A scalable framework for transforming multiple data sources to the OHDSI Common Data Model

Janos G. Hajagos

Background

The OHDSI Common Data Model (CDM) is increasingly becoming the standard format for representing EHR (Electriconic Health Record) and administrative claims data[1]. Transforming source data into the OHDSI CDM requires writing complex ETL (Extract Transform Load) scripts. Of a particular challenge is that rows from a source table are mapped to different domains (target tables) in the CDM. This mapping process is driven by the domain of the `concept_id` in the concept table. As an example, the ICD10-CM diagnosis code of Z01.818 (pre-procedure visit) will be mapped to a procedure domain while the ICD10-CM diagnosis code I25.10 (Atherosclerotic heart disease w/o hemodynamically effective stenoses) will be mapped to the condition domain. For each data source we need to write this row domain logic. We have developed an intermediatory data model, which we call Prepared Source Format (PSF) which is easier to target for ETL writers and can be used to streamline data mapping to the OHDSI CDM.

Methods

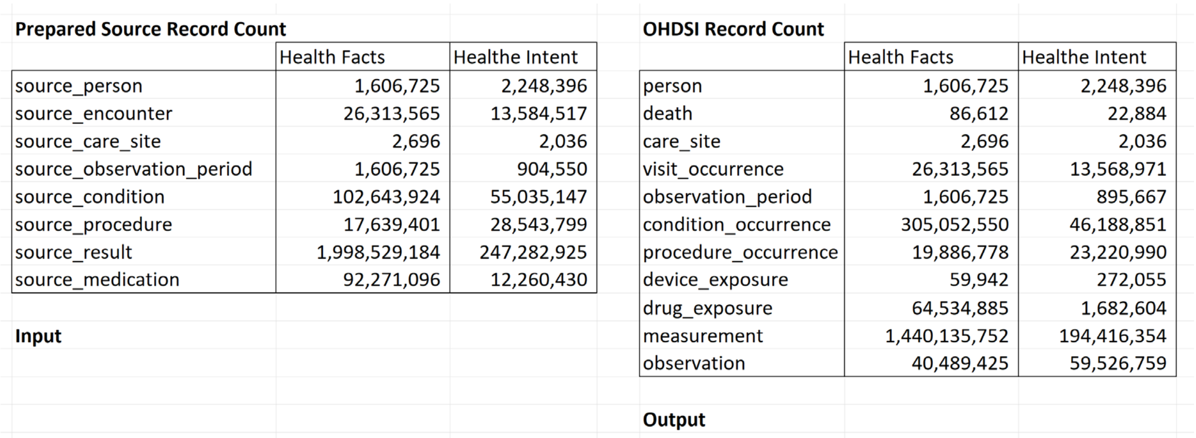
The current version of PSF format has the following defined tables (source_person, source_encounter, source_observation_period, source_condition, source_procedure, source_result, source_medication). Fields/column names in the PSF format are prefixed with either **s_** and **m_**. Fields with **s_** prefix are the source representation and fields with the **m_** prefix indicates values that the ETL writer has mapped for transformation purposes. Codes are defined using a code value paired with the object identifier value (OID). As an example, in the source_condition table a diagnosis code is defined as: **s_condition_code=I25.10** and **s_condition_code_type_oid=2.16.840.1.113883.6.90**.

Two separate data sources were mapped to the PSF format. The first source is a subset from Health Facts (a de-identified EHR database from Cerner) for patients who had at least one inpatient visit. The mapping to PSF was done directly in an Apache SPARK environment. The second data source is from Cerner's Healthe Intent population health platform and was mapped to PSF within the platform's SQL based workflow tool.

The mapper is written using PySpark and runs on the Databrick Runtime 9.1 LTS (Apache Spark 3.1.2, Scala 2.12). A scalable of cluster of 8 nodes with 16 cores and 56Gb memory was used to run the mapping process. Input data was staged in Microsoft Azure Blog Storage as compressed CSV files. For HealtheIntent data was further de-identified using a hashing function and fixed date and time. The mapper wrote files to Apache Parquet Format which can be queried directly with SPARK SQL or exported to other formats, such as a relational database. Both PSF datasets were transformed into OHDSI CDM version 5.3.1.

Results

Two different sources were first mapped to the PSF and then successfully mapped to the OHDSI CDM. Both sources represent realistic volumes of EHR data that would need to be handled for an observational study.



Prepared Source Record Count

	Health Facts	Healthe Intent
source_person	1,606,725	2,248,396
source_encounter	26,313,565	13,584,517
source_care_site	2,696	2,036
source_observation_period	1,606,725	904,550
source_condition	102,643,924	55,035,147
source_procedure	17,639,401	28,543,799
source_result	1,998,529,184	247,282,925
source_medication	92,271,096	12,260,430

Input

OHDSI Record Count

	Health Facts	Healthe Intent
person	1,606,725	2,248,396
death	86,612	22,884
care_site	2,696	2,036
visit_occurrence	26,313,565	13,568,971
observation_period	1,606,725	895,667
condition_occurrence	305,052,550	46,188,851
procedure_occurrence	19,886,778	23,220,990
device_exposure	59,942	272,055
drug_exposure	64,534,885	1,682,604
measurement	1,440,135,752	194,416,354
observation	40,489,425	59,526,759

Output

Figure 1. Record counts of mapping PSF data from two source Health Facts and Healthe Intent into the OHDSI CDM

Conclusion

The goal was to develop a process to streamline the mapping of EHR data into the OHDSI CDM. The intermediary format (PSF) does not completely remove all the complexity of writing ETL scripts to the OHDSI CDM. It simplifies the process of the domain/table specific targeting which can be complex to implement and requires understanding of the OHDSI CDM design principles[2]. By decoupling the initial ETL to the OHDSI specific transformation new sources can be more easily mapped to the OHDSI CDM.

References/Citations

1. Reinecke, I., Zoch, M., Reich, C., Sedlmayr, M., & Bathelt, F. (2021). The Usage of OHDSI OMOP - A Scoping Review. Studies in Health Technology and Informatics, 283, 95–103. https://doi.org/10.3233/SHTI210546
2. Observational Health Data Sciences and Informatics. (2019). The Book of OHDSI. https://ohdsi.github.io/TheBookOfOhdsi/TheBookOfOhdsi.pdf