Une infrastructure générique basée sur les apports du Web Sémantique pour l'analyse des bases médico-administratives

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

1 Département Epidémiologie et biostatistique, EHESP, Sorbonne Paris Cite, France.
{yann.rivault, nolwenn.lemeur}@ehesp.fr
2 Université Rennes 1, IRISA UMR6074, Rennes.
olivier.dameron@univ-rennes1.fr
3 EHESP, EA 7348 MOS Management des organisations en santé, France.

Résumé : Suite aux difficultés de gestion et d'exploration de données issues des bases médico-administratives françaises, une méthode basée sur les technologies du Web Sémantique a été mise en place. L'objectif est d'être en mesure de gérer et d'explorer des données de parcours de soins de ville et d'hôpital, de manière pertinente et efficace, pour faciliter l'analyse statistique. Dans le cadre de cette étude, l'analyse porte sur la description et la prévention des évènements indésirables suite à une opération chirurgicale prise en charge en ambulatoire. A terme une infrastructure profitant à toute étude statistique portant sur une analyse de ce genre de données sera proposée.

Mots-clés : Web Sémantique, PMSI, SNIIRAM, Ontologie.

1 Introduction

Malgré une finalité à prédominance administrative ou tarifaire des principaux systèmes d'information médico-administratifs, de nombreuses études ont montré leur utilité dans l'analyse statistique médicale comme le présente le Bulletin d'Epidémiologie Hebdomadaire dans son hors-série de décembre 2013[1]. Ainsi, le Programme de Médicalisation des systèmes d'information (PMSI) est un dispositif permettant initialement de mesurer l'activité et les ressources des établissements hospitaliers français, afin de calculer leur allocation budgétaire. Le Système National d'Information Inter-Régimes de l'Assurance Maladie (SNIIRAM) a quant à lui pour objectif la gestion des prestations et actes médicaux remboursés par l'Assurance Maladie. Plus récemment ces systèmes d'information ont été utilisés pour étudier des trajectoires de soins[2][3]. Cependant ces études portaient sur l'un ou l'autre des systèmes d'information et/ou sur un nombre restreint de pathologies ou de traitements médicamenteux[4][5].

Nous proposons une infrastructure intégrant les données hospitalières (PMSI) et les données de soins de ville (SNIIRAM) pouvant supporter une étude nécessitant de combiner les deux types de données. Nous illustrons cette infrastructure par un cas d'étude portant sur l'analyse des parcours des soins de ville pour décrire et prévenir la survenue d'évènements indésirables liés aux interventions chirurgicales pouvant être prises en charge en ambulatoire. Les difficultés principales rencontrées dans un tel contexte sont d'une part la grande quantité des données hétérogènes (grand nombre de variables et de patients) et d'autre part le besoin d'agréger certaines données selon des degrés de précision dépendant des études (par exemple pour identifier les patients souffrant d'un type de maladie et ayant été traités par une classe de médicaments). Pour pallier à cette dernière difficulté, le projet HAIKU[6], de l'équipe de

D.BUCKERIDGE de L'université McGill de Montréal, propose l'utilisation des techniques du Web Sémantique.

Sur cette base, l'infrastructure que nous proposons pour intégrer SNIIRAM et PMSI afin d'optimiser les analyses de trajectoires de soins « ville-hôpital » reposera sur l'utilisation de technologies empruntées au Web Sémantique.

2 Méthodes

2.1 Infrastructure basée sur le Web Sémantique

Ce paragraphe détaille les différentes technologies du Web Sémantique utilisées dans ce projet.

Le *Resource Description Framework* (RDF)[1] est un formalisme permettant à la fois de représenter des données ainsi que les associations entre ces données, et d'intégrer des données issues de sources différentes. Il est bien adapté à la représentation de données massives et à l'intégration de sources de données variées, comme le montre la *linked data initiative*[2], qui répertoriait en Août 2014, 570 bases de données au format RDF, pouvant contenir pour certaines plusieurs milliards de triplets. Ces bases couvrent des domaines très variés mais les sciences de la vie en représentent une part importante.

