans DrugBank).

6.2.3 vers le répertoire du médicament

Le thesaurus décrit généralement des fractions thérapeutiques comme "amoxicilline" mais parfois des ingrédients précis comme "périndopril tert-butylamine" dont l'ingrédient, "périndopril", est absent du thesaurus.

Dans le répertoire du médicament, les spécialités pharmaceutiques n'ont pas toujours de fraction thérapeutique. Par exemple, les spécialités contenant de la "pravastatine sodique" (SA) ne sont pas décrites comme contenant de la "pravastatine" (FT). La relation entre la "pravastatine" du thesaurus et la "pravastatine sodique" du répertoire du médicament est "RxNorm:form_of" (figure 6.5).

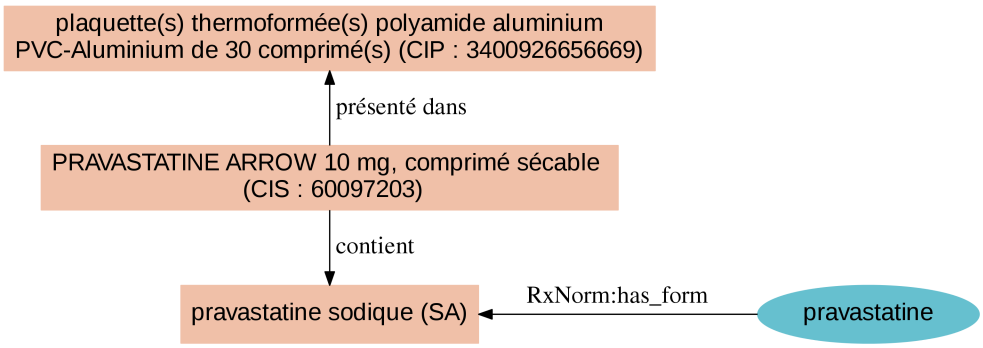

FIGURE 6.5 – Lien entre la pravastatine décrite dans le thesaurus des interactions (en bleu) et la pravastatine sodique décrite dans le répertoire du médicament. La fraction thérapeutique de la pravastatine sodique, la pravastatine, n'est pas présente dans le répertoire. SA : substance active.

Le thesaurus dans sa version de juillet 2013 contient 1020 molécules. Parmi celles-ci, 765 (75%) sont présentes (owl:sameAs) dans le répertoire du médicament. 218 (21%) correspondent à une fraction thérapeutique non présente dans le répertoire du médicament. Ces molécules sont présentes dans le répertoire sous une forme ("RxNorm:form_of") : de chlorydrate (116), de sel sodique (24), de maléate (13), d'acétate (13) et d'autres. Les molécules du thesaurus non trouvées dans le répertoire sont au nombre de 37. Il s'agit en majorité de barbituriques (barbital, allobarbital...) et de benzodiazépines (camazépam, fludiazepam...) qui ne sont pas commercialisés en France. On note aussi la présence du nonoxynol-9 qui est considéré comme un excipient de certaines spécialités comme la povidone iodée TEVA 10% solution vaginale.

6.3 Comparaison avec Merged-PDDI

A l'étape précédente, 958 molécules décrites dans le thesaurus de l'ANSM ont reçu un identifiant DrugBank si la relation était "owl:SameAs" ou "RxNorm:has_form" ou "RxNorm:form_of" entre la molécule du thesaurus et le concept de DrugBank. Parmi ces 958 molécules, 953 sont présentes dans la base Merged-PDDI d'Ayvaz et al. [8] regroupant 12 sources publiques américaines. Seuls les couples d'interaction contenant deux de ces 953 molécules en commun ont été sélectionnés dans les deux sources pour réaliser la comparaison. Dans ces sous-ensembles, le thesaurus de l'ANSM de janvier 2016 compte 42 954 couples de molécules à risque d'interaction et la base Merged-PDDI 13 851. Le chevauchement entre les deux sources est donné sur la figure 6.6. Au total, 5651 interactions (13%) décrites dans le thesaurus sont présentes dans la base Merged-PDDI et 5651 interactions (41%) décrites dans la base Merged-PDDI sont présentes dans le thesaurus de l'ANSM. La source ONC

high-priority, générant des alertes de contre-indication est couverte à 78% (394/504) par le thesaurus, sans prendre en compte les différences de niveau de contrainte.

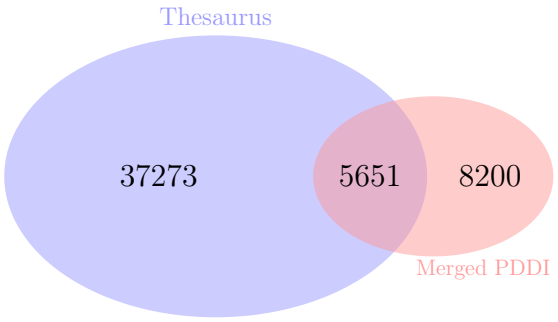

FIGURE 6.6 – Comparaison des interactions décrites entre le thesaurus de l'ANSM de janvier 2016 et la base Merged-PDDI (regroupant 12 sources) d'Ayvaz et al. sur 953 molécules en commun. Les différences de niveau de contrainte ne sont pas prises en compte.

6.4 Analyse des délivrances médicamenteuses

6.4.1 Description des données

Le thesaurus de juillet 2013 a été utilisé pour l'analyse des interactions. Il se situe dans la période où les prescriptions médicamenteuses ont été délivrées. Dans cette version du thesaurus, 996 molécules sont toujours présentes dans le thesaurus de janvier 2016 et 24 n'y figurent plus : la majorité concerne des molécules qui ne sont plus commercialisées (nelfinavir, nialamide), pour d'autres le nom a changé : "étorecoxib" est devenu "étoricoxib", "bupropione" a été changé en "bupropion". Cette version contient 50 534 couples de molécules à risque d'interaction ; 801 couples molécule1-molécule2 où molécule1 et molécule2 sont identiques ont été exclus.

Pour 111 couples à risque d'interaction les deux molécules sont associées dans un même médicament. Par exemple, pour le médicament PRAXINOR, l'association de ses deux molécules est une contre-indication. Il contient 100mg de cafédrine chlorydrate et 5mg de théodrénaline chlorydrate. Ces deux substances sont classées dans la classe thérapeutique des sympathomimétiques indirects et l'association de deux molécules de cette classe est contre-indiquée : risque de vasoconstriction et/ou de crises hypertensives. Ce médicament n'est plus commercialisé depuis le 27/11/2013.

L'échantillon du dossier pharmaceutique contient 7 015 571 délivrances et 19 586 205 codes CIP. Parmi ces codes CIP, 18 916 995 (96,6%) étaient dans un format correct, c'est-à-dire 7 chiffres. Parmi les codes CIP au format correct, 17 592 898 (93%) étaient connus

du répertoire du médicament de l'ANSM et étaient donc des médicaments. Une recherche manuelle a montré que les dix premiers codes CIP inconnus par ordre de fréquence correspondaient à des lancettes pour patients diabétiques. Pour 1 632 130 (9,3%) médicaments, la voie d'administration était locale et 5579 étaient des médicaments homéopathiques. La figure 6.7 montre le diagramme de flux des inclusions des codes CIP.

Parmi les 13 305 médicaments différents dans ces données, 12 886 ne sont ni des médicaments homéopathiques, ni des médicaments administrés par voie locale. Ils ont donc un risque potentiel d'interaction. Parmi ces derniers, 10 032(82%) contiennent au moins une substance présente dans le thesaurus des interactions.

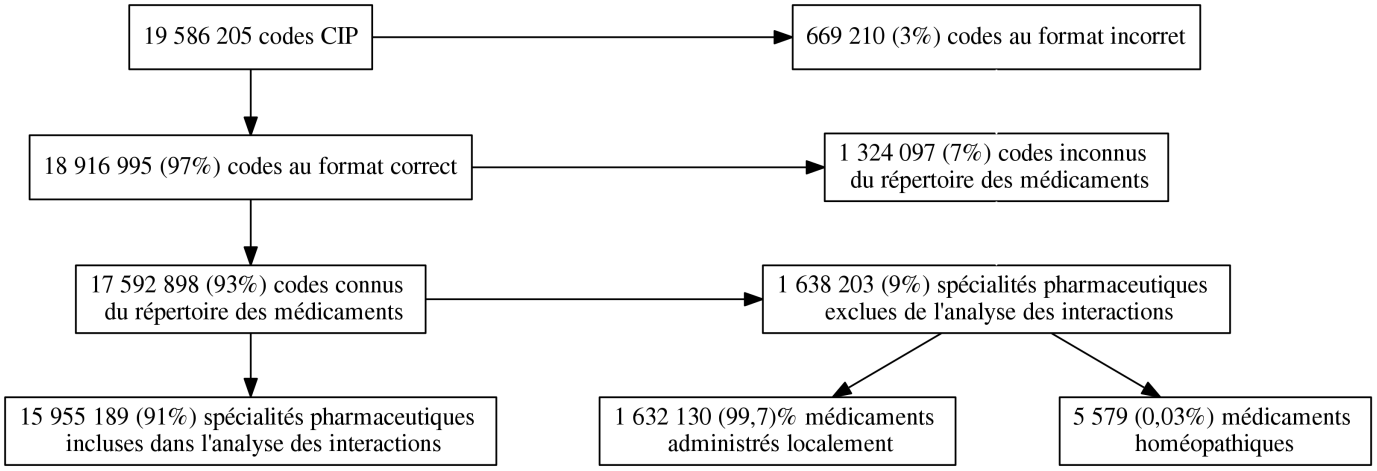

FIGURE 6.7 – Diagramme de flux des inclusions des codes CIP

Une figure descriptive sur la répartition de l'âge (figure A.2) est présentée en annexe.

