Conversion vers OMOP d'une extraction de 4 millions de patients du SNDS (2008 - 2016)

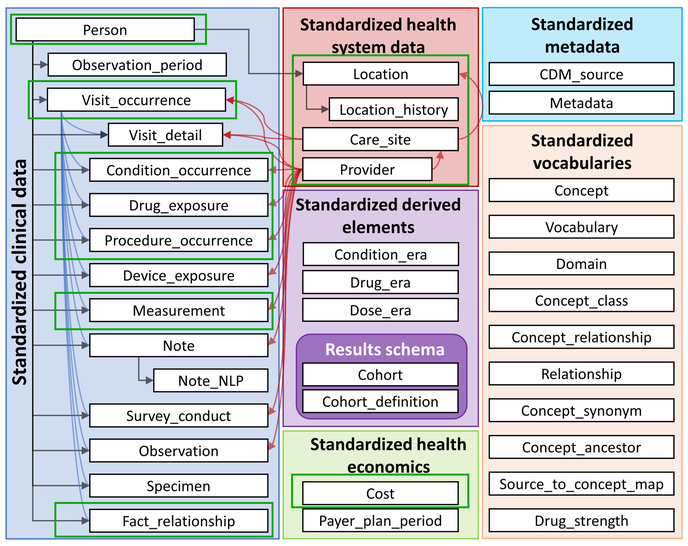
Dinh-Phong Nguyen[1 - 2], Matthieu Doutreligne[1], Adrien Parrot[3], Antoine Lamer[4], Nicolas Paris[3],

(1) Direction de la recherche, des études, de l'évaluation et des statistiques, 75350 Paris 07, France ; (2) Sorbonne Université, faculté de médecine, F-75013 Paris, France ;
(3) AP-HP, Web INnovation Données - Direction des Systèmes d'Information, Paris, France ; (4) Université Lille, CHU Lille, EA2694 - Évaluation des Technologies de Santé et des Pratiques Médicales, F-59000 Lille, France.

Introduction

La consolidation progressive des bases de données médico-administratives nécessite davantage d'interopérabilité et de standardisation. OMOP (Observational Medical Outcomes Partnership) est un modèle de données commun initialement conçu pour faciliter la recherche observationnelle entre plusieurs centres exploitant des données de santé. Le but est de pouvoir partager un même programme d'analyses statistiques qui puisse fonctionner sur tous les sites. A travers le consortium OHDSI (Observational Health Data Sciences and Informatics, [5]) OMOP est désormais utilisé pour des usages très variés tels que le pilotage des systèmes de soins ou l'étude des effets indésirables médicamenteux. Il fournit un cadre structurel et conceptuel s'appuyant sur des terminologies de référence telles que SNOMED pour les diagnostics, RxNorm pour les médicaments et LOINC pour les résultats de laboratoire [6].

Figure 1. Modèle de données OMOP v6.0. En vert, les tables cibles constituées par notre ETL



En proposant une structure de données et une sémantique communes, il permet une portabilité instantanée des outils d'analyse, facilitant la réalisation d'études dans des sites différents. Nous présentons une conversion du Système national des données de santé (SNDS) à ce format dans le cadre d'interCHU, un réseau français de bases de données au format OMOP.

Données

Les données utilisées proviennent d'un échantillon aléatoire du SNDS de 3 160 997 patients âgés de 18 à 120 ans de 2008 à 2016, et couvrant l'ensemble des composantes DCIR et PMSI du SNDS.

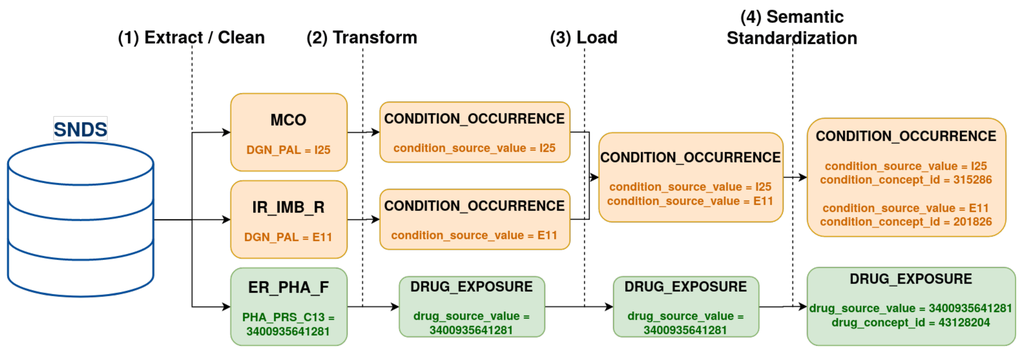
Méthodologie

Nous proposons un alignement en deux temps :

