Syntactic and Semantic harmonization of the French National healthcare database (SNDS)

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

[1]Plateforme des Données de Santé (Health Data Hub), Paris, France, opensource@health-data-hub.fr

[2]Bordeaux PharmacoEpi, INSERM CIC-P 1401, Université de Bordeaux, Bordeaux, France, nicolas.Thurin@u-Bordeaux.fr

*Contributed equally

Introduction

- The **SNDS** is one of the world's largest healthcare database, encompassing **outpatients claims**, **hospital discharge summaries**, and **national death registry** for the whole French population
- SNDS relies on a complex structure and numerous specific vocabularies : e.g., **CCAM** and **CSARR** (procedures), **NABM** (laboratory tests), **LPP** (medical devices), **CIP** and **UCD** (drugs).
- **Data standardization** is needed to **improve the reuse of the SNDS** for real-world evidence generation and **promote script and program sharing** .

Methods

❖ Syntactic harmonization

SNDS to OMOP CDM v5.3.1 ETLs drafted by experts from the Université de Bordeaux and HDH team.

❖ Semantic harmonization

1. **Translation** of source concepts (Deepl)
2. **Proofreading** and **correction** of the English translation
3. **Mapping** to the standard OMOP concepts with **USAGI** by medical residents and experts

French ontology	Level of mapping
CCAM/CSARR	80 % of the most occurrent source concepts (2019-2020, inpatient and outpatient) : mapping at the code level Others : mapping at the chapter level
CIP / UCD / NABM / ENT_PRV / SOR_MOD / IR_SPE_V / CT_IND	Mapping at the code level
LPP	Mapping at the chapter level

4. **Cross-review** of the mapping

❖ Syntactic harmonization

The following tables of the OMOP CDM v5.3.1 were generated:

- PERSON
- OBSERVATION_PERIOD
- VISIT_OCCURRENCE
- VISIT_DETAIL
- CONDITION_OCCURRENCE
- DRUG_EXPOSURE
- PRCEDURE_OCCURRENCE
- DEVICE_EXPOSURE
- OBSERVATION
- DEATH
- LOCATION
- CARE_SITE
- PROVIDER

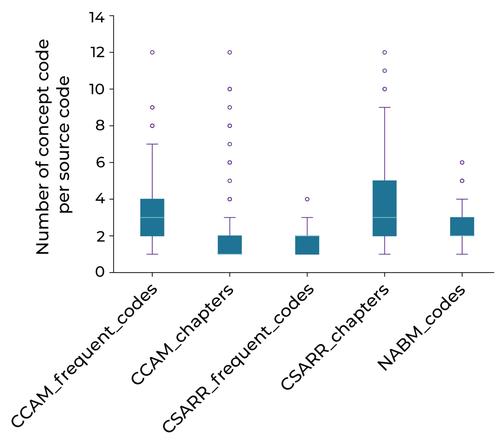

Figure 1. Level of equivalence for CCAM, CSARR and NABM codes

Results

❖ Semantic harmonization

The following tables of the OMOP CDM v5.3.1 were generated:

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 764 / 764 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

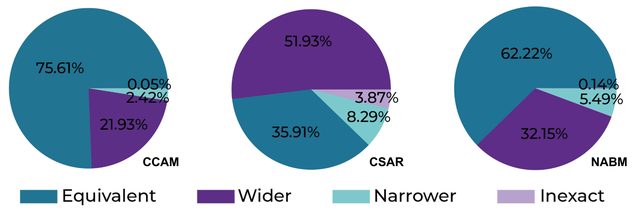

Figure 2. Number of target concepts per source code / chapter

- Regarding CCAM codes, **22%** of the targets are **wider than the source code**, showing this ontology is particularly detailed (Figure 1).
- The **most frequent CCAM codes** are mapped to a **median of 3 codes,** while chapters with less detail are mapped to 1 code in median (Figure 2).

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 SNDS using OHDSI tools, making its power available for health outcome research**