6.4.2 Dénombrement des interactions et des alertes

Comme un couple de molécules peut être issu de plusieurs couples de protagonistes dans le thesaurus, plusieurs niveaux de contrainte peuvent lui être associés. Nous avons décidé de garder le niveau de contrainte le plus élevé sauf si ce dernier dépendait du contexte. Par exemple, pour le couple métoprolol - quinidine (tableau 6.3), le niveau de contrainte choisi pour le dénombrement est "précaution d'emploi" car la contre-indication entre ces deux molécules dépend de la condition "insuffisance cardiaque". Ce choix a été réalisé afin de ne pas surestimer le nombre de contre-indications médicamenteuses. Cependant ce couple génèrerait cinq alertes dans un LAP : une contre-indication, trois précautions d'emploi et une notification à prendre en compte. Dans le cas où ce couple était présent sur une délivrance, nous avons dénombré un couple d'interaction de niveau "précaution d'emploi" et cinq alertes générées.

La figure 6.8 montre le diagramme de flux du dénombrement des interactions et des alertes.

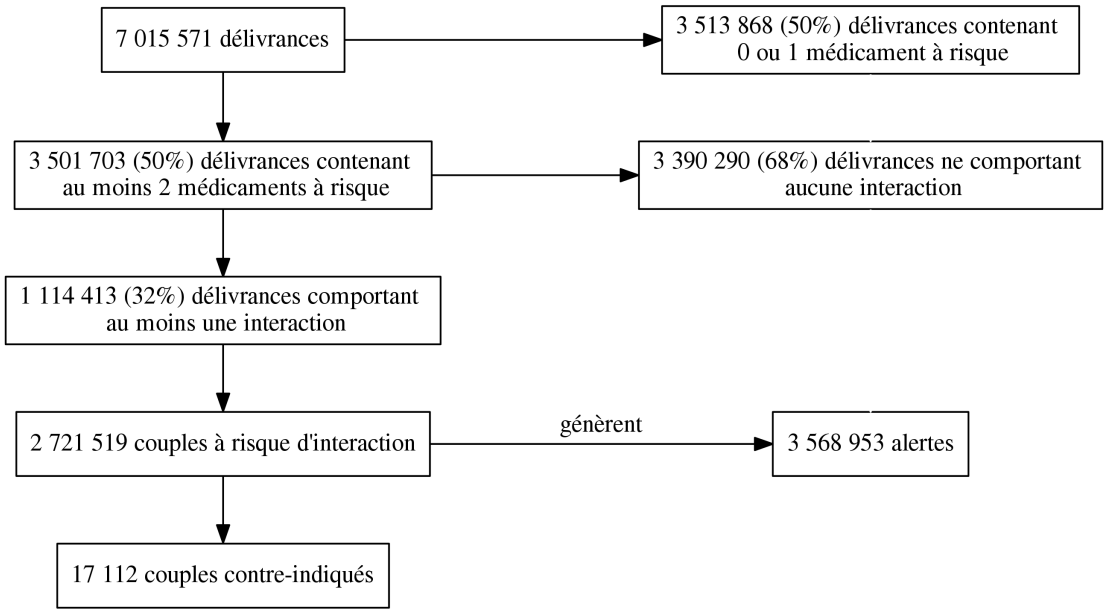

FIGURE 6.8 – Diagramme de flux du dénombrement des interactions et des alertes

Au total, 3 501 703 (50%) délivrances contenaient au moins deux spécialités pharmaceutiques possédant un risque potentiel d'interaction (figure A.3). Le programme dénombre 1 114 413 délivrances avec au moins un risque avéré d'interaction d'après les informations du thesaurus. En moyenne, 32% des délivrances avec un risque potentiel d'interaction comportent un risque avéré d'interaction d'après le thesaurus. Cependant, cette fréquence varie fortement avec le nombre de médicaments (figure A.4). Par exemple, 70% des délivrances avec 6 médicaments à risque comportent au moins une interaction.

Seulement 14 279 couples de molécules sur les 50 534 décrits dans le thesaurus sont détectés. Parmi ces 14 279 couples, 208 génèrent 50% des 2 721 519 risques d'interaction dénombrés. Ces 2 721 519 interactions génèreraient 3 568 953 alertes dans un LAP. Une figure descriptive du nombre d'alertes en fonction du nombre de médicaments à risque est présentée en annexe (figure A.6)

Les dix couples les plus fréquents sont présentés dans le tableau 6.8.

Niveau de contrainte	molécule1	molécule2	N délivrance
PE	acide acétylsalicylique	hydrochlorothiazide	57 862
PE	acide acétylsalicylique	furosémide	46 030
PE	acide acétylsalicylique	perindopril tert-butylamine	43 477
PE	diosmectite	médicaments administrés par voie orale	42 880
PE	acide acétylsalicylique	ramipril	34 667
PE	acide acétylsalicylique	valsartan	25 193
PE	acide acétylsalicylique	clopidogrel	24 435
PE	acide acétylsalicylique	irbesartan	23 162
PE	fluindione	paracétamol	22 883
PE	acide acétylsalicylique	candesartan cilexetil	19 391

Tableau 6.8 – Couples en interaction les plus fréquents sur des données de délivrance d'après les informations du thesaurus des interactions de l'ANSM de juillet 2013. N : nombre. PE : précaution d'emploi.

Les interactions sont des contre-indications dans 0.6% (17 112) des cas (figure A.5). 16 069 (94%) contre-indications sont générées par la combinaison d'un médicament de la classe "médicaments susceptibles de donner des torsades de pointes" avec un médicament de la classe "torsadogènes (sauf antiparasitaires, neuroleptiques, méthadone)". Les dix couples contre-indiqués les plus fréquents sont issus de ces deux couples de protagonistes et sont présentés dans le tableau 6.9. Le risque de toutes ces associations est une mort subite par troubles du rythme ventriculaire.

Niveau de contrainte	molécule1	molécule2	N délivrance
CI	cyamémazine	escitalopram	3024
CI	dompéridone	escitalopram	1961
CI	amiodarone	escitalopram	1283
CI	escitalopram	sotalol	787
CI	amiodarone	dompéridone	762
CI	amisulpride	escitalopram	685
CI	citalopram	cyamémazine	659
CI	escitalopram	tiapride	544
CI	dompéridone	sotalol	482
CI	citalopram	dompéridone	464

Tableau 6.9 – Couples contre-indiqués les plus fréquents sur des données de délivrance d'après les informations du thesaurus des interactions de l'ANSM de juillet 2013. N : nombre, CI : contre-indication.

En réalisant l'analyse avec la source ONC high-priority qui considère l'association de

deux molécules comme une contre-indication absolue, 21 035 alertes seraient générées. Le tableau 6.10 montre les dix couples les plus fréquents d'après cette source.

molécule1	molécule2	N délivrance	Niveau de contrainte thesaurus
simvastatine	vérapamil	3 619	PE
simvastatine	diltiazem	3 221	PE
simvastatine	amiodarone	2 873	PE
dompéridone	escitalopram	1 961	CI
amiodarone	escitalopram	1 283	CI
flécaïnide	sotalol	1 256	PC
escitalopram	flécaïnide	1 099	NA
amiodarone	flécaïnide	1 009	PC
escitalopram	sotalol	787	CI
amiodarone	dompéridone	762	CI

Tableau 6.10 – Couples les plus fréquents sur des données de délivrance d'après la source d'interaction américaine ONC high-priority générant une alerte bloquante (contre-indication). Le niveau de contrainte du thesaurus est donné pour information. PE : précaution d'emploi, CI : contre-indication, PC : à prendre en compte, NA : couple non décrit dans le thesaurus.

Le nombre de contre-indications générées par la source ONC high-priority sans que le thesaurus ne décrive d'interaction est de 1448. Le tableau 6.11 montre qu'il s'agit principalement d'interactions liées à la substance flécaïnide et au risque de torsades de pointes.

molécule1	molécule2	N délivrance
escitalopram	flécaïnide	1 099
flécaïnide	dompéridone	753
flécaïnide	citalopram	272
flécaïnide	sulpiride	57
flécaïnide	azithromycine	55
simvastatine	fluconazole	55
flécaïnide	halopéridol	44
flécaïnide	clarithromycine	32
flécaïnide	pimozide	7
flécaïnide	moxifloxacine	7

Tableau 6.11 – Couples les plus fréquents sur des données de délivrance d'après la source d'interaction américaine ONC high-priority générant une alerte bloquante (contre-indication) non présents dans les thesaurus des interactions de l'ANSM de juillet 2013 et janvier 2016. N : nombre.

6.5 Reproductibilité

Les publications scientifiques sont de plus en plus souvent accompagnées du partage des données et du code source développé pour les analyser, dans la limite des règles de confidentialité[46]. La transparence et la reproductiblité sont des marques de qualité de la recherche scientifique. Les systèmes de gestion de version comme Git permettent de partager des travaux de recherche et de collaborer avec d'autres chercheurs[46]. Dans cet esprit, nous mettons à la disposition des chercheurs les programmes suivants :

- IMthesaurusANSM [1] est un paquet R permettant de transformer le contenu du thesaurus des interactions dans un format structuré.
- LinkedThesaurus [2] est un dossier contenant l'ontologie et les programmes développés pour transformer le format structuré d'un thesaurus en RDF et les liens réalisés vers l'UMLS, DrugBank et le répertoire des médicaments de l'ANSM.
- FreqDDI [3] est le programme utilisé pour mesurer la fréquence des interactions sur des données de délivrance.