RDF Schema (RDFS)[3] est une extension de RDF permettant de représenter des connaissances sous forme d'ontologies en décrivant des hiérarchies de classes. Le *Web Ontology Language* (OWL)[4] est lui-même une extension de RDFS permettant de représenter les propriétés de ces classes. Cela permet d'agréger les données selon des catégories de plus en plus générales. La plupart des ontologies sont représentées en OWL (et sont donc automatiquement aussi des ontologies RDFS) même si elles en utilisent rarement la richesse du formalisme (mais en les faisant directement en OWL, on se laisse la possibilité de les enrichir plus tard). Les données en RDF, RDFS et OWL sont typiquement stockées dans des *triplestores*, qui sont l'équivalent des bases de données relationnelles et qui permettent d'interroger les données en tenant compte de la hiérarchie des classes grâce à des requêtes SPARQL[5] (similaires aux requêtes SQL). Il existe de nombreux *triplestores* ; nous avons utilisé Fuseki[6].

2.2 Cas d'étude des complications après chirurgie ambulatoire

Dans ce paragraphe, nous détaillons l'extraction des données des patients du PMSI et du SNIIRAM, leur représentation en RDF et leur lien avec des ontologies.

Les données sont extraites du PMSI et du SNIIRAM du 01/01/2012 au 31/12/2013. Les patients étudiés ont subi un acte chirurgical pouvant être pris en charge en ambulatoire (soit 203 actes)[7]. Les parcours de soins incluent les soins de ville 6 mois avant et 6 mois après l'opération, la chirurgie, et une éventuelle ré-hospitalisation pour complication. Toutes les ré-hospitalisations dans un délai de trois mois après la chirurgie ont été dénombrées. Parmi ces ré-hospitalisations, celles qui correspondent à une complication de la chirurgie antérieure ont

1 http://www.w3.org/TR/rdf11-concepts/
2 http://lod-cloud.net/
3 http://www.w3.org/TR/rdf-schema/
4 http://www.w3.org/TR/owl2-overview/
5 http://www.w3.org/TR/sparql11-overview/
6 http://jena.apache.org/documentation/fuseki2/index.html

été identifiées sur la base des diagnostics principaux et associés (codifiés selon la Classification statistique Internationale des Maladies, version 10 ou CIM-10[7]).

Du fait de leur évolution constante et de leur complexité croissante, la connaissance complète de certaines codifications et nomenclatures médicales est compliquée. Les ontologies relatives à la CIM-10[8] pour les diagnostics médicaux et au Code Identifiant de Prestation (CIP)[9][10] pour les médicaments sont utilisées pour interroger les données selon un niveau hiérarchique d'une classe de maladies, ou d'une classe de médicaments (les antibiotiques par exemple).

Les données de patients extraites du PMSI et du SNIIRAM extraites au format matriciel « individus × variables » ont été transformées au format RDF de manière automatisée par des programmes Python. L'ensemble de ces données ainsi que les ontologies ont été chargées sur un serveur SPARQL Jena Fuseki (Jena 2.13.0)[8] installé localement (à cause des restrictions de confidentialité).

Dans un objectif de comparaison, la dilatation intraluminale d'un accès vasculaire arterioveineux d'un membre avec pose d'endoprothese, par voie arterielle transcutanee (EZAF002 dans la Classification Commune des Actes Médicaux) a été étudiée d'une part par des algorithmes R[9] (R version 3.1.2, 2014/10/31) et d'autre part par des requêtes SPARQL réalisées aussi bien avec R (package *rrdf*[10]) que Jena Fuseki. Ces 2 approches avaient les mêmes finalités à savoir l'identification des patients opérés en ambulatoire ou non, et des patients ré-hospitalisés pour complication (Figure 1).

3 Résultats et discussion

Les données extraites du PMSI 2012 comprennent 2 590 565 patients adultes (âge > 18) pour un total de 15 597 374 actes médicaux dont 5 311 058 chirurgies (potentiellement réalisées en ambulatoire : durée de séjour < 1 jour), soit plus de 2 chirurgies par patient en moyenne. Les données de remboursements de soins de villes de ces 2 590 565 patients sont de l'ordre du milliard de prestation. Explorer et interroger ces données très volumineuses de manière intelligente est un premier pas vers l'analyse statistique. La méthodologie basée sur le Web Sémantique que nous avons mis en place est un moyen d'y parvenir.

