OHDSI 2021 Abstract

Pragmatic OMOP CDM

Gregory Klebanov, Melanie Philofsky
Odysseus Data Services, Inc. MA, USA

Background

The Observational Health Data Sciences and Informatics (OHDSI) collaborative, is a global community that advocates for better health research through open-source analytics and data standardization. OMOP common data model (CDM) is a representation of biomedical data that standardizes entities, attributes, and relationships across multiple sources. The use of CDM-formatted data permits researchers across the OHDSI network to generate meaningfully comparable inferences from the same analyses.

One of the richest data sources for creation of OMOP databases is Hospital data. However, the data is typically spread across many systems, including EHR, billing, labs, case forms, surveys, registries and many more. Converting all data presents a real challenge due to the sheer amount of effort required.

In addition to this, the biggest challenge is that even after a long and costly OMOP CDM ETL project, there are still gaps being discovered in OMOP data when new studies are being executed. This leads to endless changes in the OMOP CDM ETL code, re-running full refresh and full re-testing of data. This is not only time and resource consuming but also presents a challenge to complete data fixes on-time as required for that study.

Instead, we have started to apply a more pragmatic approach to the OMOP conversion and refreshes:

- Convert only a core set of data required for the majority use cases
- Apply a method allowing dynamically inserting missing data into OMOP CDM without requiring a change to the ETL code and a full refresh of data

Methods

The OMOP CDM database is an enabling capability to support research - both internal and external. Thus, a Study should be a core driving source of the requirements for data that must be present in OMOP CDM. It is clearly impossible to predict all possible variations of studies that will be coming.

There is a vast variety of data that exist in Hospital settings and performing an initial OMOP CDM conversion without a clear and well-defined scope in mind could end up a very costly effort while still resulting in data missing as required for a specific study, thus rendering the effort wasted and database unusable for research without an additional significant effort. Instead, the OMOP CDM database should be looked at from a maturity perspective and a more pragmatic approach to enrichment applied to OMOP CDM database while it is going through the maturity phases:

1) MVP - initial OMOP CDM ETL completed
2) Evolving - first 1-2 years of on-going refreshes
3) Mature - > 2 years

During the **MVP Lifecycle Phase** (aka “Initial OMOP CDM ETL”), it is important for the scope to be driven by the upcoming research requirements - from both internal and external stakeholders. It recommended to identify the target business area and business stakeholders, the upcoming first batch of studies and

work with business stakeholders from day one to identify and document the data elements that will be required.

The following best practice should be applied to mature OMOP CDM during the **Evolving Lifecycle Phase**:

1) Establish a mature refresh cycle, including frequency of refreshes, instance archiving cadence, scope to ETL code updates.
2) Continuously work with business stakeholders to create a roadmap for the upcoming research
3) Use study protocol to identify research attributes that are required to be present in data.
4) Insert any missing data elements into OMOP CDM without performing a full refresh by updating those patient records with missing data only.
5) Incorporate changes into the full OMOP CDM ETL code during the next refresh cycle.

These steps will allow OMOP CDM instances to be continuously enriched with new data - or existing data to be continuously harmonized - while still allowing effective research to be performed on OMOP CDM without a delay.

Currently, there are only two approaches that are being applied to refresh data - full and incremental refresh, with full refresh being the most prevalent. However, adding missing data elements through a full refresh is a costly effort that involves code changes, full ETL re-run and full re-testing. Instead, a more "surgical" approach can be applied to add missing data elements or fix existing data points by applying changes to only those person records and data elements that are being affected.

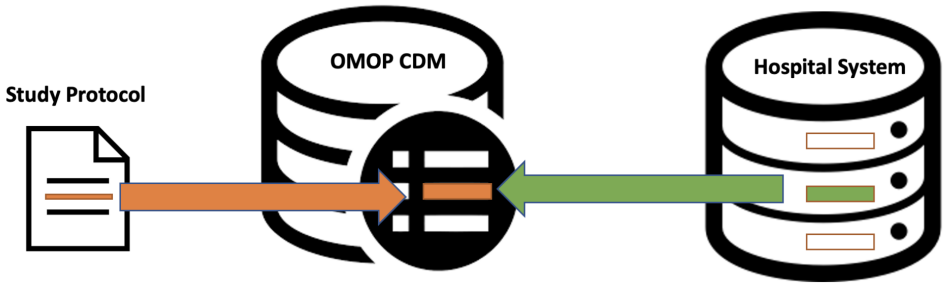

Figure 1: Surgical approach to targeted data enrichment

This will ensure a quick enrichment as required to support a specific study and is driven by a precise set of data elements as outlined in a study protocol.

Results

By focusing on defining the scope from the study perspective, our initial OMOP CDM ETL effort has decreased, while our team is still able to create an OMOP CDM that is relevant to the target study from Day 1. We are also focused on continuous enrichment of the OMOP CDM database during the subsequent support and maintenance lifecycle - both through full refreshes as well as targeted updates as required by the research portfolio.

Our ETL team is continuously working to refine a core set of data elements to be required for the initial Hospital OMOP CDM ETL ("MVP Phase") as well as improving technical solutions and methods for an effective "surgical" update of the OMOP CDM data during Evolving Phase.

Conclusions

The pragmatic approach to the initial OMOP CDM ETL and OMOP CDM refreshes minimizes the total effort and associated costs, while maximizing the value through all phases in the OMOP CDM lifecycle.