Ces programmes sont susceptibles d'évoluer. Git permet d'enregistrer toutes les modifications réalisées, faites par l'auteur ou de futurs collaborateurs, et de revenir au moment initial de leur publication.

1. https://github.com/scossin/IMthesaurusANSM
2. https://github.com/scossin/LinkedThesaurus
3. https://github.com/scossin/FreqDDI

Discussion

Un format ouvert

A notre connaissance, il s'agit de la première étude s'intéressant aux données du thesaurus et à ses liens vers d'autres sources de données. Le format PDF actuel du thesaurus limite son utilisation. Le bulletin d'information du centre régional de pharmacovigilance et d'information sur le médicament de Bordeaux numéro 107 [47] juge le thesaurus comme « un indispensable à consulter aussi souvent que possible, même si le format n'est pas adapté à la pratique ». Un format ouvert et réutilisable a été demandée en 2014 par la commission Open Data mandatée par Marisol Touraine.

La transformation d'un format fermé en format ouvert est une étape délicate et potentiellement source d'erreurs. La retranscription correcte des données lors de sa transformation devrait être vérifiée et vérifiable ce qui n'est pas le cas des solutions propriétaires. Une erreur à cette étape est susceptible d'entraîner la non-détection d'une interaction médicamenteuse. Dans leur analyse sur la sécurité d'utilisation des bases de données médicamenteuses des LAP et des LAD, Berthelot et al. [48] ont constaté des défaillances liées à cette étape entraînant une non-détection d'interactions médicamenteuses. La solution proposée dans l'article est la création par l'ANSM d'un référentiel structuré des interactions médicamenteuses directement intégrable dans les bases de données médicamenteuses.

Dans ce travail, nous proposons un format ouvert du thesaurus dans un but de recherche. Les données du thesaurus sont protégées par Copyright et peuvent être « reproduites ou traduites à des fins de recherche ou d'étude personnelle, mais ne peuvent être ni vendues ni utilisées à des fins commerciales ». Une publication officielle du thesaurus au format ouvert doit être une initiative de l'ANSM. Les identifiants uniformes de ressource (URI) utilisés par l'ANSM pourraient être de la forme : `http://ansm.sante.fr/resource/domperidone`. Il est important de permettre le déréférencement d'un URI lorsqu'un navigateur cherche à y accéder. Il est donc nécessaire de disposer des droits sur le nom de domaine (`http://ansm.sante.fr`).

La formalisation des connaissances proposée dans ce travail est limitée et ne donne pas accès à une machine à toutes les informations décrites dans le thesaurus notamment les facteurs modificateurs du risque d'une interaction, les mécanismes et les conduites à tenir. Des études ont montré que la prise en compte d'éléments contextuels améliorait la spécificité des alertes émises par un LAP [3]. Améliorer la représentation des connaissances sur les interactions médicamenteuses est un sujet de recherche d'actualité en informatique médicale. Des ontologies existent pour représenter les mécanismes d'interaction comme DIO[49](Drug inter-

action ontology) et DINTO[50] (Drug interaction ontology) mais n'ont pas été développées pour la pratique clinique. La représentation proposée permet d'identifier les concepts basiques de molécule, de famille et d'interaction avec leurs relations. La représentation en OWL permet d'inférer des relations hiérarchiques entre les classes thérapeutiques du thesaurus qui ne sont pas explicites. Certaines relations hiériarchiques n'existent pas à cause d'une molécule de différence entre deux grandes classes ce qui peut être le signe d'une incohérence de classification. Toutes les relations inférées n'ont pas été vérifiées, certaines relations peuvent être le fruit du hasard, notamment pour les petites classes. Un format OWL du thesaurus faciliterait la maintenance du référentiel car un raisonneur pourrait vérifier la cohérence d'une hiérarchie. Plusieurs applications manipulant le langage RDF faciliteraient aussi la visualisation de son contenu et la recherche d'information pour ses utilisateurs.

Seules les connaissances sur les interactions retenues par le groupe de travail sont présentes dans le PDF. Les interactions discutées mais non retenues et l'absence d'interactions démontrée par des études (interactions négatives) n'apparaissent pas. Il n'est pas possible de connaître l'historique et les anciennes versions ne sont plus accessibles en ligne. Les technologies du web sémantiques permettraient de gérer les interactions non retenues, les interactions négatives et les historiques de chaque interaction.

Des données liées

La publication de données du thesaurus liées sémantiquement à d'autres sources permet de les intégrer automatiquement à un réseau global d'information sur les médicaments. Ces liens permettent de récupérer des informations d'autres sources. Par exemple, le lien réalisé vers un concept de l'UMLS (et RxNorm) permet de récupérer les codes ATC pour rechercher des interactions par famille ATC, de récupérer le descripteur MeSH pour rechercher des articles scientifiques, de connaître la traduction d'une molécule dans d'autres langues ou de rechercher un médicament commercialisé aux Etats-Unis. Le lien réalisé vers DrugBank fournit de nombreuses informations biochimiques sur chaque molécule qui peuvent être utiles à la compréhension du mécanisme de l'interaction. DrugBank fournit aussi d'autres liens vers des sources d'information. Enfin, le lien réalisé vers le répertoire du médicament permet de connaître les médicaments commercialisés en France contenant une molécule du thesaurus.

Les liens ont été réalisés en utilisant les libellés des termes pour trouver des correspondances dans l'UMLS, DrugBank et le répertoire des médicaments. Le grand volume de données biomédicales contenu dans l'UMLS a permis de trouver tous les concepts correspondants sauf pour la monmectite. La présence de terminologies françaises, notamment la terminologie MeSH traduite par une équipe INSERM [51], a aussi facilité les liens. Cependant, certains liens ont été réalisés manuellement et auraient nécessité une validation par des experts du domaine. La non-vérification de ces liens est la principale limite de cette étape. La publication du code source permet aux chercheurs de détecter et corriger d'éventuelles erreurs réalisées.

On constate que la source RxNorm est bien structurée ce qui facilite la recherche d'information sur les médicaments. Le répertoire du médicament et le thesaurus de l'ANSM ne séparent pas clairement les principes actifs et leurs différentes formes. La difficulté de trouver des liens exacts entre ces sources reflète le manque de structuration de ces données. Comme le médicament est universel, la structure de RxNorm pourrait être utilisée pour créer une terminologie française des médicaments.

Comparaison des sources d'interaction

Contrairement à la France, il n'existe pas de référentiel national aux Etats-Unis sur les interactions médicamenteuses. Des travaux récents [8][9] ont cherché à regrouper différentes sources ouvertes. Intégrer ces informations permet de réunir la description du mécanisme, les évènements indésirables cliniques, la conduite à tenir et le niveau de contrainte qui sont plus ou moins détaillés dans chacune des sources. Cette intégration permet ensuite de comparer leur chevauchement. Il existe cependant de nombreuses limites à cette comparaison qui sont inhérentes à la création des sources et leur objectif : certaines couvrent un nombre limité de molécules, d'autres sont générées par traitement automatique de la langue, d'autres sont créées par un panel d'expert. Le chevauchement des 14 sources internationales est présenté dans une matrice dans l'article de Ayvaz et al. qui trouvent des chevauchements faibles. Les sources spécifiques comme ONC ne sont pas complètement incluses dans des sources généralistes comme NDF-RT ou DrugBank. Peters et al. [28] montrent un chevauchement faible de 24% entre DrugBank et NDF-RT. Environ 60% des alertes ONC étaient présentes dans NDF-RT.

Dans notre étude, la comparaison a été effectuée sur les molécules communes entre le thesaurus et les autres sources pour ne pas prendre en compte des différences liées à leur commercialisation. La première observation est le nombre élevé d'interactions (37 273) décrites dans la source française par rapport à l'ensemble des autres sources internationales (8200). Une des raisons possibles est l'utilisation de nombreuses classes thérapeutiques dans le thesaurus (168) ce qui conduit à générer de nombreux couples de molécules en interaction, notamment le niveau "à prendre en compte" (figure 6.3). L'actualisation des connaissances peut aussi être un biais dans cette comparaison car les sources ont des dates de publication différentes. Concernant les contre-indications, 110 couples (22%) des alertes de ONC high-priority[30] ne sont pas présentes dans le thesaurus des interactions, sans prise en compte du niveau de contrainte du thesaurus. L'absence d'interactions non retenues dans le thesaurus ne permet pas de conclure si ces 110 contre-indications d'après ONC ont été jugées non pertinentes cliniquement ou si elles n'ont pas été discutées en groupe de travail.

Les défauts d'une source d'information sur les interactions peuvent être les suivants :

— Manquer une interaction cliniquement significative entre deux molécules

— Envoyer un excès d'informations aux cliniciens qui n'arrivent pas à filtrer l'information pertinente en un temps limité, rendant la source inutilisable en pratique

— Envoyer un niveau d'alerte trop élevé conduisant le clinicien à modifier son attitude thérapeutique alors que la balance bénéfice-risque est favorable.
— Envoyer un niveau d'alerte trop bas conduisant le clinicien à ne pas modifier son attitude thérapeutique alors que la balance bénéfice-risque est défavorable.

