A simplified ETL approach to transforming the MIMIC database into the OMOP Common Data Model in SQLite

Luis Alberto Robles Hernandez, Juan M. Banda

Background

Electronic Health Records (EHR) have played an important role in transforming clinical data into knowledge in order to improve patient care(1). As a consequence, several articles describing the transformation of a clinical database into a common data model, such as Observational Medical Outcomes Partnership (OMOP), have been published(2–5), which allow converting certain types of open-access databases to a standard data model which unifies the data in a known schema with semantic consistency. However, most of these resources are not maintained properly or are outdated. Therefore, when performing the transformation process to the OMOP data model, some records are mapped to deprecated or non-existent entities. Additionally, the implementation process of some of these ETL approaches is long and complex since numerous technical configurations are required prior to its implementation.

This work was carried out based on the approach of Paris et al.(2), in which their main objective was to transform a Medical Information Mart for Intensive Care (MIMIC) database(6) into an OMOP database in order to evaluate the benefits of this transformation. In this paper, we propose a simplified and updated version of that ETL, with the purpose of facilitating this transformation process. As an evaluation, we present a mapping coverage comparison between the original approach and our approach.

Methods

We used the first comprehensive and publicly available mapping from MIMIC to OMOP by Paris et al.(2), found available on Github (7). Originally, their approach uses the PostgreSQL RDBMS for its implementation with additional configuration steps, making it complicated to install on certain types of computers or when not having administrator rights. Our approach uses SQLite which does not require any installation or configuration, and most of the work performed was to update certain compatibility issues found.

Note that our approach used version 1.4.0 of the MIMIC database, a version similar to the one proposed by the authors previously mentioned (1.4.21) and with minor differences with respect to the number of records. The MIMIC database contains the following (with relation to each of the OMOP tables):

Table 1. MIMIC-OMOP data flows

OMOP tables	Number of rows (n)	MIMIC tables
CARE_SITE	93	transfers, service
COHORT_ATTRIBUTE	334,117	callout
CONCEPT	11,535,651	d_cpt, d_icd_procedures, d_labitems
CONDITION_OCCURRENCE	716,595	admissions, diagnosis_icd
DEATH	15,759	patients, admissions

DRUG_EXPOSURE	24,943,776	prescriptions, inputevents_cv, inputevents_mv
MEASUREMENT	365,449,324	chart/lab/microbiology/in/output events
NOTE	2,082,294	noteevents
OBSERVATION	6,721,040	admissions, chartevents, datetimevvents, drgcodes
OBSERVATION_PERIOD	58,076	patients, admissions
PERSON	46,520	patients, admissions
PROCEDURE_OCCURRENCE	1,063,525	cptevents, procedureevents_mv, procedure_icd
PROVIDER	7,567	caregivers
SPECIMEN	40,142,391	chartevents, labevents, microbiologyevents
VISIT_OCCURRENCE	58,976	admissions
VISIT_DETAIL	407,460	admissions, transfers, service

As a first step, a data definition language (DDL) was executed in the SQLite terminal in order to create the OMOP database (NOTE: OMOP CDM version 5.4 is being used for this process) including all its tables and columns. Then all the vocabulary tables were loaded from concepts obtained from Athena (8). Additional tables including manual mappings to the database under the MIMIC III schema were also loaded. It is important to emphasize that some of these manual mappings provided by Paris et al.(2) contain deprecated/outdated concepts, and a considerable number of them were updated in the current approach.

Once having the MIMIC data, the OMOP database, and the updated tables including manual mappings, the next step consisted of the extraction, transformation, and loading of the MIMIC database to the OMOP CDM database. This process was composed of moving all the records from the MIMIC database into the OMOP database into their respective tables, as shown in Table 1. Moreover, these records were also mapped to concepts previously loaded from the OHDSI vocabulary, if possible. And as an alternative, some of these records were mapped to MIMIC Local concepts.

As a final step, an evaluation process was carried out in order to mainly analyze the following parameters:

- Number of records mapped to a concept.
- Number of patients should match in both MIMIC and OMOP.
- Number of admissions should match in both databases.
- Incomplete or missing data.

