Real World Challenges When Using Real World Data: Creating a Multi-Institutional Database in OMOP

Michael N. Cantor, Deepika Sharma

Background

Transforming diverse clinical data into the common OMOP data model facilitates interoperability and standard approaches to analysis(1). In 2019, the Regeneron Genetics Center moved from receiving de-identified EHR data in its native, non-standard format, to receiving data in OMOP. While this change simplified the approach to transforming clinical data into a usable format (e.g., the ability to use a single codebase for data curation and processing), the transition to OMOP also led to several data integration challenges. These challenges were similar to other groups attempting to integrate diverse datasets into a national database(2, 3). Understanding the expectations around integration diverse datasets in OMOP and approaches to overcoming integration challenges should inform future OMOP transformations and analysis.

Methods

The RGC received data from 5 academic collaborators, all of whom transformed data from an Epic® EHR system to OMOP. Of note, all of the transformations were performed locally at the respective institutions, not with the help of outside consultants or third-party vendors (e.g. IQVIA). We examined issues that arose when attempting to create binary trait matrices (generally disease diagnosis codes) and quantitative trait matrices (generally lab results). We noted differences in encoding and formatting.

Results

The majority of the encoding issues arose in our analysis of quantitative traits from the measurement table. We found that lab values were often encoded with multiple different standard OMOP codes, both across and within institutions. For example, Body Mass Index (BMI) may be represented in LOINC or SNOMED CT , each of which has a different OMOP concept code. Similarly, *hematocrit by automated count* or *hematocrit of blood* are two of four potential OMOP concept codes that may be used for the same measurement. Many measurement tables and entries were also lacking high or low ranges and units. Different encodings ranged from 1 (Alkaline Phosphatase, among others) to 5 (anti-Smith antibody). Additionally, the unit of measurement for the same OMOP concept code/lab measurements was either missing or different for different patients. Finally, certain tables had different formats (different numbers of columns, different positions in the columns for certain data types). Specific examples of formatting issues included missing ‘year’, ‘month’, and ‘day’ columns in the person table.

Condition_occurrence tables also varied in their content and mapping approaches. Some tables contained the source ICD9 CM or ICD10 CM codes. Some conditions mapped to their SNOMED CT equivalent, while others mapped diseases to the OMOP concept entry for their ICD9 CM or ICD10 CM code. Additionally, the ETL process sometimes led to less-specific encoding, such as defaulting to 3 digit ICD9 CM or ICD10 CM codes that caused an unexpected increase in the count of patients with certain 3 digit diagnosis codes

Conclusion

OMOP is a valuable tool to facilitate large scale observational research. However, the real world implementation of the OMOP standards can lead to additional challenges for its practical use. Within an institution, the informatics group that defines the data transformation rules should work closely with the

IT group that implements the rules and transforms the data to ensure the clinical validity and quality of the transformed data. Additionally, because there are often multiple standard encodings for a clinical concept, OHDSI should publish a "standard of standards" set of encodings to help the community arrive at uniform mappings and avoid potential data loss. As institutions go beyond sharing summary statistics and share data, having this uniform approach will greatly facilitate large-scale studies.

References/Citations

1. Hripcsak G, Shang N, Peissig PL, Rasmussen LV, Liu C, Benoit B, et al. Facilitating phenotype transfer using a common data model. J Biomed Inform. 2019;96:103253.
2. Bradwell KR, Wooldridge JT, Amor B, Bennett TD, Anand A, Bremer C, et al. Harmonizing units and values of quantitative data elements in a very large nationally pooled electronic health record (EHR) dataset. J Am Med Inform Assoc. 2022;29(7):1172-82.
3. Cholan RA, Pappas G, Rehwoldt G, Sills AK, Korte ED, Appleton IK, et al. Encoding laboratory testing data: case studies of the national implementation of HHS requirements and related standards in five laboratories. J Am Med Inform Assoc. 2022.