Ce travail n'a pas cherché à expliquer les différences constatées entre le thesaurus et les autres sources. L'évaluation d'une source d'information sur les interactions est une tâche délicate et complexe, réalisable uniquement par des experts du domaine.

De nombreuses études américaines ont montré des discordances entre plusieurs sources pour la pratique clinique [52]. Par exemple, Wang et al. [53] ont testé trois logiciels américains leaders sur le marché pour la détection des interactions médicamenteuses en selectionnant 59 couples de médicaments contre-indiqués d'après la FDA. Seulement 68% des contre-indications étaient couvertes par toutes les sources. Hazlet et al. [54] ont testé neuf programmes d'interaction médicamenteuse installés dans des officines. En prenant 16 interactions médicamenteuses bien décrites dans la littérature, ils trouvèrent des sensibilités allant de 0.44 à 0.88. Les explications avancées sont la difficulté de maintenir une base de connaissance, la présence de différences de jugement inter-observateurs et la politique des vendeurs de logiciels qui, pour se couvrir d'éventuelles poursuites judiciaires, prévilégient la sensibilité à la spécificité [52][3][55].

Le thesaurus est un référentiel national français intégré dans les principaux logiciels d'aide à la prescription en France. Il ne devrait donc pas y avoir de grandes différences entre ces logiciels pour la détection des interactions. Cette hypothèse reste à confirmer.

Analyse de délivrances

Comparaison des sources L'analyse des délivrances par la source ONC high-priority permet d'identifier des différences avec le thesaurus en situation réelle.

Concernant les différences de couverture, 1448 contre-indications auraient été générées par ONC high-priority sans que le thesaurus ne signale d'interaction. Celles-ci concernent principalement la molécule flécaïnide qui présente un risque de torsades de pointes d'après ONC [1].

La principale différence de niveau de contrainte concerne l'association entre le vérapamil (inhibiteur calcique) et la simvastatine (statine). D'après le thesaurus, il s'agit d'une précaution d'emploi (niveau 2) : « ne pas dépasser la posologie de 20 mg/j de simvastatine ou utiliser une autre statine non concernée par ce type d'interaction ». La différence de niveau de contrainte est ici liée à une condition, la posologie de la simvastatine. Une hypothèse est que cet élément contextuel a été volontairement omis dans la source ONC pour forcer les cliniciens à choisir une autre statine et éviter tout risque de rhabdomyolyse. D'après la

1. ONC renvoie vers la liste des médicaments torsadogènes décrite par CredibleMeds

source commerciale Multum c'est le niveau le plus élevé (niveau 3) mais la recommandation est d'ajuster les doses.

Cette approche pragmatique de comparaison des sources est cependant limitée par l'absence de médicaments hospitaliers dans les données.

Fréquence des contre-indications A partir d'un échantillon du dossier pharmaceutique, nous trouvons en moyenne 32% de délivrances présentant un risque d'interaction parmi les "délivrances à risque", c'est-à-dire celles contenant au moins deux médicaments avec un risque potentiel d'interaction. Ces dernières représentent 50% des 7 millions de délivrances présentes dans les données.

Le nombre de délivrances contenant une contre-indication était de 17 112 soit 0.24% (24 pour 10 000) du nombre total de délivrances. La représentativité de ces résultats est limitée à la zone géographique (huits départements anonymisés), à la période juin à août 2013 et à des médicaments dispensés en ville. Les interactions seraient plus fréquentes à l'hôpital car les traitements y sont plus nombreux et complexes [56].

Une analyse similaire [57] avait été réalisée en 1999 dans le Nord-Pas-de-Calais sur une période de trois mois, sur des données de remboursement en ville. Les auteurs dénombrent 14 390 contre-indications sur plus de 5 millions de délivrances (27 pour 10 000) en utilisant le référentiel des interactions de l'ANSM (AFSSAPS).

Tobi et al. [58] ont proposé des définitions pour la co-prescription, la médication concomittante et la co-médication. La co-prescription est définie comme la présence de plusieurs médicaments sur une même ordonnance. La médication concomittante correspond aux médicaments prescrits par un ou plusieurs médecins dans une même période. La co-médication est la prise de plusieurs médicaments par le patient au même moment respectant ou non les prescriptions médicales. Notre analyse s'est limitée à la co-prescription car nous n'avions pas d'identifiant patient pour analyser la médication concomittante.

Gagne et al. [59] ont analysé la médication concomittante à partir de la liste d'interactions de Malone et al. [60] contenant seulement 25 couples et une base de données régionale italienne de remboursement de 2004. Sur une période d'un an, ils ont dénombré 8894 risques d'interaction parmi 7902 individus. Zwart-van Rijkom et al. [56] ont analysé les prescriptions dans un hôpital danois sur une période d'un an avec leur référentiel national contenant 331 couples. Sur 21 277 patients qui ont reçu au moins un médicament, 5909 (27,8%) présentaient un risque d'interaction médicamenteuse. Parmi 208 187 prescriptions, 20 058 (9,6%) contenaient un couple de molécules connu pour interagir.

Ces résultats montrent la faisabilité d'utiliser le format ouvert du thesaurus pour dénombrer des interactions médicamenteuses à partir de données de prescription ou de délivrance et nous proposons un algorithme basé sur les méthodes de recherche d'information pour réaliser ce dénombrement. Une étude de plus grande envergure menée avec le service de pharmacovigilance du CHU de Bordeaux, prenant en compte la médication concomittante à partir des

données du SNIIRAM, est en cours au moment de l'écriture de ce travail.

Fréquence des alertes Phansalkar et al. [12] ont analysé les fichiers journaux [2] pour identifier les alertes fréquentes et ignorées. Envoyer un excès d'information peut en effet être contre-productif et surcharger le travail des cliniciens. Les 50 couples les plus fréquents généraient plus de 50% des alertes et celles-ci étaient ignorées dans 95% des cas. Ces alertes avaient conduit l'auteur à se poser la question de les désactiver. Dans notre échantillon, 208 couples ont généré 50% des alertes.

La surcharge cognitive est un problème multifactoriel, notre étude ne permet pas d'analyser l'utilisation du thesaurus, intégré dans un LAP, en situation réelle. Cependant, certains couples de molécules peuvent générer plusieurs alertes (jusqu'à 5) et donc favoriser la surcharge cognitive du professionnel de santé.

2. log en anglais, désigne l'enregistrement des évènements d'un programme

Conclusion

Nous proposons dans ce travail un format ouvert du thesaurus des interactions médicamenteuses de l'ANSM qui utilise les standards et les technologies du web sémantique. Les données ont été liées vers l'UMLS, DrugBank et le répertoire du médicament pour enrichir les connaissances, faciliter la recherche d'information et offrir la possibilité de développer de nouvelles applications. Ces liens ont permis de comparer le référentiel national avec d'autres sources sur les interactions et d'analyser des délivrances de médicaments. Ces dernières analyses fournissent des retours d'information intéressants sur le contenu du thesaurus et sur les alertes qu'il génère lorsqu'il est intégré dans un logiciel d'aide à la prescription.

La représentation informatique des connaissances sur les interactions médicamenteuses nécessite d'être améliorée pour prendre en compte les éléments contextuels et fournir des informations claires et spécifiques aux cliniciens.

Annexes

Ontologie

Développement

L'ontologie contient seulement trois classes : DDI (drug-drug interaction), Famille et Molécule. La classe DDI possède trois 'data property' dont la portée (range) est une chaîne de caractère (xsd:string) : le niveau de l'interaction, sa description et son mécanisme.

Les relations entre ressources (object property) sont les suivantes :

— 'contient' fait le lien entre une famille et une molécule ou entre deux familles. La relation 'appartient' est sa relation inverse. On dit qu'une famille contient une autre famille ou une molécule. La relation est transitive : si une famille A contient une famille B qui contient une famille C alors la famille A contient la famille C.

— 'participe1' et 'participe2' font le lien entre une interaction et les protagonistes 1 et 2 de cette interaction respectivement. Deux relations ont été créées pour définir la notion de participation à une interaction. La raison de ce choix est liée à une erreur de raisonnement lorsqu'on crée une seule relation 'participe' : le raisonneur infère que toute instance participant à une interaction 'interagit' avec lui-même. Or, dans le thesaurus, certaines classes interagissent avec elle-même mais pas toutes. Comme la relation 'interagit' est une 'SuperProperty Of (Chain)', il est impossible de la définir 'irreflexive'. Ce problème est similaire à la difficulté de définir la relation 'hasBrother' entre deux individus sans inférer qu'un individu 'hasBrother' lui-même [1].

— 'interagit' fait le lien d'interaction. La relation est définie par les quatre compositions suivantes :

1. participe1 o invparticipe2
2. participe1 o invparticipe2 o contient
3. appartient o participe1 o invparticipe2
4. appartient o participe1 o invparticipe2 o contient

Le rond signifie "suivi de". La relation est créée si le chemin entre deux instances existe. La figure A.1 représente graphiquement les différents chemins qui seront empruntés par un raisonneur pour créer la relation "interagit" entre deux molécules, deux classes ou entre une classe et une molécule. C'est une relation symétrique : si A interagit avec B alors B interagit avec A.