Results

Table 2 and **Table 3** show the basic characterization of the MIMIC-OMOP population for both the approach proposed by Paris et al.(2) and the approach proposed in this work. We can highlight that in the original approach there is a difference in the number of ICU stays after the transformation process, while in the proposed approach, this number remains the same. Moreover, the number of admissions should match the total of emergency, elective, newborn, and urgent admissions, which is not the case for the original approach.

Table 2. Baseline characteristics of MIMIC versus OMOP (approach by Paris et al.(2))

Items	**MIMIC 1.4.21**	**MIMIC-OMOP**
Persons	46,520	46,520
Admissions	58,976	58,976
ICU stays	71,575	61,532
Female gender	20,399 (43.85%)	20,399 (43.85%)
Male gender	26,121 (56,15%)	26,121 (56,15%)
Age	64 years, 4 months	64 years, 4 months
Emergency	42,071	42,071
Elective	7,706	7,706
Newborn	-	N/A
Urgent	-	N/A

Table 3. Baseline characteristics of MIMIC versus OMOP (current approach)

Items	**MIMIC 1.4.0**	**MIMIC-OMOP**
Persons	46,520	46,520
Admissions	58,976	58,976
ICU stays	87,721	61,532
Female gender	20,399 (43.85%)	20,399 (43.85%)
Male gender	26,121 (56,15%)	26,121 (56,15%)
Age	64 years, 4 months	64 years, 4 months
Emergency	42,071	43,407*
Elective	7,706	15,569*
Newborn	7,863	N/A
Urgent	1,336	N/A

***** Both "emergency" and "urgent" from the MIMIC database is considered to be under the "emergency" category, whereas "elective" and "newborn" are considered to be under the "elective" category in the OMOP schema according to the manual mapping provided by Paris et al. (9)

Furthermore, as shown in Figure 1 we can observe the number of records successfully mapped to a concept from any of the loaded vocabularies from Athena. Ideally, every record should be mapped with any of these concepts, and by making a comparison between the approach proposed by Paris et al.(2) we can highlight that the number of mapped records is greater in the approach proposed in this work.

