How vocabulary updates can affect individual OMOP instances

Daniel Park, Elise Gatsby, Benjamin Viernes, Kushan Hewa, Scott L DuVall, Kristine E Lynch

Background

Vocabulary updates to the OMOP common data model (CDM) are frequent, and may include depreciations or additions of CONCEPT_IDs, and changes to CONCEPT_NAME, CONCEPT_CODE or DOMAIN_ID. This is a critical part of ensuring the OMOP CDM remains up to date for controlled vocabularies and relationships. However, such changes can negatively impact an organization's automated Extract-Transform and Load (ETL) process, and careful design and process enforcement is needed to minimize the downstream risk of data gaps and errors. Although using CONCEPT_CODE is "OHDSI's recommended best practice",[1] here we present examples of the downstream consequences of a specific scenario where the CONCEPT_ID for a given concept stays consistent between vocabulary updates but various attributes change (e.g. CONCEPT_NAME, DOMAIN_ID) and describe proactive solutions.

Methods

The process of mapping VA source concepts to OMOP standard concepts varies by what is available in the native source data. Generally, this is done by specifying the DOMAIN_ID and VOCABULARY_ID and matching the text of the CONCEPT_CODE to the text identifier for the value to load. For example, we match the text for provider specialty from source data directly to a CONCEPT_CODE where DOMAIN_ID is Provider and VOCABULARY_ID is NUCC (National Uniform Claim Committee).

We evaluated the number of vocabulary updates that occurred from March 1, 2021 to April 1, 2021 across both the entire CDM and those used within the VA OMOP instance. The type of vocabulary change by volume and the degree to which VA OMOP concepts were impacted by these updates were evaluated.

Results

From March 1, 2021 to April 1, 2021 there were 114,732 vocabulary updates involving static CONCEPT_IDs accompanied by various attributes changes. The type of change varied by concept, but the most common change was a domain change (e.g., drug to observation) followed by name change (e.g., oral to oral route (Figure 1)). Of the total changes, <1% (n=813) corresponded to concepts existing in the VA OMOP instance with SNOMED concepts most affected (Figure 2a). Changes to Route concepts impacted 2.9 billion rows in the Drug Exposure table due to the addition of 'route' to the end of each individual CONCEPT_NAME (e.g., Oral to Oral Route) resulting in a decrease to the proportion of routes mapped from 51% to 1.5% (Figure 2b).

Figure 1. Concept name change for medication route breaks existing source to target mapping

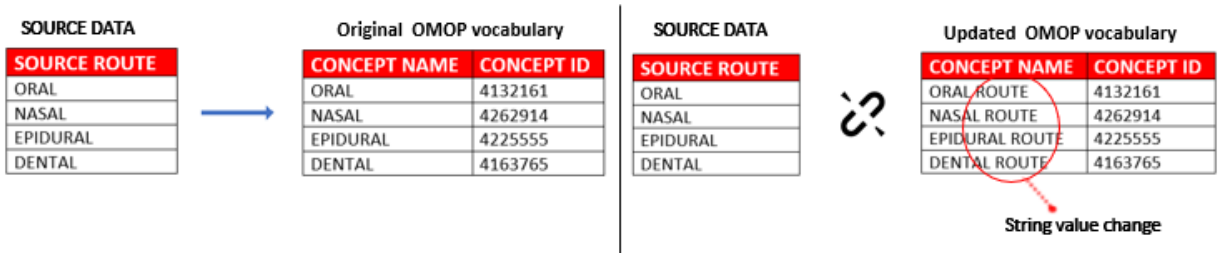
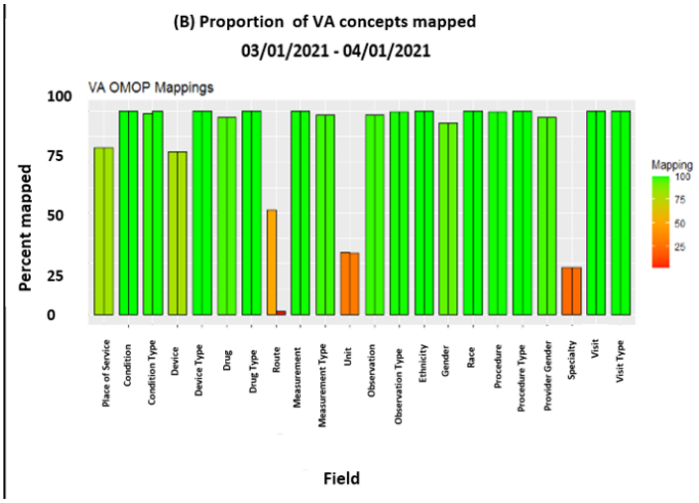


Figure 2. Vocabulary changes from March 1, 2021 through April 1, 2021 (A) and the effect on VA mappings (B)

(A) Frequency of OMOP CDM Vocabulary Updates

03/01/2021 - 04/01/2021

VOCABULARY ID	OMOP VOCABULARY	VA VOCABULARY
ICD10CM	530	349
ICD10PCS	80	80
ICD9CM	115	101
NDC	3,145	44
SNOMED	76,632	239
Others	34,230	-



There are several solutions that could mitigate the impact of these changes, both those conducted from within the OHDSI community and among the site implementers. For OHDSI, extremely careful assessment of the impact of CONCEPT_CODE changes should be conducted, and field advisories should be issued, as these can impact ETL processes severely. Changes in CONCEPT_CODE should be avoided where possible, and where another organization governs those changes (NLM, Regenstrief, etc.), analysis and reporting of the impacts of these changes would be helpful.

For the site implementers, there are number of mitigation strategies that can be employed. First, anticipatory quality assurance can be conducted to compare the proportion of concepts mapped to previous refreshes (Figure 2b). Second, for those meta-data entries in which the mapping to CONCEPT_ID changes, manual review for those used by the CDM instance may be warranted. Third, since VA currently string-matches medication routes to CONCEPT_NAME, when concept name changed the mapping chain broke. This resulted in the percent of Drug Exposure's route mapped decreasing from 51.3% to 1.5% (Figure 2B) and the number of route concepts mapped from 33 to 6. Our future patch will be to assign a CONCEPT_CODE to the string rather than matching string to string.

Conclusion

We demonstrate the local impact of a vocabulary update on mapping proportions across concepts. Ideally, vocabulary changes would be infrequent and minimally affect any given OMOP instance; however, each individual ETL team needs to be able to adapt to content updates. Designing ETL architecture that is minimally dependent on vocabulary stability can reduce the impact of the necessary vocabulary updates.

References/Citations

1. Observation Health Data Sciences and Informatics. (2019 October.) 2019 Tutorial – Extraction, Transformation, and Load Process (ETL). https://ohdsi.org/extraction-transformation-and-load-process-etl/