Dans un premier temps nous nous sommes intéressés aux données hospitalières. Leur représentation en RDF correspond à plus de 140 millions de triplets. Leur chargement dans le serveur Jena Fuseki a été effectué en 45 minutes avec un ordinateur portable Dell Latitude E6230, 8Go RAM. Ces 140 millions de triplets n'ont pas pu être chargés dans R. L'étude de comparaison entre R et Jena Fuseki a dû être restreinte à un acte chirurgical : la dilatation intraluminale d'un accès vasculaire arterioveineux d'un membre avec pose d'endoprothese, par voie arterielle transcutanee. La représentation en RDF des données de cette étude correspond à 219 320 triplets. Les temps d'exécution des requêtes étaient assez convenables, de l'ordre d'une seconde pour Fuseki et de l'ordre de 5 secondes pour R avec le package *rrdf*. Toutefois, les données étant stockées par R dans la mémoire vive, l'utilisation du format RDF en est limité, et on ne peut interroger des millions de triplets comme le fait Fuseki.

Parallèlement, les algorithmes sous R étaient bien plus complexes (boucles itératives et jointures multiples) et donc beaucoup moins performants que les requêtes SPARQL équivalentes. Les temps d'exécution de ces algorithmes étaient de l'ordre de plusieurs dizaines de secondes.

Les résultats de l'extraction auraient pu être représentés dans une base de données relationnelle, les requêtes SQL auraient été d'un niveau de complexité similaire à nos requêtes SPARQL, mais il aurait été difficile d'effectuer les tâches d'agrégation par transitivité.

7 http://www.atih.sante.fr/cim-10-fr-2015-usage-pmsi
8 http://jena.apache.org/documentation/fuseki2/index.html
9 http://www.r-project.org/
10 https://github.com/egonw/rrdf

L'analyse des données hospitalières pour l'acte choisit dans l'étude comparative nous a permis de dénombrer et comparer la survenue d'évènements indésirables entre une prise en charge en ambulatoire et une hospitalisation longue. La Figure 1 représente la répartition des types d'hospitalisation selon différents critères d'étude : opération ambulatoire vs hospitalisation longue, ré-hospitalisation ou non, complication ou non.

