Mapping the 66.6M-patient French Nationwide Healthcare Database (SNDS) to the OMOP CDM v5: a feasibility study

Nicolas Thurin¹,², Marine Pénichon¹, Régis Lassalle¹, Patrick Blin¹, Cécile Droz-Perroteau¹, Martijn Schuemie³, Nicholas Moore¹,²

¹Bordeaux PharmacoEpi, INSERM CIC1401, Université de Bordeaux, Bordeaux, France; ²INSERM U1219, France; ³Janssen R&D, Titusville, New Jersey, USA

Background

- SNDS (Système National des Données de Santé) is the French Nationwide Healthcare System Database
  - Covers about 99% of the French population (about 66.6 million persons) from birth (or immigration) to death (or emigration)
  - Merger of information from reimbursed claims (SNIRAM) linked to the national hospital-discharge summary (PMSI) and the national death registry (CépiDC)
  - Includes: demographics; physician or paramedical visits (without medical indication); dispensed drugs and medical devices; lab tests (without results); chronic medical conditions; hospitalization with ICD10 codes for primary, linked and associated diagnoses, date and duration; reimbursed procedures

- EGB (Échantillon Généraliste de Bénéficiaires) is the 1/97th random permanent representative sample of SNDS
  - 810 000 persons
  - Planned 20-year longitudinal data
  - Same structure than SNDS

- EGB is an ideal candidate for assessing the possibility to apply the OMOP CDM to the SNDS

Methods

1. Extraction of an EGB sample (extraction of all the myocardial infarctions between 2009 and 2014)
2. Data management of the 74 resulting datasets to 14 transitional ones
   - Socio demographic data
   - Medical visits
   - Chronic medical conditions
   - Hospitalization diagnoses
   - Drug dispensions
   - Test labs
   - Inpatient procedures
3. ETL design
   - Theoretical mapping of the variables from EGB to OMOP CDM v5 (Figure 1.)
   - Assessment of the mapping from source codes and thesauri to OMOP standardized vocabulary (Table 1.)
4. Effective mapping of data
   - Variables mapping
   - Generation of the treatment duration for the drug_exposure table according to drug dispensions data
   - Codes mapping whenever present in OMOP vocabulary

Results

ETL design

<table>
<thead>
<tr>
<th>EGB</th>
<th>CDM V5</th>
</tr>
</thead>
<tbody>
<tr>
<td>socio-demographics</td>
<td>person</td>
</tr>
<tr>
<td>long term condition</td>
<td>observation_period</td>
</tr>
<tr>
<td>principal and related diagnosis of hosp. stays</td>
<td>location</td>
</tr>
<tr>
<td>associated diagnosis of hosp. stays</td>
<td>death</td>
</tr>
<tr>
<td>public hosp. inpatient lab tests</td>
<td>condition_occurrence</td>
</tr>
<tr>
<td>private hosp. inpatient lab tests</td>
<td>care_site</td>
</tr>
<tr>
<td>public hosp. outpatient lab tests</td>
<td>procedure_occurrence</td>
</tr>
<tr>
<td>private hosp. outpatient procedures</td>
<td>visit_occurrence</td>
</tr>
<tr>
<td>public hosp._outpatient procedures</td>
<td>measurement</td>
</tr>
<tr>
<td>non-hosp. outpatient lab tests</td>
<td>provider</td>
</tr>
<tr>
<td>non-hosp. outpatient procedures</td>
<td>drug_exposure</td>
</tr>
</tbody>
</table>

Figure 1. EGB data Mapping Approach to CDM v5 (clinical part only, without devices)

Effective mapping of data

- To date, 4 data tables have been generated
  - person
  - drug_exposure
  - location
  - observation_period
  - Drug treatment duration for the drug_exposure table has been generated using medians of treatment duration and taking into account the 3-month drug boxes.

Table 1. Presence of EGB and SNDS vocabularies in the OMOP vocabularies

<table>
<thead>
<tr>
<th>Domains</th>
<th>EGB / SNDS vocabularies</th>
<th>Presence is the OMOP vocabularies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Condition</td>
<td>IC10</td>
<td>Yes</td>
</tr>
<tr>
<td>Drug</td>
<td>ATC</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>CIP</td>
<td>No</td>
</tr>
<tr>
<td>Device</td>
<td>LPP</td>
<td>No</td>
</tr>
<tr>
<td>Procedure</td>
<td>CCAM</td>
<td>No</td>
</tr>
<tr>
<td>Measurement</td>
<td>NABM</td>
<td>No</td>
</tr>
</tbody>
</table>

ATC: Anatomical Therapeutic Chemical; CCAM: Classification Commune des Actes Médeciaux; CIP: Code Identifiants de Présentations; ICD: International Classification of Diseases; LPP: Liste des Produits et Préséntations; NABM: Nomenclature des Actes de Biologie Médicale; UCD: Unité Commune de Dispensation

- Most of the national nomenclatures used in France are not included in the OMOP vocabularies
  - CIP is used to identify drug packages dispensed in community pharmacies
  - UCD is used to identify expensive drugs dispensed in the hospitals
  - LPP is mainly used to code reimbursed medical devices
  - CCAM is used to code outpatient and inpatient procedures
  - NABM is used to code inpatient and outpatient lab tests

Conclusions

- Most of the standardized clinical data tables can be generated following OHDSI Extract Transform and Load (ETL) process. However the mapping of the French terminologies for procedures, lab tests, drugs and devices to OMOP Standard Vocabularies is required to complete fully the transformation of the clinical part of EGB and SNDS and enable the use of the whole set of OHDSI tools.