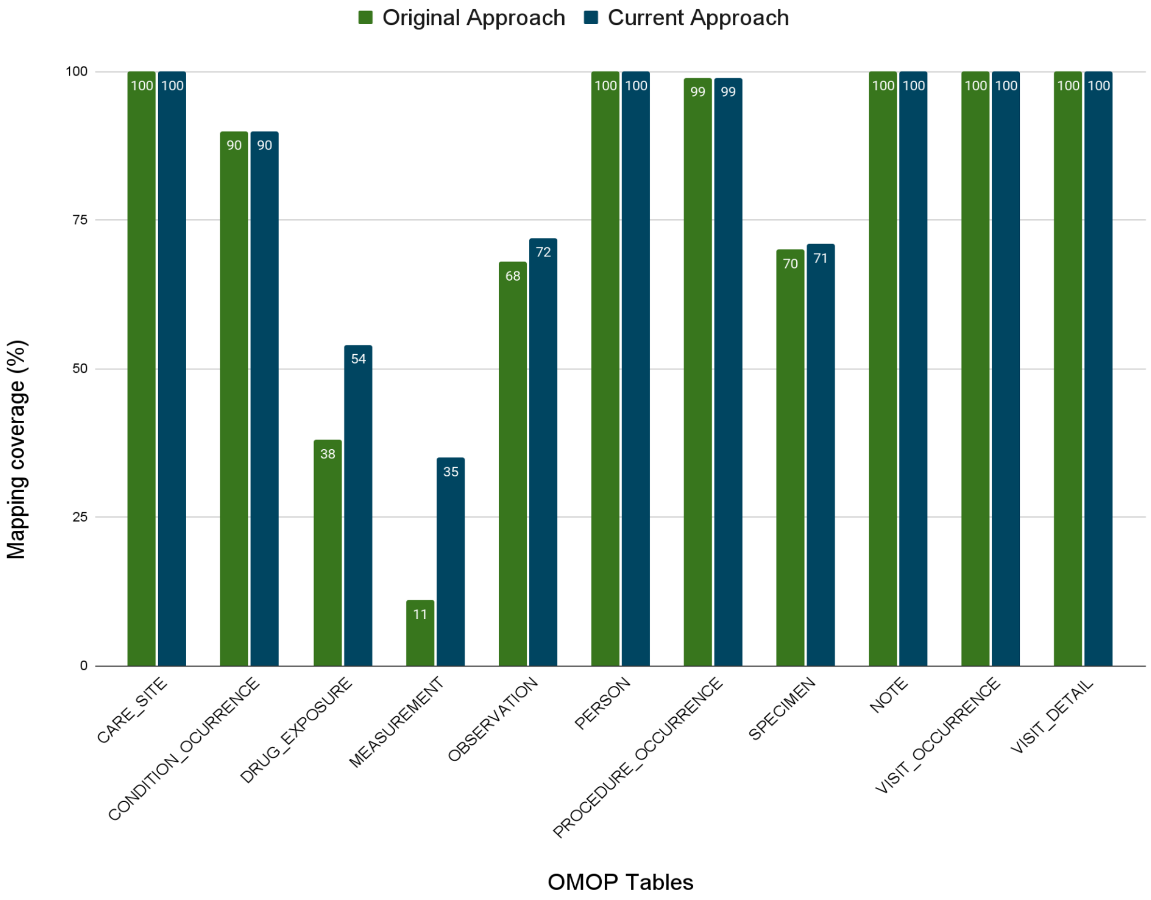

Figure 1. Terminology mapping coverage and comparison versus original approach(2)

Conclusion

In this work, we demonstrated that this approach obtained comparable results with respect to Paris et al.(2). Our evaluation for this was presented in the results section, showing similar or even higher overlaps after our ETL process. Moreover, the methodology used in this approach is simpler as SQLite was used in which no complex installation or configuration was required, greatly simplifying the process and allowing people with no administrator access to the computers to locally complete an ETL process. We make all our updated ETL codes and script available for all researchers to use via our GitHub repository(10).

References

1. Ross MK, Wei W, Ohno-Machado L. “Big data” and the electronic health record. Yearb Med Inform. 2014 Aug 15;9:97–104.

2. Paris N, Lamer A, Parrot A. Transformation and Evaluation of the MIMIC Database in the OMOP Common Data Model: Development and Usability Study. JMIR Med Inform. 2021 Dec 14;9(12):e30970.

3. Maier C, Lang L, Storf H, Vormstein P, Bieber R, Bernarding J, et al. Towards Implementation of OMOP in a German University Hospital Consortium. Appl Clin Inform. 2018 Jan;9(1):54–61.

4. FitzHenry F, Resnic FS, Robbins SL, Denton J, Nookala L, Meeker D, et al. Creating a common data model for comparative effectiveness with the Observational Medical Outcomes Partnership. Appl Clin Inform. 2015 Aug 26;6(3):536–47.

5. Lamer A, Depas N, Doutreligne M, Parrot A, Verloop D, Defebvre MM, et al. Transforming French Electronic Health Records into the Observational Medical Outcome Partnership’s Common Data Model: A Feasibility Study. Appl Clin Inform. 2020 Jan;11(1):13–22.

6. Alistair E.W. Johnson,Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, Roger G. Mark. MIMIC-III, a freely accessible critical care database [Internet]. 2016 [cited 2022 Jun 21]. Available from: http://dx.doi.org/10.1038/sdata.2016.35

7. MIMIC-OMOP [Internet]. [cited 2022 Jun 21]. Available from: https://github.com/MIT-LCP/mimic-omop

8. Athena [Internet]. [cited 2022 Jun 21]. Available from: https://athena.ohdsi.org/search-terms/terms

9. admission_type_to_concept.csv [Internet]. [cited 2022 Jun 21]. Available from: https://github.com/MIT-LCP/mimic-omop

10. mimic-omop: Mapping the MIMIC-III database to the OMOP schema (using SQLite) [Internet]. Github; [cited 2022 Jun 23]. Available from: https://github.com/thepanacealab/mimic-omop