1. http://stackoverflow.com/questions/19559651

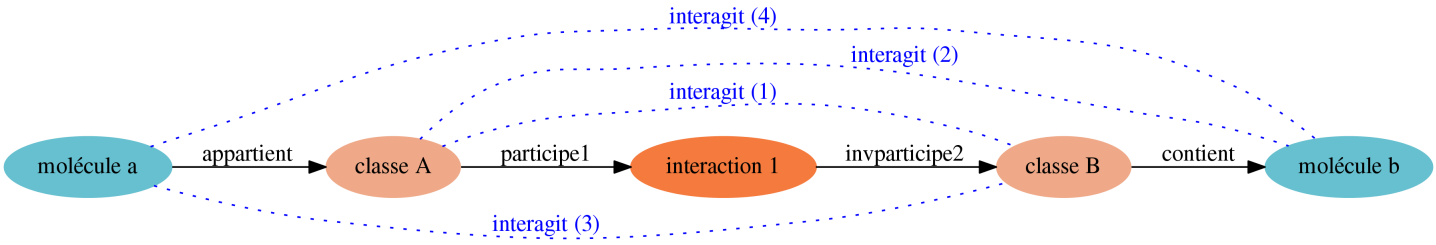

FIGURE A.1 – La relation symétrique "interagit" est créée par un raisonneur si un des chemins, défini dans l'ontologie, existe entre deux instances. Ce schéma représente les différents cas de figure où la relation "interagit" est créée. Entre parenthèses figure le numéro de la composition définie plus haut.

Hiérarchie de familles

Nous souhaitons représenter les classes thérapeutiques du thesaurus comme des instances de la classe Famille. Or, il est nécessaire de les définir comme des classes pour qu'un raisonneur puisse infèrer les relations hiérarchiques 'rdfs:subClassOf'. Nous procédons donc en deux étapes. A la première étape, les familles sont définies comme des classes OWL dans un fichier temporaire. Le raisonneur HermiT est lancé en ligne de commande et infère les relations 'rdfs:subClassOf' entre les familles du thesaurus. A la deuxième étape, les familles sont définies comme des instances de la classe Famille. Les relations 'rdfs:subClassOf' inférées par HermiT plus haut sont transformées en relation ':appartient'. La hiérarchie de familles est ainsi présente dans notre ontologie au travers de notre relation ':contient' (ou sa relation inverse ':appartient') et les classes thérapeutiques du thesaurus sont représentées comme des instances et non des classes OWL.

Pour qu'un raisonneur puisse inférer les relations 'rdfs:subClassOf' entre les classes thérapeutiques du thesaurus à la première étape, il faut faire l'hypothèse du monde fermé. Or, le langage OWL fait l'hypothèse d'un monde ouvert par défaut. Sous cette hypothèse, l'ensemble des individus appartient à un univers ouvert : un auteur n'énumère pas forcément tous les individus d'une classe, soit parce qu'il n'est pas exhaustif, soit parce qu'il ne connaît pas l'ensemble des individus ou que l'ensemble des individus n'est pas énumérable. Or, dans notre cas, nous souhaitons raisonner sous l'hypothèse d'un monde fermé car les auteurs du thesaurus cherchent à énumérer l'ensemble des molécules pour chacune des classes thérapeutiques. Pour raisonner sous cette dernière hypothèse, il est nécessaire de définir les classes thérapeutiques du thesaurus comme des classes dites "énumérées" ou "fermées" en langage OWL. Le raisonneur HermiT peut ainsi inférer les relations de subsomption entre les familles et affirmer par exemple que la classe des diurétiques de l'anse est 'rdfs:subClassOf' de la classe des diurétiques.

Instanciation et inférence

Un programme prend en paramètre la version d'un thesaurus et instancie notre ontologie. L'ontologie est ensuite ouverte avec le logiciel Protégé™. Nous vérifions les résultats des inférences réalisées par un raisonneur intégré dans l'application. Il est possible de demander les explications d'une inférence. Par exemple, pour l'inférence "diurétiques de l'anse" 'interagit' avec "anti-inflammatoires non stéroïdiens", l'explication du raisonneur est la suivante :

- **Car** "diurétiques de l'anse" ':appartient' "diurétiques des hypokaliémants"
- **Car** "diurétiques des hypokaliémants" ':appartient' "diurétiques"
- **Car** ':appartient' est une relation transitive
    - **Donc** "diurétiques de l'anse" ':appartient' "diurétiques"
- **Car** "anti-inflammatoires non stéroïdiens" ':participe1' "anti-inflammatoires non stéroïdiens interagit avec diurétiques" (libellé de l'interaction créé dans le programme R)
- **Car** "diurétiques" ':participe2' "anti-inflammatoires non stéroïdiens interagit avec diurétiques"
- **Car** ':invparticipe2' est l'inverse de ':participe2'
- **Car** ':interagit' correspond à la composition : ':appartient' suivi de ':participe1' suivi de ':invparticipe2'
    - **Donc** "diurétiques de l'anse" ':interagit' "anti-inflammatoires non stéroïdiens"

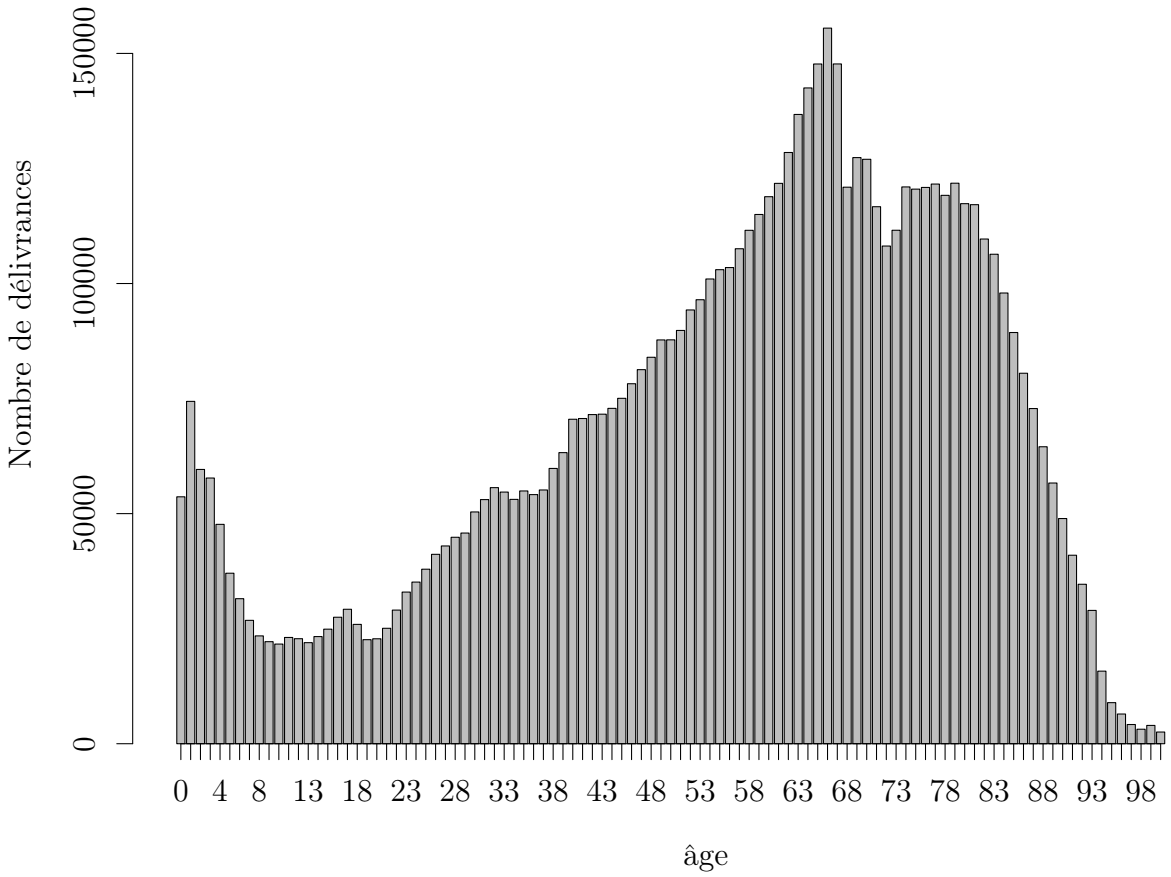

FIGURE A.2 – Distribution de l'âge des patients dans les données de délivrance.

