Analysis on Common OMOP Conversion Hiccups and ETL Challenges

Xialin Wang[1], Jing Li[1], Mui Van Zandt[1], Frank DiMartini[1], Anthony Reckard[1], Qi Yang[1]
[1]Real World Solutions, IQVIA, Durham, NC, USA

IMS Health & Quintiles are now IQVIA™

Background

In order to convert from source data to the OMOP CDM, extract-transform-load (ETL) process are implemented. Due to wide variety of unique source data, ETL process is difficult to master and a standard criterion is needed across databases. Data quality checks is the final and critical step to ensure a standard quality of data conversion. Analysis and summary of common hiccups we have seen during OMOP conversions are useful to be incorporated in data quality check process as guideline to improve conversion efficiency and accuracy.
Among the conversion projects we have done around the world, the following are the most common hiccups: 1) Incomplete CDM tables 2) Vocabulary issues 3) Data inconsistencies and abnormality 4) SQL environment issues.

Methods

Quality Control Methods

- OHDSI Tools
  - Achilles: Provided descriptive statistical analysis with reporting and data quality checks, executed with each refresh, and sent to clients as part of deliverable if purchased OMOP data asset
  - Data Quality Dashboard: Followed the Kahn Framework on Data Quality Assessment, checked the data quality from plausibility, conformance, and completeness, provided visualization of data checks and utilized configurable data check threshold
- IQVIA Proprietary Tools
  - CDM QC Check: provided in-depth quality assurance on all domains and serves as an ETL code review, executed along with each table once the table is completed
  - On-Premise Check: examined if the OMOP data abide to real-world scenarios, executed immediately after an OMOP conversion and prior to Achilles and Data Quality Dashboard
  - Business Validation: performed at the end of OMOP conversion, and compared between source data and converted OMOP CDM database to check if conversion is done properly

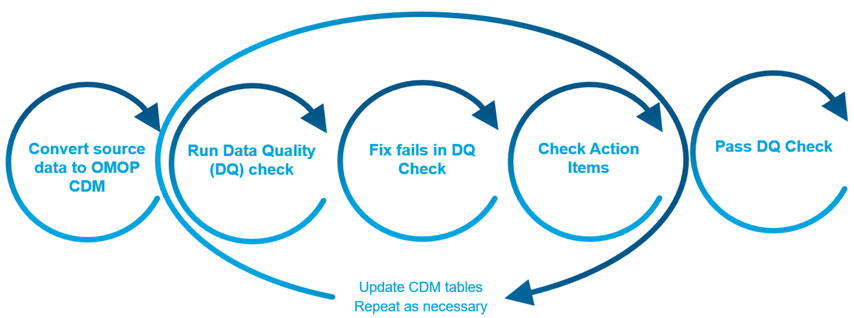

Figure 1. Data quality check process

Results

Summary of Common Hiccups in OMOP Conversion

- We summarized the most common hiccups of ETL conversions to OMOP CDM format across different sites in two continents, and categorized the issues into four categories.
- Four common categories of hiccups in OMOP conversions:
  - **Incomplete CDM tables**: multiple OMOP CDM tables are missing due to misunderstanding on how to populate the table
  - **Vocabulary issues**
    1. Multiple records for one concept mapping: pick one of the multiple standard vocabulary mapping to create the OMOP CDM record instead of one record per mapping.
    2. Non-standard vocabulary: data partners use proprietary coding system; codes mapped to OMOP vocabulary aren't mapped to a "Standard" concept.
    3. Invalid concepts: invalid concept_id were selected; no OMOP standard vocabulary mapping available.
    4. Wrong type_concept_ids: use of the wrong type_concept_id or misunderstanding the definition of this field
  - **Data inconsistencies and abnormality**
    1. Abnormal values: unconventional values exist in data asset (i.e. negative, null, or 0 as value_as_number)
    2. Incorrect logic of observation_period or drug_exposure days_supply calculation
       - Observation period does not cover entire period of time where events are recorded for a person.
       - Incorrect logic is applied to calculate drug days_supply. Negative, 0 or null values exist.
    3. Multiple input on records: some records contain multiple coding systems and text. A hierarchy must be selected to avoid duplicate records.
    4. Event dates outside of birth date or death date.
    5. Person table doesn't include all patients in all event table.
  - **SQL environment issues**
    1. MSSQL, ORACLE, PostgreSQL, Redshift and other environments require different syntax for certain SQL queries.
    2. Days_supply calculation in drug_exposure domain: DATEDIFF is applied to start date and end date in MSSQL, CAST function is applied to start date and end date to calculate days_supply in ORACLE.

Conclusions

As the conclusion, results from data quality check on multiple OMOP conversions around the world all presented issues in the same categories of errors, in addition to database-specific issues. To help others with their data conversion, the most common failed items were summarized to serve as a guideline for future OMOP conversion projects. Although the ETL challenges vary from databases to databases, it is recommended that data quality checks either via OHDSI tools or an organization's own tools be used to ensure the transformation prevents these kinds of issues. Also, this analysis will continue to be updated as more information and hiccups are encountered.