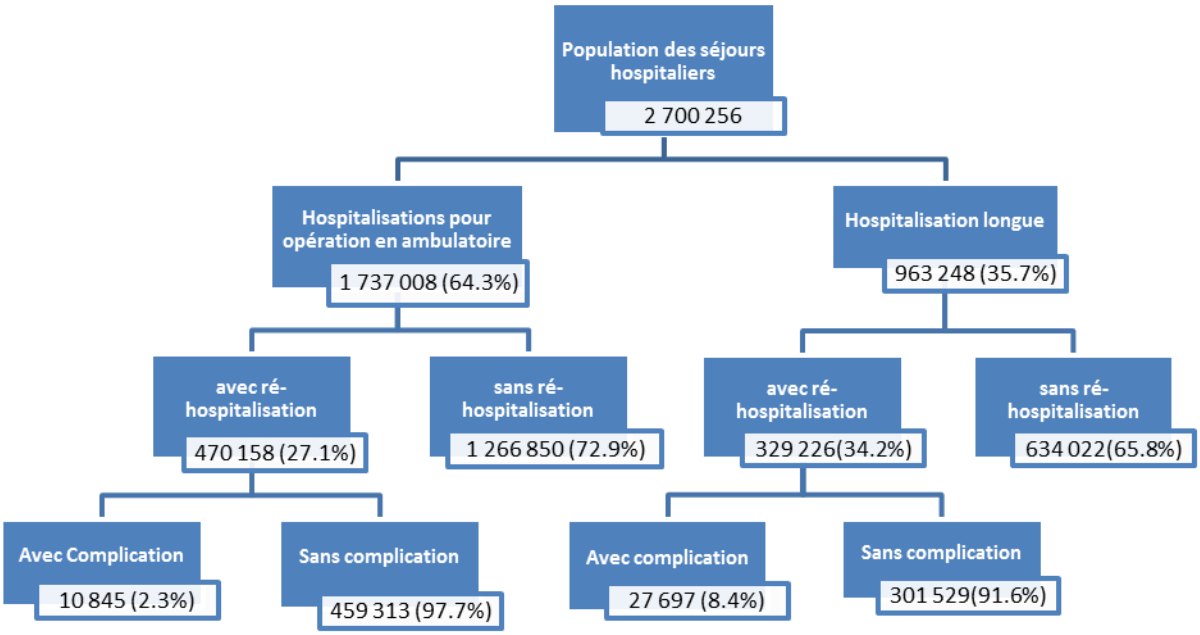

FIGure **1** – Répartition des hospitalisations

Après cette phase descriptive des ré-hospitalisations, nous souhaiterions être en mesure de répondre à des questions plus spécifiques. Un exemple simple serait « Est-ce que les patients ré-hospitalisés pour bactériémie ont reçu un traitement antibiotique prophylactique ? ». A l'origine, les diagnostics des séjours hospitaliers (PMSI) sont codés selon la nomenclature CIM10. Les traitements médicamenteux (SNIIRAM) sont codés selon la nomenclature CIP. L'infrastructure intègre PMSI, SNIIRAM et les ontologies telles que NDF et CIM10 ce qui nous permettra d'appréhender le détail des diagnostics et des médicaments par familles.

4 Conclusion

La méthodologie basée sur le Web Sémantique a apporté des solutions aux différents problèmes rencontrés. L'infrastructure mise en place par cette méthodologie pourrait être par la suite utilisée pour toute autre analyse de données issues du PMSI et du SNIIRAM.

En termes d'application, les sous-populations sélectionnées grâce à des requêtes feront l'objet d'analyses statistiques sous R. Dans un premier temps, nous comparerons les incidences de ré-hospitalisations pour complications suite à une chirurgie, prise en charge en ambulatoire ou non. Dans un second temps nous définirons des algorithmes d'identification

de patients présentant des complications, n'ayant pas été ré-hospitalisés, mais pris en charge par la médecine de ville. Enfin nous essayerons d'identifier les déterminants de telles complications en nous intéressant notamment aux éventuels écarts de suivi des recommandations post-opératoires.

Remerciements

Nous tenons à remercier l'Agence Nationale de sécurité du médicament et des produits de santé et le centre d'expertise PEPS (Plateforme Epidémiologie des Produits de Santé) pour le financement de Yann Rivault.

Références

[1] InVS. (2013).Apports des bases médico-administratives pour l'épidémiologie et la surveillance : regards croisés France-Québec, *Bulletin épidémiologique hebdomadaire*, Décembre 2013, N° Hors-série.

[2] QUANTIN C. et al. (2013). A data mining approach for grouping and analyzing trajectories of care using claim data: the example of breast cancer. *BMC Medical Informatics and Decision Making*.

[3] ALLEMAND H. et al. (2014). Perforations and haemorrhages after colonoscopy in 2012: a study based on comprehensive French helth insurance data (SNIIRAM). *clinics and research in hepatology and gastroenterology journal*.

[4] ALLA F. et al. (2015). Homeopathy in France in 2011-2012 according to reimbursements in the French national health insurance database (SNIIRAM). Fam Pract.

[5] ALLA F. (2015). Compliance with pregnancy prevention plan recommendations in 8672 French women potential exposed to acitretin. Pharmacoepidemiology and drug safety.

[6] L. BUCKERIDGE D. L. et al. (2012). HAIKU : A Semantic Framework for Surveillance of Healthcare-Associated Infections. Procedia Computer Science.

[7] Haute Autorité de Santé. (2012). Ensemble pour le développement de la chirurgie ambulatoire – Socle de connaissances – Synthèse. http://www.has-sante.fr/portail/upload/docs/application/pdf/2012-04/synthese_-_socle_de_connaissances.pdf.

[8] World Health Organization. (2015). International Classification of Diseases, Version 10, http://bioportal.bioontology.org/ontologies/ICD10.

[9] J. LINCOLN MD Michael. (2015). National Drug File - Reference Terminology, http://bioportal.bioontology.org/ontologies/NDFRT.

[10] R. FERGERSON. (2015). Anatomical Therapeutic Chemical Classification, http://bioportal.bioontology.org/ontologies/ATC.