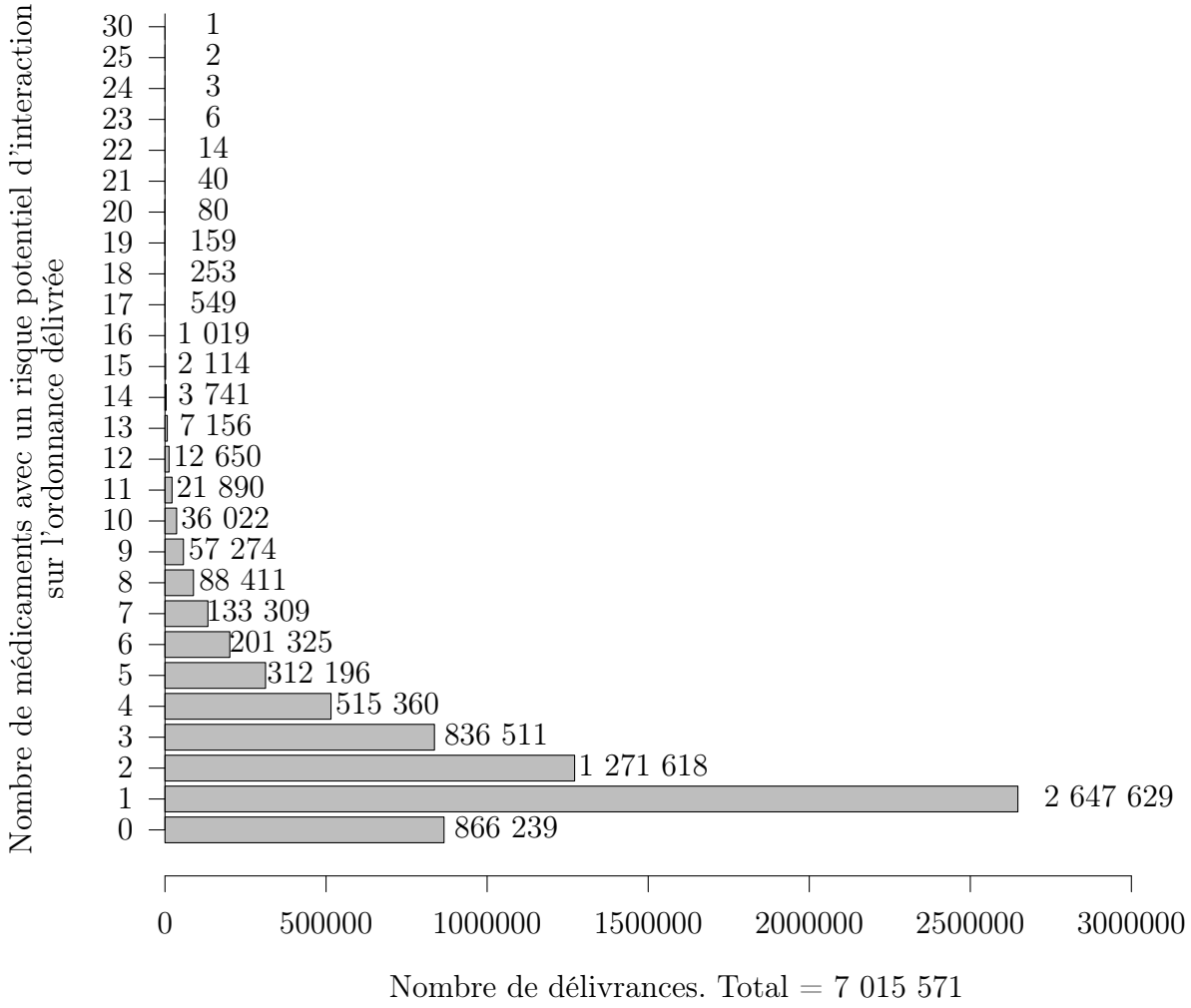

FIGURE A.3 – Nombre de délivrances en fonction du nombre de médicaments avec un risque potentiel d'interaction.

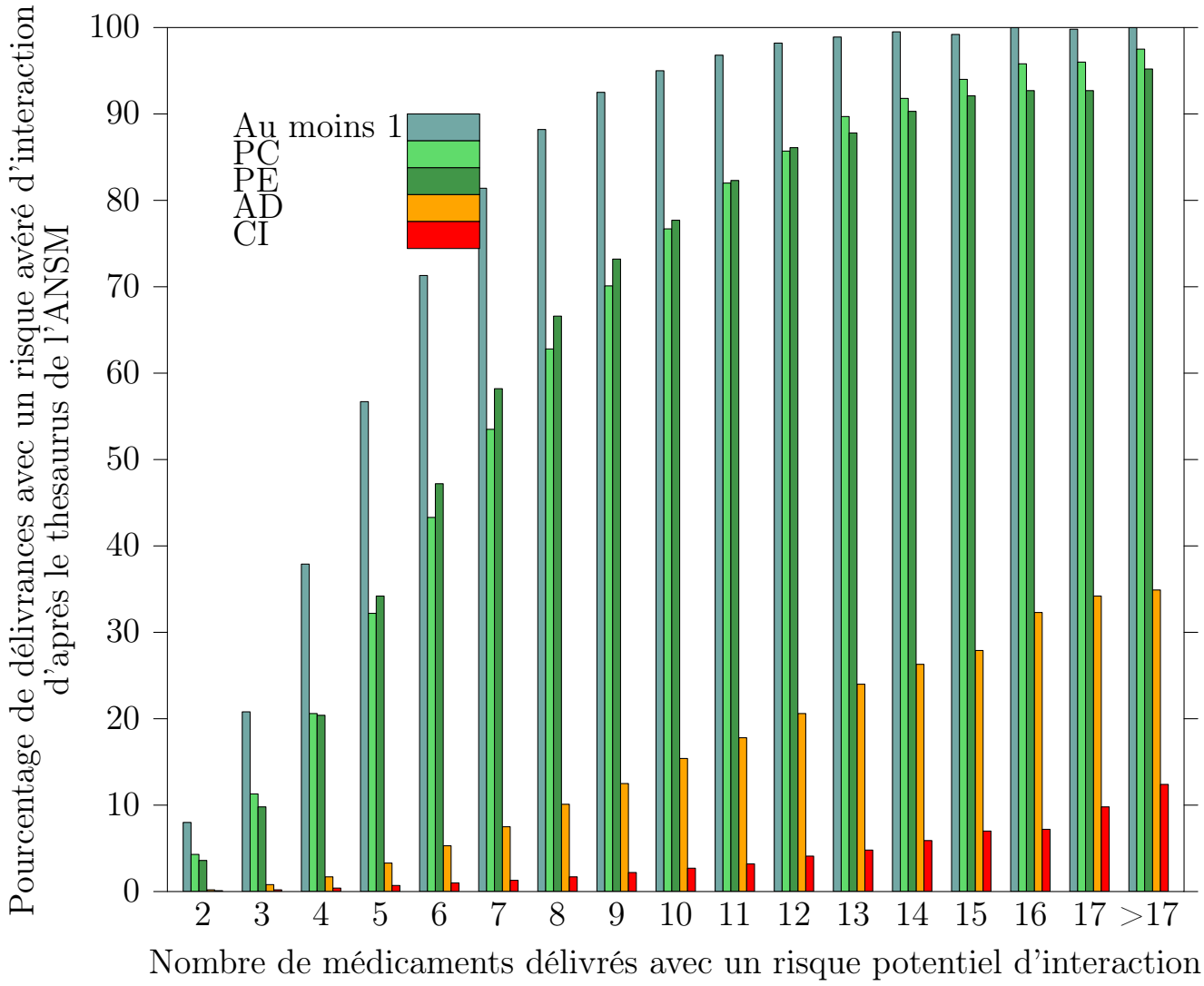

FIGURE A.4 – Fréquence du risque d'interaction médicamenteuse d'après le thesaurus des interactions de l'ANSM en fonction du nombre de médicaments à risque potentiel d'interaction sur l'ordonnance délivrée.

