Syntactic and Semantic Harmonization of the French National Healthcare Database (SNDS)

Lorien Benda[1], Régis Lassalle[2], Cécile Roseau[1], Stéphanie Combes[1], Cécile Droz-Perroteau[1], Nicolas Thurin[1]

[1]Plateforme des Données de Santé (Health Data Hub), 75015, Paris, France

[2]Bordeaux PharmacoEpi, INSERM CIC-P 1401, Université de Bordeaux, 33000, Bordeaux, France

Background

France hosts one of the world's largest continuous homogeneous claims databases: The *Sytème National des Données de Santé* (SNDS).[1] Using a unique pseudonymized identifier, the SNDS merges reimbursed outpatient claims from all French healthcare insurance schemes with hospital-discharge summaries from public and private hospitals, and the national death registry. This medico-administrative database covers 98.8% of the French population, over 66 million people, from birth (or immigration) to death (or emigration). Under certain conditions, SNDS data can also be used to enrich other research databases through data linkage. The use of the SNDS for research purposes has long been under-exploited due to access rules but also because of its permanent evolving complex structure (180 tables, 4500 variables, 250 To of data) and ontologies. The SNDS relies on numerous specific French vocabularies e.g., CCAM and CSARR for procedures, NABM for laboratory tests, LPP for medical devices, CIP and UCD for drugs. Data standardization through the implementation of common data models (CDM) appeared as promising features to improve the reuse of the SNDS for real-world evidence generation and the sharing of scripts.

The Bordeaux PharmacoEpi platform from the Université de Bordeaux and the Health Data Hub have been working on the transformation of the SNDS to the OMOP CDM with the support of EHDEN. The Extract Transform and Load (ETL) process counted two components:

- The syntactic harmonization, involving the alignment of data structures and variable labels;
- The semantic harmonization, involving the alignment of the ontologies used in the SNDS to the OMOP standard vocabularies.

Methods

In the frame of the syntactic harmonization, two SNDS to OMOP CDM v5.3.1 ETLs were drafted with the support of SNDS experts, and OMOP experts by the Université of Bordeaux and the HDH teams. These two ETLs were then compared for consistency, in order to ensure the quality of the process.

Regarding the semantic harmonization, source concepts were translated by an automatic translator (Deepl) and then sent for alignment to medical residents. The residents were asked to proofread the English translation, and to use Usagi to perform the mapping to the corresponding standard OMOP standard concept. The levels of alignment equality had to be filled in. In a second step, alignments were reviewed by a different resident. For the procedure domain, mapping at the code level was restricted to the 80 % of the most occurrent source concepts in 2019-2020 (outpatient and inpatient). The remaining concepts were mapped at the chapter level. For the drug, the measurement, the visit and the provider

domains, source concepts were mapped at the code level, as well as for some algorithm-derived condition source concepts. For medical devices, the alignment was done at the chapter level.

Results

The syntactic harmonization led to the generation of the following tables of the OMOP CDM: PERSON, OBSERVATION_PERIOD, VISIT_OCCURRENCE, VISIT_DETAIL, CONDITION_OCCURRENCE, DRUG_EXPOSURE, PROCEDURE_OCCURRENCE, DEVICE_EXPOSURE, OBSERVATION, DEATH, LOCATION, CARE_SITE, PROVIDER

Table 1 presents the summary of the semantic harmonization work. From the initial 8 179 CCAM and 566 CSARR source concepts, all have been translated into English, and respectively 686 and 98 were mapped to a standard SNOMED-CT concepts, as well as all chapter level terms. For the drug domain, mapping of source concepts at the product level (CIP and UCD) is currently ongoing. Alignment at the ingredient level is already available via product ATC codes. It was initially planned to rely on a pre-existing alignment between NABM and LOINC for laboratory tests.[2] However, since unlike LOINC the French NABM codes do not include laboratory test results, it was decided to map the 963 NABM concepts to SNOMED-CT in the procedure domain. The overall LPP chapter level terms were mapped to SNOMED-CT in the Device domain, too much details being provided at the code level. All specific SNDS codes qualifying from where a patient was admitted or discharged to, and health provider specialty were mapped at the code level.

French ontology	Meaning	Main target domains	Number of mapped source concepts
CIM10	Hospital discharge codes	Conditions	Included in OMOP vocabulary
CCAM	Medical procedures	Procedure Observation Spec Anatomic Site	686 / 8 179 concept codes 1 387 / 1 387 chapters codes
CSARR	Physical and speech therapy	Procedure	98 / 566 concept codes 94 / 94 chapters codes
ATC	Drug (ingredient level)	Drug	Included in OMOP vocabulary
CIP / UCD	Drug (box and dispensing unit level)	Drug	Ongoing
NABM	Laboratory test (no results)	Measurement procedure	973 / 973 concept codes
LPP	Medical devices	Device	0 / 29 161 concept codes 399 / 399 chapters codes
ENT_PRV	where the patient was admitted from	Visit	9 / 9 concept codes
SOR_MOD	where the patient was discharged to	Visit	8 / 8 concept codes
IR_SPE_V	Healthcare provider specialties	Provider	96 / 96 concept codes
CT_IND	Algorithm-derived major comorbidities flags	Condition	202 / 202 concept codes

Table 1. Summary of French vocabularies mapping

Conclusion

Syntactic harmonization has been successfully conducted. Semantic harmonization was made complex by the level of detail captured by the French ontologies and is currently being improved. The current ETL already enables the execution of federated real-world study in the SNDS using OHDSI tools, making its power available for health outcome research.

Acknowledgment

The Health Data Hub and Bordeaux PharmacoEpi warmly thank the medical residents for their work on the mapping : Alexandre Kitic, Nicolas Kitic, Raphaël Lee, Sara Tunon de Lara, François Bourquard,

Abbreviations: ATC = Anatomical, Therapeutic, Chemical Classification; CCAM = Classification Commune des Actes Médicaux; CIM10 = classification internationale des maladies, 10e révision; CIP = Code Identifiant de Présentation; CSARR = Catalogue Spécifique des Actes de Rééducation et Réadaptation; LPP = Liste des Produits et Prestations; NABM = Nomenclature des Actes Biologiques; UCD = Unité Commune de Dispensation; LOINC = Logical Observation Identifiers Names & Codes; SNOMED - CT = Systematized Nomenclature of Medicine Clinical Terms.

References

1. Bezin J, Duong M, Lassalle R, Droz C, Pariente A, Blin P, Moore N. The national healthcare system claims databases in France, SNIIRAM and EGB: Powerful tools for pharmacoepidemiology. Pharmacoepidemiol Drug Saf. 2017 Aug;26(8):954-962. doi: 10.1002/pds.4233. Epub 2017 May 24. PMID: 28544284
2. https://bioloinc.fr/bioloinc/KB/