- Structurel : extraction des données et alimentation du schéma de la base de données OMOP à l'aide d'un ETL (Extract-Transform-Load) inspiré de celui de SCALPEL3 [2]. Le code a été validé à hauteur de 82 % à l'aide de tests unitaires.
- Sémantique : conversion des codes sources (CIM10, CCAM, CIP13, ...) en champs standards correspondant aux terminologies internationales de référence (SNOMED, RxNorm, LOINC).

Nous avons mutualisé les connaissances des utilisateurs du SNDS pour établir les spécifications de l'alignement structurel et avons centralisé les correspondances terminologiques existantes sur une plateforme collaborative [3], [4]. Une chaîne de traitement unique a été implémentée alignant 12 des 22 tables cibles cliniques, économiques et d'offre de soins (cf. Figure 1) en s'appuyant sur les technologies à l'état de l'art pour le traitement de données massives (scala, Spark). Cette partie structurelle de l'alignement est schématisée Figure 2 : (1) L'*extract* permet de **lire et de nettoyer les sources** (lignes en erreur, lignes pour information ou doublons sur le PMSI supprimés); (2) le *transform* consiste à **mapper les colonnes** du schéma d'origine vers les colonnes standards d'OMOP; (3) le *load* permet de concaténer les différentes tables sources pour **créer une table OMOP cible**. Le code est modulaire, ce qui permet d'aligner simplement des tables supplémentaires en suivant cette logique (extracteurs, transformeurs, chargement).
La partie sémantique de l'alignement correspond à l'étape (4) Figure 2. Elle a été implémentée grâce à un serveur d'alignement collaboratif permettant de s'accorder sur les correspondances des nomenclatures françaises vers les nomenclatures internationales et de normaliser le processus avec des tables pivots communes.

Figure 2. Schéma de la transformation structurelle et sémantique vers OMOP



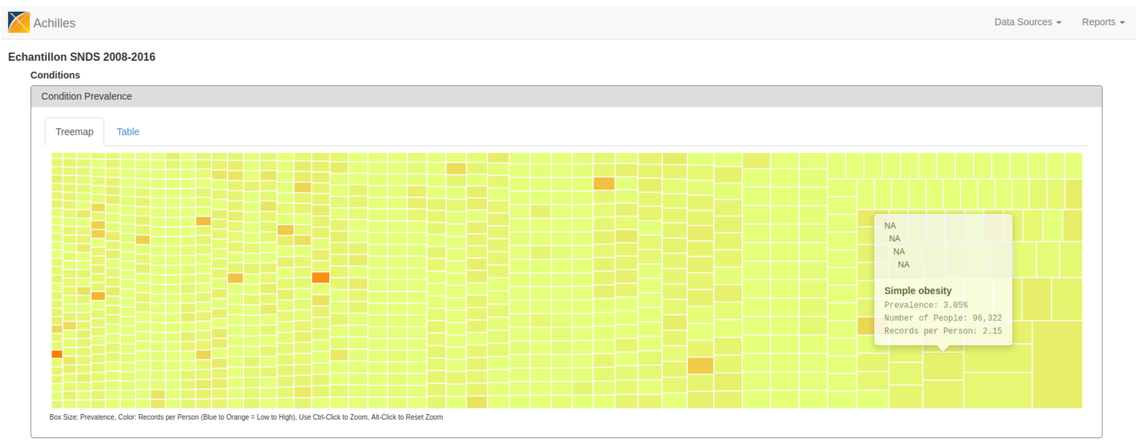
Résultats

La base OMOP constituée comprend notamment 1,48 milliard de visites, 19,57 millions de diagnostics principaux, 606,65 millions d'achats de médicaments. Les principales terminologies du SNDS ont été alignées vers les standards internationaux à hauteur de 74 % de tous les codes. Les taux d'alignement dans les tables cibles sont les suivants :

- médicaments (variable drug_concept_id) : 89,3 %,
- spécialités médicales (variable specialty_concept_id) : 29,5%,
- actes médicaux (variable procedure_concept_id) : 1,3 %,
- visites (variable visit_concept_id) : 59,1 %,
- diagnostics (variable condition_concept_id) : 99,9 %,
- biologie (variable measurement_concept_id) : 0 %.

