Application of OMOP Common Data Model to Disease Registry Data

Vojtech Huser[1], Maria Rogozhkina[1], Vlad Korsik[1], Teresa A. Simon[2], Peter Moorthamer[2], Dan Kiselev[1], Anastasia Vakhmistrova[1], Eugene Paulenkovich[1], Alexander Davydov[1], Michel Van Speybroeck[2]

[1]Odysseus Data Services (an EPAM company), [2]Johnson and Johnson

Introduction

Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) was created to harmonize representation of healthcare data collected in disease registry data sources. The majority of the OMOP model use is in routine healthcare data. Since 2018, the Clinical Trial and Registry Working Groups of the Observational Health Data Sciences and Informatics (OHDSI) consortium were developing conventions of extending the model in research settings and use cases (clinical trial and registry).

We describe our experience and challenges in mapping and semantic harmonization, specifically disease measure instruments for 5 disease registry data sets. The rationale for converting site data into OMOP CDM is to use single analytical code and toolset across sites. (this is also called: portability of analysis benefit of using a CDM).

Selected encountered themes/issues are described below.

Materials-Methods-Results

Source semantic profile

The volume of semantic work for the project was determined by the number of data elements to be analyzed upon conversion. The average metrics and characteristics of each site are represented in the table below. The integration used 5264 site-level custom concepts and 498 network-level custom concepts (Registry vocabulary).

Site	Median number of unique codes per vocabulary per site	Average number of unique codes per vocabulary per site	Min number of unique codes per vocabulary per site	Max number of unique codes per vocabulary per site	Total number of unique codes per site
Site 1	17	58	2	169	403
Site 2	15	71	1	328	847
Site 3	33	80	2	217	797
Site 4	21	66	1	479	2375
Site 5	19	44	1	229	842
TOTAL					**5264**

Table 1. Site-level metrics

The majority (94% of codes) of concepts were related to Event Domains, with Conditions and Measurements as predominant ones. The distribution of source domains is described in the table below.

Source Domain Type	Source Domain	Median unique code per voc	Avg unique code per voc	Min unique code per voc	Max unique code per voc	Total unique code per voc	Percentage
Dimension	Geography	13	81	13	217	243	4.62%
Dimension	Race	11	16	5	37	63	1.20%
Dimension	Ethnicity	4	4	4	4	4	0.08%
Dimension	Gender	2	2	2	2	8	0.15%
Event	Measurement	114	116	5	320	1272	24.16%
	Condition	29	84	1	479	2106	40.01%
	Observation	22	62	3	328	1185	22.51%
	Drug	17	34	13	124	236	4.48%
	Meas Value	26	26	2	50	104	1.98%
	Unit	8	9	3	17	36	0.68%
	Procedure	4	4	4	4	4	0.08%
	Route	3	3	3	3	3	0.06%
TOTAL						**5264**	**100%**

Table 2. Source domain metrics (voc=vocabulary)

OMOP does not formally allow capture of information known to be not present. However, our research analytical use case requires the manipulation of negative and unknown facts, therefore, we implemented an approach to extensively pre-coordinate at the source level. This resulted in approximately 80% of the code being generated through permutations with negation/unknown status attributes.

The multilanguage nature of semantic data elements, which resulted from the geographical separation of sites, was curated once the element was introduced to the Vocabulary environment either by using a translation provided by the data producer or by using an automated translation pipeline.

Semantic data mapping

Concepts retrieved from data collection forms (DCFs) undergone semantic mapping according to basic OMOP CDM rules to populate proper landing tables. The table-of-origin, as well as information about the date's completeness was considered as reliable information to distinguish actual vs historical facts. We accepted both pre-coordination and post-coordination as valid options for storing semantic mappings within the OMOP CDM.

The disease activity/severity instruments-related source entries, where the origination (i.e. questionnaire name) of the question-answer was obvious, were covered with both: mapping to clinical facts (incl. historical) as well as represented as measurements. Scenarios that were not straightforward or complex (same data collected in 2 separate areas for different purposes) were treated as clinical facts only to prevent misinterpretation while querying Standard Concepts.

The lack of standardized terminologies/concepts to fully capture the original meaning (semantics) for established analytical use cases was the primary reason for creating concepts within the Registry terminology. The majority of the *de novo* created standard terms belong to the disease activity/severity instrument field.

The main parameters of the created terminology are presented below:

For many of the registries we used the following vocabulary pipeline: sources data codes → custom common terminology → standard terminology. Other registries were supported by concepts in existing OHDSI vocabulary and still others triggered an addition of concepts to OHDSI OMOP extension vocabulary.

OMOP does not formally allow capture of information known to be not present. (negative information). In this project, a specific extension of OMOP standard was applied. We allowed negative information to be represented as well as recorded this uncertain information (unknown whether the fact happened or not when known that the patient was tested/asked). Registries substantiated this use case as a need to distinguish between: 'asked/tested but answer in unknown' vs 'don't know if asked/tested'. This is representative of a turnaround closed world model (usual OMOP) into the <wide> open world model. Other challenges included multiple source languages (besides English).

The distributed data ETL logistics team was responsible for the ETL code execution.

Discussion

As OMOP CDM-based data integration approach is applied to multiple types of data, in this case disease registry data, a few best practices should be considered.

- For this project, registry data from each site was mapped independent of the next. When mapping disease measure scores, ensure the location of these data for each registry and ensure all those concepts are mapped in the same manner across all the registries (i.e., on the registry network level).
- Consider site variation in data collection cardinality (=frequency of data collection over time) for instruments measuring disease activity/severity. Are they repeated measures in some registries while only a one time for other registries.
- Consider carefully data integration implications of situations when a diagnosis is treated as both a condition/symptom of the disease and or a comorbidity elsewhere in the database.
- Registry use case and research context required adjustment of some CDM conventions. Specifically, we allowed concepts that capture negative information (concepts for 'not present' facts).

References

1. Puttmann D, De Keizer N, Cornet R, Van Der Zwan E, Bakhshi-Raiez F. FAIRifying a Quality Registry Using OMOP CDM: Challenges and Solutions. Stud Health Technol Inform. 2022;294:367-371. doi:10.3233/SHTI220476
2. Hallinan CM, Ward R, Hart GK, et al. Seamless EMR data access: Integrated governance, digital health and the OMOP-CDM. BMJ Health Care Inform. 2024;31(1):e100953. Published 2024 Feb 21. doi:10.1136/bmjhci-2023-100953