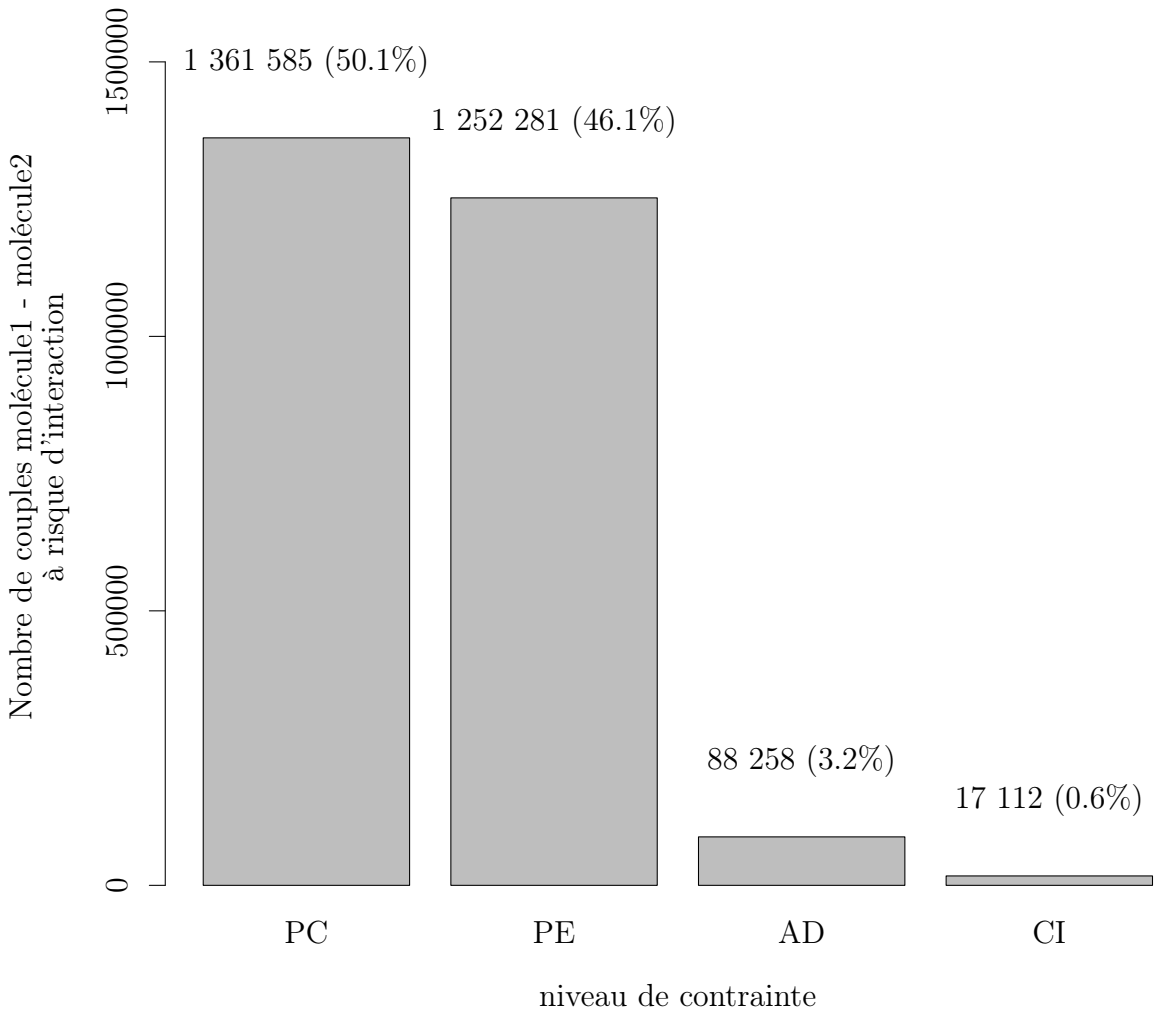

FIGURE A.5 – Niveau de contrainte des couples de molécules à risque d'interaction dénombrés par notre programme, utilisant le format ouvert du thesaurus des interactions de l'ANSM, sur des délivrances médicamenteuses entre juin et août 2013. PC : à prendre en compte, PE : précaution d'emploi, AD : association déconseillée, CI : contre-indication

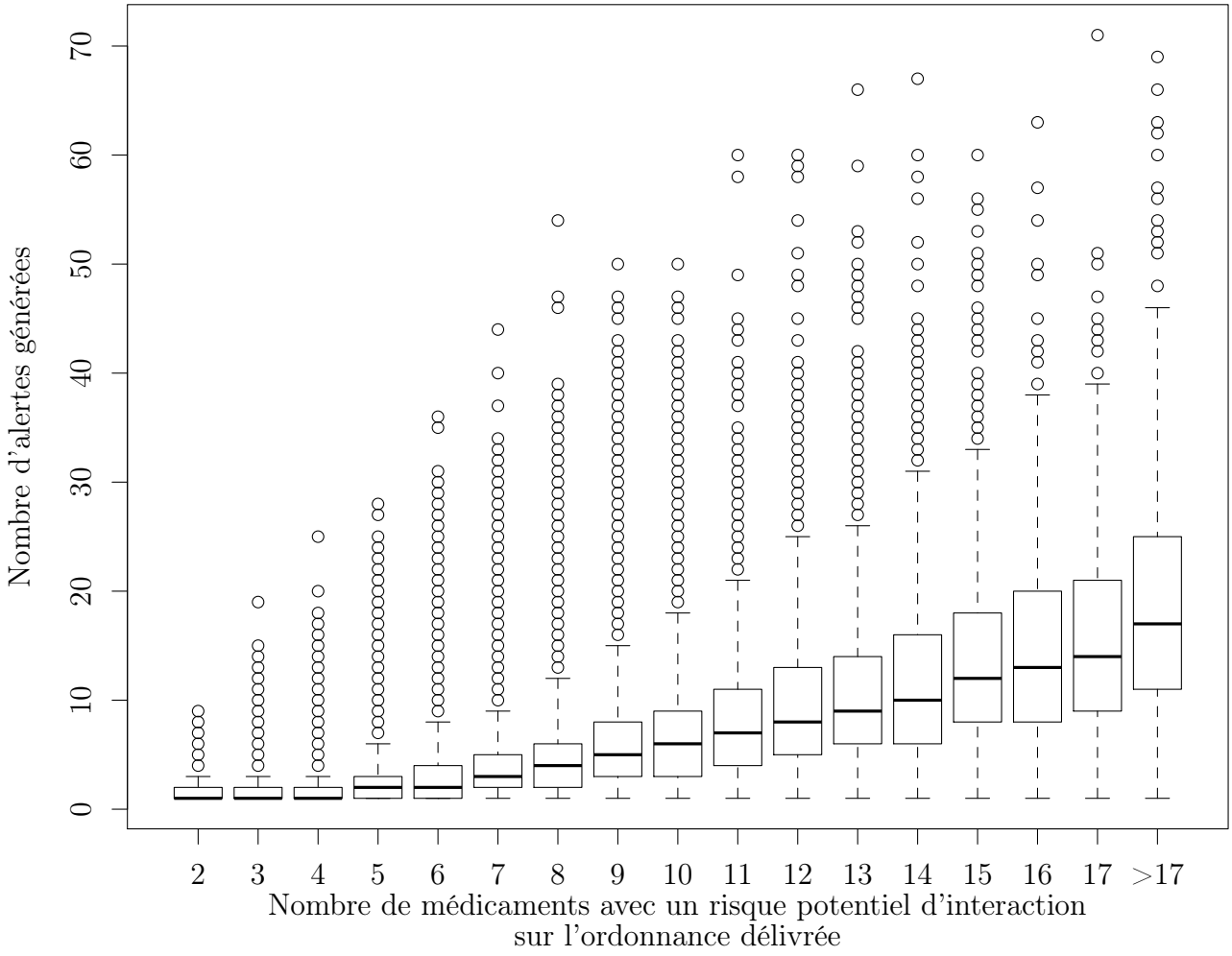

FIGURE A.6 – Nombre d'alertes générées parmi les délivrances comportant au moins un risque d'interaction médicamenteuse en fonction du nombre de médicaments à risque potentiel d'interaction sur l'ordonnance délivrée

Bibliographie

[1] Sultana J, Cutroneo P, Trifirò G. Clinical and economic burden of adverse drug reactions. Journal of Pharmacology & Pharmacotherapeutics. 2013 Dec ;4(Suppl1):S73–S77. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3853675/.

[2] McDonnell PJ, Jacobs MR. Hospital admissions resulting from preventable adverse drug reactions. The Annals of Pharmacotherapy. 2002 Sep ;36(9):1331–1336.

[3] Scheife RT, Hines LE, Boyce RD, Chung SP, Momper JD, Sommer CD, et al. Consensus recommendations for systematic evaluation of drug-drug interaction evidence for clinical decision support. Drug Safety. 2015 Feb ;38(2):197–206.

[4] Mathieu N. Thèse de pharmacie - Interactions médicamenteuses : de la théorie à la réalité. Faculté de Nancy ; 2008.

[5] Ansari J. Drug Interaction and Pharmacist. Journal of Young Pharmacists : JYP. 2010 ;2(3):326–331. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2964764/.

[6] Routledge PA, O'Mahony MS, Woodhouse KW. Adverse drug reactions in elderly patients. British Journal of Clinical Pharmacology. 2004 Feb ;57(2):121–126. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1884428/.

[7] van Roon EN, Flikweert S, le Comte M, Langendijk PNJ, Kwee-Zuiderwijk WJM, Smits P, et al. Clinical relevance of drug-drug interactions : a structured assessment procedure. Drug Safety. 2005 ;28(12):1131–1139.

[8] Ayvaz S, Horn J, Hassanzadeh O, Zhu Q, Stan J, Tatonetti NP, et al. Toward a complete dataset of drug-drug interaction information from publicly available sources. Journal of Biomedical Informatics. 2015 Jun ;55:206–217.

[9] Banda JM, Kuhn T, Shah NH, Dumontier M. Provenance-Centered Dataset of Drug-Drug Interactions. arXiv:150705408 [cs]. 2015 Jul ;ArXiv: 1507.05408. Available from: http://arxiv.org/abs/1507.05408.

[10] Rapport de la Commission Open Data en santé ;. Available from: http://drees.social-sante.gouv.fr/IMG/pdf/rapport_final_commission_open_data-2.pdf.

[11] CNAMTS - DDRI - DSM : onze associations médicamenteuses formellement contre-indiquées ;. Available from: http://www.urps-ml-paysdelaloire.fr/APIMED/uploads/pdf/Prescription%20m%C3%A9dicamenteuse%20chez%20la%20personne%20%C3%A2g%C3%A9e/CNAM-2003.asso.med.CI.pdf.

[12] Phansalkar S, van der Sijs H, Tucker AD, Desai AA, Bell DS, Teich JM, et al. Drug-drug interactions that should be non-interruptive in order to reduce alert fatigue in electronic

health records. Journal of the American Medical Informatics Association: JAMIA. 2013 May ;20(3):489–493.

[13] Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. Journal of Biomedical Informatics. 2008 Oct ;41(5):706–716.

[14] Berners-Lee T, Hendler J, Lassila O. The Semantic Web ; 2001.

[15] Robu I, Robu V, Thirion B. An introduction to the Semantic Web for health sciences librarians. Journal of the Medical Library Association. 2006 Apr ;94(2):198–205. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1435839/.

[16] Gruber TR. A translation approach to portable ontology specifications. Knowledge Acquisition. 1993 Jun ;5(2):199–220. Available from: http://www.sciencedirect.com/science/article/pii/S1042814383710083.

[17] Samwald M, Jentzsch A, Bouton C, Kallesøe CS, Willighagen E, Hajagos J, et al. Linked open drug data for pharmaceutical research and development. Journal of Cheminformatics. 2011 ;3(1):19.