Comme premier cas d'usage, nous avons utilisé les requêtes d'Achilles [1], un logiciel de description automatique pour des données au format OMOP. La Figure 3 montre le résultat visuel obtenu pour les diagnostics : chaque case concerne un diagnostic, l'aire est proportionnelle à la prévalence et la couleur illustre le nombre moyen de cas par personne. Nous testons actuellement les requêtes pour l'âge de primo-apparition des concepts médicaux : diagnostics, médicaments, visites.

Figure 3. Prévalences des diagnostics dans l'échantillon SNDS au format OMOP



Discussion et conclusion

Afin de conforter la pertinence de la base OMOP, nous finalisons la production d'indicateurs de morbidité dans le cadre d'une collecte européenne pour Eurostat. Nous les déclinons sur la base SNDS brute et la base SNDS-OMOP afin de proposer un algorithme partageable de calcul des indicateurs.
Le développement de correspondances vers des standards comme OMOP permet de faciliter l'utilisation du SNDS. La perspective d'une base de données standardisée offre de nombreux avantages, tels qu'une meilleure qualité des données, des outils et des méthodes partageables au niveau international, la facilitation de la validation externe d'algorithmes, ainsi qu'une visibilité accrue au sein d'une communauté internationale.
Des efforts conséquents restent à fournir sur : (i) la correspondance sémantique, a minima pour des études françaises alignées, (ii) le développement de premières études de cas mettant à l'épreuve la base standard constituée.
Une autre limite que nous voyons à l'utilisation d'une base standard est la diversité des environnements de calcul, nécessitant donc tout de même quelques adaptations pour un fonctionnement optimal. Nous recommandons l'utilisation massive de requêtes de type SQL, facilement transposables d'un environnement à un autre.

Références

[1] *Achilles, logiciel de caractérisation automatisée de la base OMOP*. Availble at `https://github.com/OHDSI/Achilles`. URL: `https://github.com/OHDSI/Achilles`.

[2] Emmanuel Bacry, Stéphane Gaïffas, Fanny Leroy, Maryan Morel, Dinh Phong Nguyen, Youcef Sebiat, and Dian Sun. "SCALPEL3: a scalable open-source library for healthcare claims databases". In: *arXiv:1910.07045 [cs]* (Oct. 2019). arXiv: 1910.07045. URL: `http://arxiv.org/abs/1910.07045` (visited on 10/18/2019).

[3] *Documentation collaborative du SNDS*. Available at `https://documentation-snds.health-data-hub.fr/`. URL: `https://documentation-snds.health-data-hub.fr/` (visited on 02/10/2020).

[4] *Gitlab interCHU/snds-structural-mapping*. en. `https://framagit.org/interchu/snds-structural-mapping/tree/master/drees/mappings_documentation/results`. URL: `https://framagit.org/interchu/snds-structural-mapping/tree/master` (visited on 02/10/2020).

[5] George Hripcsak, Jon D Duke, Nigam H Shah, Christian G Reich, Vojtech Huser, Martijn J Schuemie, Marc A Suchard, Rae Woong Park, Ian Chi Kei Wong, Peter R Rijnbeek, Johan van der Lei, Nicole Pratt, G Niklas Norén, Yu-Chuan Li, Paul E Stang, David Madigan, and Patrick B Ryan. "Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers". In: *Studies in health technology and informatics* 216 (2015), pp. 574–578. ISSN: 0926-9630. URL: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4815923/` (visited on 05/27/2019).

[6] *The Book of OHDSI*. Available at `https://ohdsi.github.io/TheBookOfOhdsi/`. URL: `https://ohdsi.github.io/TheBookOfOhdsi/` (visited on 02/02/2020).

Drees — Direction de la recherche, des études, de l'évaluation et des statistiques
Statistique publique
Liberté • Égalité • Fraternité — République Française
Ministère des Solidarités et de la Santé
Ministère du Travail
Ministère de l'Action et des Comptes publics