MIMIC-III into OMOP : 48h hackathon evaluation

N. Paris[1*], A. Parrot[1$], T. Pollard[2], A.E.W. Johnson[2]

[1]WIND-DSI, AP-HP, Paris, France

[2]Laboratory for Computational Physiology, MIT Institute for Medical Engineering and Science

[*]nicolas.paris@aphp.fr, [$]adrien.parrot@caramail.fr

Background

- MIMIC is an open intensive care unit (ICU) dataset developed by the MIT Lab for Computational Physiology, containing deidentified health data associated with **50,000 ICU patients** (demographics, vital signs, biology, medication, reports)
- The MIT Lab plan to adopt a common data model to enhance reproductibily and to facilitate integration with international datasets[1]
- The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) has been shown to be an effective way to standardize observational health databases [2]

1 Objectives

1. Evaluate the feasibility of transforming electronic health records (EHR) into OMOP-CDM.
2. Assess the completeness of vocabulary mapping
3. Propose improvements of OMOP-CDM in the context of datas from ICUs
4. Share the experience of the use of OMOP-CDM in APHP Big Data Platform

2 Methods

- Data Source : MIMIC-III version 1.4
- CDM version : OMOP-CDM version 5.3 which defines 15 standardized clinical data tables, 3 health system data tables, 2 health economics data tables, 5 tables for derived elements and 12 tables for standardized vocabulary
- **Freely postgreSQL-based ETL implementation**[3]
- Introduction of unit row id for all MIMIC database : mimic_id
- Concepts mapping : manually (based on csv files, visible on our github and check by physicians) and automatically with omop concept_relationship table. We used the OMOP mapping for NDC-RxNorm, ICD9-SNOMED, CPT4-SNOMED.
- **Concept driven methodology** : the domain of each local concept drive the concept to the right table
- **Unit tests** implemented during the ETL process : pgTAP
- Constant dialogue with OMOP and MIMIC communities
- Split notes in section to make natural language processing (NLP) easier
- Modification of OMOP-model : add columns

Field	Type	Required
offset_begin	integer	No
offset_end	integer	No
section_source_value	text	No
section_source_concept_id	integer	No

Table 1: NOTE_NLP proposal

Field	Type	Required
admitting_concept_id	integer	No
admitting_source_value	character varying(50)	No
admitting_source_concept_id	integer	No
discharge_to_concept_id	integer	No
discharge_to_source_value	character varying(50)	No
discharge_to_source_concept_id	integer	No

Table 2: VISIT_OCCURRENCE proposal

3 Results

About the conversion to OMOP-CDM

- OMOP-provided mapping : we check 100 items for each mapping used (NDC, ICD9 and CPT4). ICD9 and CPT4 are correcly mapped to SNOMED. But only 85% of NDC are linked to a correct RxNorm code. In part due to incorrect NDC code (from MIMIC), in part because only **78% of NDC codes are mapped to Rxnorm**

OMOP tables	Number of rows	MIMIC-III tables	Mapping
PERSONS	46520	patients, admissions	100%
DEATH	14849	patients, admissions	100%
VISIT_OCCURRENCE	58976	admissions	100%
VISIT_DETAIL	271808	transfers, service	100%
MEASUREMENT	366226116	chart / lab / events / outputevents	70 %
OBSERVATION	6721040	admissions, drgcodes, chart / datetimeevents	70%
DRUG_EXPOSURE	24934758	prescriptions, inputevents_cv /_mv	62%
PROCEDURE_OCCURRENCE	1063525	cptevents, procedure events_mv / _icd	99%
CONDITION_OCCURRENCE	716595	admissions, diagnosis_icd	94 %
NOTE	2082294	notevents	0%
NOTE_NLP	16350855	noteevents	NA
COHORT_ATTRIBUTE	2628838	callout	0%
CARE_SITE	93	transfers, service	100%
PROVIDER	7567	caregivers	100%
OBSERVATION_PERIOD	58976	patients, admissions	NA
SPECIMEN	39874171	chart / labevents / microbiologyevents	71 %

Table 3: Table mapping from MIMIC III source data to OMOP-CDM and % of standard mapping

Remark : we didn't use the health economics data tables (not provided in MIMIC)

- All the rows marked as error are removed : $< 5\%$ of rows per table
- 50% of the MIMIC columns are not used in OMOP models. Mostly due to the redundant or derived informations. The main concern could be the timestamp when the measurements contain a lot of it.
- Add many derived values (BMI...) and scores (SOFA, SAPSII...)
- Data transformation was done by 2 developers and praticians in 500 hours

About APHP data platform and hackathon

- 25 teams, 160 participants had 48 hours to undertake a clinical project using the OMOP MIMIC-III database throught **15000 requests**. They had the opportunity to create mixed teams : clinicians brought the questions which need data mining, along with their expertise of the data ; data scientists judged the technical feasibility and eventually implement the various analysis needed
- AP-HP calculation clusters, able to access to the data pre-loaded in Jupyter environments, where will be installed the most popular tools and libraries in R and Python, with Hadoop Spark

Figure 1: Big Data Platform

4 Conclusion

- ETL implementation:
  → Simple, postgreSQL/csv based ETL for community improvement & sharing[3]
  → Developped by AP-HP engineer & intensivist, supported by MIMIC & OMOP experts
  → We encourage others teams to help us to improve the concept mapping and ETL code
- HACKATHON:
  → OMOP suffers from too verbose SQL queries
  → Model relatively easy to understand (couple of hours)
- PERPECTIVES:
  → Evaluation OHDSI softwares : ACHILLES, ATLAS...
  → Proposals : denormalized model to improve performances, ICU centric tables, quality table, temporal column for drug_exposure and dose_era tables
  → AP-HP hospitals research data warehouse (10 millions patients) will use OMOP-CDM

References

[1] Alistair E.W. Johnson, Tom J. Pollard, and al. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035, May 2016.

[2] Erica A Voss and al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *JAMIA*, 22(3):553–564, May 2015.

[3] MIMIC3 omop cdm. https://github.com/MIT-LCP/mimic-omop. Accessed: 2018-01-03.