[18] Ratnam J, Zdrazil B, Digles D, Cuadrado-Rodriguez E, Neefs JM, Tipney H, et al. The application of the open pharmacological concepts triple store (open PHACTS) to support drug discovery research. PloS One. 2014 ;9(12):e115460.

[19] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research. 2004 Jan ;32(Database issue):D267–270.

[20] 2016AA UMLS® Release Notes and Bugs [Technical Documentation] ;. Available from: https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/notes.html.

[21] Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: RxNorm at 6 years. Journal of the American Medical Informatics Association : JAMIA. 2011 ;18(4):441–448. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3128404/.

[22] Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Research. 2006 Jan ;34(Database issue):D668–672.

[23] Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, et al. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Research. 2014 Jan ;42(Database issue):D1091–D1097. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3965102/.

[24] Avis aux titulaires d'autorisation de mise sur le marché de médicaments à usage humain et aux pharmaciens responsables des établissements pharmaceutiques mentionnés à l'article R. 5124-2 CSP ;.

[25] Rapport sur la gouvernance et l'utilisation des données de santé ;. Available from: http://social-sante.gouv.fr/IMG/pdf/Rapport_donnees_de_sante_2013.pdf.

[26] Haute Autorité de Santé - Agrément des Bases de données sur les Médicaments ;. Available from: http://www.has-sante.fr/portail/jcms/c_672761/fr/agrement-des-bases-de-donnees-sur-les-medicaments.

[27] Rubrichi S, Quaglini S, Spengler A, Russo P, Gallinari P. A system for the extraction and representation of summary of product characteristics content. Artificial Intelligence in Medicine. 2013 Feb ;57(2):145–154.

[28] Peters LB, Bahr N, Bodenreider O. Evaluating drug-drug interaction information in NDF-RT and DrugBank. Journal of Biomedical Semantics. 2015 May ;6(1):19. Available from: http://link.springer.com.docelec.u-bordeaux.fr/article/10.1186/s13326-015-0018-0.

[29] 2016AA UMLS® Release Notes and Bugs [Release Notes Documentation] ;. Available from: http://evs.nci.nih.gov/ftp1/NDF-RT/ReadMe.txt.

[30] Phansalkar S, Desai AA, Bell D, Yoshida E, Doole J, Czochanski M, et al. High-priority drug–drug interactions for use in electronic health records. Journal of the American Medical Informatics Association : JAMIA. 2012 ;19(5):735–743. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3422823/.

[31] Crowther NR, Holbrook AM, Kenwright R, Kenwright M. Drug interactions among commonly used medications. Chart simplifies data from critical literature review. Canadian Family Physician. 1997 Nov ;43:1972–1981. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2255205/.

[32] Takarabe M, Shigemizu D, Kotera M, Goto S, Kanehisa M. Network-based analysis and characterization of adverse drug-drug interactions. Journal of Chemical Information and Modeling. 2011 Nov ;51(11):2977–2985.

[33] Tatonetti NP, Ye PP, Daneshjou R, Altman RB. Data-driven prediction of drug effects and interactions. Science Translational Medicine. 2012 Mar ;4(125):125ra31.

[34] Boyce R, Collins C, Horn J, Kalet I. Computing with evidence Part II: An evidential approach to predicting metabolic drug-drug interactions. Journal of Biomedical Informatics. 2009 Dec ;42(6):990–1003.

[35] Kilicoglu H, Shin D, Fiszman M, Rosemblat G, Rindflesch TC. SemMedDB: a PubMed-scale repository of biomedical semantic predications. Bioinformatics. 2012 Dec ;28(23):3158–3160. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3509487/.

[36] Dusetzina SB, Tyree S, Meyer AM, Meyer A, Green L, Carpenter WR. Linking Data for Health Services Research: A Framework and Instructional Guide. AHRQ Methods

for Effective Health Care. Rockville (MD): Agency for Healthcare Research and Quality (US) ; 2014. Available from: http://www.ncbi.nlm.nih.gov/books/NBK253313/.

[37] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria ; 2015. Available from: https://www.R-project.org/.

[38] van Hage WR, with contributions from: Tomi Kauppinen, Graeler B, Davis C, Hoeksema J, Ruttenberg A, et al.. SPARQL: SPARQL client ; 2013. R package version 1.16. Available from: https://CRAN.R-project.org/package=SPARQL.

[39] Wickham H. stringr: Simple, Consistent Wrappers for Common String Operations ; 2015. R package version 1.0.0. Available from: https://CRAN.R-project.org/package=stringr.

[40] Sharpsteen C, Bracken C. tikzDevice: R Graphics Output in LaTeX Format ; 2016. R package version 0.10-1. Available from: https://CRAN.R-project.org/package=tikzDevice.

[41] Wickham H, Chang W. devtools: Tools to Make Developing R Packages Easier ; 2016. R package version 1.10.0. Available from: https://CRAN.R-project.org/package=devtools.

[42] Wickham H, Danenberg P, Eugster M. roxygen2: In-Source Documentation for R ; 2015. R package version 5.0.1. Available from: https://CRAN.R-project.org/package=roxygen2.

[43] RStudio Team. RStudio: Integrated Development Environment for R. Boston, MA ; 2015. Available from: http://www.rstudio.com/.

[44] Shearer R, Motik B, Horrocks I. HermiT: A highly-efficient OWL reasoner. In: ResearchGate. vol. 432 ; 2008. Available from: https://www.researchgate.net/publication/221218516_HermiT_A_highly-efficient_OWL_reasoner.

[45] MCIA. Mésocentre de Calcul Intensif Aquitain ;. [Online ; accessed 26-Aout-2016]. http://www.mcia.univ-bordeaux.fr/index.php?id=56.

[46] Ram K. Git can facilitate greater reproducibility and increased transparency in science. Source Code for Biology and Medicine. 2013 Feb ;8(1):7.

[47] Bulletin d'information du centre régional de PHARMACOVIGILANCE et d'INFORMATION sur le MEDICAMENT de BORDEAUX ;. Available from: http://www.pharmacologie.u-bordeaux2.fr/documents/pharmacovigilance/INFOS_pdf/117_2016-08.pdf.

[48] Berthelot H, Simon G, Toussi M. Analyse de la sécurité d'utilisation des bases de données médicamenteuses, des logiciels d'aide à la prescription et à la dispensation. Pharcorama. 2015 Mar ;.

[49] Yoshikawa S, Satou K, Konagaya A. Drug interaction ontology (DIO) for inferences of possible drug-drug interactions. Studies in Health Technology and Informatics. 2004 ;107(Pt 1):454–458.

[50] Herrero-Zazo M, Segura-Bedmar I, Hastings J, Martínez P. DINTO: Using OWL Ontologies and SWRL Rules to Infer Drug-Drug Interactions and Their Mechanisms. Journal of Chemical Information and Modeling. 2015 Aug ;55(8):1698–1707.

[51] Le MeSH Bilingue anglais - français ;. Available from: http://mesh.inserm.fr/mesh/.

[52] Tilson H, Hines LE, McEvoy G, Weinstein DM, Hansten PD, Matuszewski K, et al. Recommendations for selecting drug-drug interactions for clinical decision support. American journal of health-system pharmacy: AJHP: official journal of the American Society of Health-System Pharmacists. 2016 Apr ;73(8):576–585.

[53] Wang LM, Wong M, Lightwood JM, Cheng CM. Black box warning contraindicated comedications: concordance among three major drug interaction screening programs. The Annals of Pharmacotherapy. 2010 Jan ;44(1):28–34.

[54] Hazlet TK, Lee TA, Hansten PD, Horn JR. Performance of community pharmacy drug interaction software. Journal of the American Pharmaceutical Association (Washington,DC: 1996). 2001 Apr ;41(2):200–204.

[55] Brochhausen M, Schneider J, Malone DC, Empey PE, Hogan WR, Boyce RD. Towards a foundational representation of potential drug-drug interaction knowledge ; 2014. .

[56] Zwart-van Rijkom JEF, Uijtendaal EV, ten Berg MJ, van Solinge WW, Egberts ACG. Frequency and nature of drug–drug interactions in a Dutch university hospital. British Journal of Clinical Pharmacology. 2009 Aug ;68(2):187–193.

[57] Guédon-Moreau L, Ducrocq D, Duc MF, Quieureux Y, L'Hôte C, Deligne J, et al. Absolute contraindications in relation to potential drug interactions in outpatient prescriptions: analysis of the first five million prescriptions in 1999. European Journal of Clinical Pharmacology. 2004 Feb ;59(12):899–904.

[58] Tobi H, Faber A, van den Berg PB, Drane JW, de Jong-van den Berg LTW. Studying comedication patterns: the impact of definitions. Pharmacoepidemiology and Drug Safety. 2007 Apr ;16(4):405–411.

[59] Gagne JJ, Maio V, Rabinowitz C. Prevalence and predictors of potential drug-drug interactions in Regione Emilia-Romagna, Italy. Journal of Clinical Pharmacy and Therapeutics. 2008 Apr ;33(2):141–151.

[60] Malone DC, Abarca J, Hansten PD, Grizzle AJ, Armstrong EP, Van Bergen RC, et al. Identification of serious drug-drug interactions: results of the partnership to prevent drug-drug interactions. Journal of the American Pharmacists Association: JAPhA. 2004 Apr ;44(2):142–151.