Make Your Tools Work for You: Customizing the Data Quality Dashboard to Identify Changes in Source Data

Melanie Philofsky, RN, MS[1]
[1] Odysseus Data Services, Cambridge, MA, USA

Background

The Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) is the semantic harmonization of observational healthcare data stored in a standardized format. The OMOP CDM is used by many organizations throughout the world to enable federated network research. The University of Colorado, Anschutz Medical Center (CU AMC) transforms electronic health record (EHR) and registry data to the OMOP CDM to contribute data for the National COVID Cohort Collaborative's (N3C) COVID-19 analytics[1] along with other national and local collaborations.

For the CU AMC, one of the main challenges in ensuring source data are comprehensively and accurately transformed to the OMOP CDM is identifying changes to source data and updating the extract, transform, and load (ETL) logic before the CDM is released to researchers. Therefore, one of the most important steps in the process is running the DQD tool on the OMOP CDM before making these data available for use. For this step, OHDSI's Data Quality Dashboard (DQD) is used to perform approximately 4,000 data quality checks[2] on the OMOP CDM. If any of the checks fail, then analysis of the failure can be undertaken to identify the source of the issue and adjustments made to the ETL logic to correct any issues found.

Methods

The DQD is preconfigured with threshold failure rates which might not be representative of the data in your CDM. Thankfully, these threshold failure rates are adjustable. There are three categories of checks: completeness, conformance, and plausibility. The completeness checks assess the percentage of data expected for a field. Completeness is dependent on the source data and the threshold should be adjusted to a level representative of the source data for a given query. Not only do you want to adjust the thresholds which fail the completeness checks to a level representative of your source data, but you also want to down adjust any predefined thresholds to a level just above your current failure rate in order to identify changes in your source data. By editing the completeness thresholds of individual checks, we were able to identify changes in the source data.

In order to monitor the completeness of the source data, we set the threshold levels of the DQD checks to 1% greater than the current failure rate for a field level check. This was done to ensure minor changes in the completeness of the source data would trigger a DQD failure notification for a particular check.

Results

The following changes to the source data or OHDSI vocabularies were identified after the tightening of the DQD failure thresholds:

1. **Addition of a new source field for a required data element in the OMOP CDM.** The first notification of a completeness failure check was for the gender_concept_id field in the person table. Upon analysis of the failure, it was discovered the value set for a person's sex had changed and the new values didn't map to OMOP's standard concepts of female or male. Further analysis of the source data revealed a new source field where a person's biological sex is stored in the source database. The ETL was altered to bring in data from this newly discovered field and completeness of the gender concept id field check rose to > 99%.
2. **Changes to the usual population whose data contribute to a dataset.** A drop in the completeness percentage for a Person's race, gender, and ethnicity field level checks lead to an investigation of the source data and subsequent discovery of many persons in the OMOP CDM who lack demographic data and have sparse clinical data. Sparse clinical data are defined by less than 3 clinical event records for a person in the OMOP CDM. Clinical events consist of records in the Condition Occurrence, Device Exposure, Drug Exposure, Measurement, Observation, or Procedure Occurrence domains. Many of these persons only have Covid immunization records in the source data. The healthcare system which contributes data to our instance of the OMOP CDM held many mass Covid immunization clinics when the vaccine first became available. By analyzing the data in the source electronic health record (EHR) system, we hypothesize the persons with sparse clinical data do not regularly receive care from the healthcare system but did receive one or more covid immunizations during the pandemic. Since the OMOP CDM is designed for longitudinal research studies, persons with sparse clinical data are deemed not suitable for research. Persons with sparse clinical data will be removed from the CDM to increase fidelity.
3. **Change in mapping to a standard concept identifier (concept_id) in a new vocabulary.** Analysis of an increase in the completeness failure rate for the Condition Occurrence table lead to the identification of a change in the mapping of a non-standard, source concept_id to a standard concept_id in a vocabulary not yet downloaded from Athena. This failure identified the need to download an additional vocabulary from Athena.
4. **Change in source data value set used in custom mapping data elements ETL'd to the CDM.** Some domains in an EHR do not have coded data elements. Therefore, these data elements and their source values must be manually mapped using an exact text string match to a standard concept_id. When there is a change in the source values, these data must be manually remapped.

Conclusions

Adjusting the DQD threshold levels to just above current failure rates assists data owners in ensuring data integrity remains high as changes to source data field use, collection of data, standard vocabulary changes, and source value sets evolve.

References

[1] Haendel MA, Chute CG, Bennett TD, Eichmann DA, Guinney J, Kibbe WA, et al. The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. J Am Med Inform Assoc. 2021 Mar 1;28(3):427–43.
Data Quality Dashboard. https://github.com/OHDSI/DataQualityDashboard. Accessed: May 12, 2023

[2] Data Quality Dashboard. https://github.com/OHDSI/DataQualityDashboard. Accessed: May 12, 2023