Methods for Managing Vocabulary Evolution in a Multi-Site Centralized Data Repository

William T Roddy[1] & German Soto[1], Ian Braun, PhD[1], Smith F. Heavner, PhD, RN, FCCM,[1,2] Kanwaljit Singh, MD[1]

[1]Critical Path Institute, Tucson, AZ

[2]Clemson University, Clemson, SC

Background

The International Neonatal Consortium – Real World Data (INC-RWD) project aims to advance evidence-based drug development tools tailored to neonates – a population traditionally underrepresented in clinical research and drug development. The initiative has collected EHR data with a focus on data generated in Neonatal Intensive Care Units. To support multi-site analyses the Critical Path Institute has harmonized these data to the OMOP Common Data Model (CDM). A key factor in selecting the OMOP CDM was its robust standardized vocabularies, which enable semantic interoperability across data sources.

The OHDSI community has engaged in discussions on the critical topic of vocabulary stability and versioning.[1,2] The process of updating source vocabularies, mappings between terms, updating standard concept status, and the domain alignment of concepts can result in variations between harmonized CDM instances. This has shown to be challenging for network studies and illustrates the importance of a unified version of the Standardized Vocabularies.[2] There has been improvement in the management of vocabulary versioning since the Vocabulary Working Group adopted an biannual vocabulary release cycle and published planned changes for each release via the Vocabulary Wiki.[3]

These improvements reflect OHDSI's commitment to this issue, but challenges remain. With no canonical migration path between vocabulary versions, these limitations reinforce the need for community-driven communication, tooling, and shared best practices to mitigate potential negative effects and support accurate vocabulary management. To date the INC-RWD project has relied on a single version of the Standardized Vocabularies. However, additions to source terminologies and other updates from the vocabulary team have begun to necessitate upgrading the harmonized CDM instances. To address these challenges, we present a methodology of upgrading the Standardized Vocabularies and evaluating potential impact on the harmonized CDMs. This method compares two versions of the Standardized Vocabularies to identify impact on ETLs and custom concept terms.

Methods

We have implemented four ETL's on EHR data sources using the v5.0 23-JAN-23 version of the standardized vocabularies. When the term originated from a terminology in the standardized vocabularies, we used the provided standard equivalent. In instances where the source terms were not available in the standardized vocabulary, we created custom concepts and established mappings to standard equivalents using appropriate relationships.[4] These mappings were manually curated in parallel with ETL development to ensure high-fidelity transformations and appropriate domain alignment (e.g., quantitative observations were placed in either the Measurement or Observation domains).

We are currently upgrading our ETLs and custom mappings to the v5.0 27-FEB-25 vocabulary release. To support this upgrade, we developed a series of SQL queries and orchestrated the workflow using dbt to evaluate the differences between the two versions of the Standardized Vocabularies and aid in the update process. One unique aspect of these queries is that they include producing statistics for the impact on specific CDM instances including the number of records and number of patients impacted.

The process included the following steps:

1. *Identify changes in concept validity:*
   For all non-standard to standard mappings used in the ETL, check if the target concept's *invalid reason* has changed. If so, attempt to re-map using mapping chaining based on transitivity rules. Unmappable concepts are flagged for manual review.

2. *Assess domain changes for mapped concepts:*
   For all mapped concepts (direct or chained), check for changes in domain_id. Flag changes that may impact ETL output, prioritizing those likely to cause data loss or misclassification. This may occur due to the new domain_id not belonging to clinical event domain or the new clinical domain not supporting the same columns as the previous data. For example, a change from the 'Condition' to 'Observation' domain is less likely to result in data loss than a change from 'Observation' to 'Meas Value' or 'Drug'.

3. *Manual curation*
   Review all flagged concepts. Identify alternative standard concepts or update mappings as needed. Required ETL logic changes are tracked and implemented via the CDM issue tracker.

4. *Execute and evaluate*
   Apply vocabulary and ETL updates. Rerun the ETL and assess data quality using record counts, concept distributions, custom checks, and the DataQualityDashboard.

Results

We have utilized our queries and update process for four of our CDM's that included over 85-million clinical event records and more than 26,000 patients. These results are based on vocabulary content in the clinical event domains. Table 1 illustrates the change of concept standard status and domain alignment for both the standard and non-standard concepts.

	Standardized Vocabulary Concepts			Custom Concepts		
	Total Concepts	Total CDM Records	N Patients	Total Concepts	Total CDM Records	N Patients
Concepts with invalid targets	680	156,578 (0.18%)	16,498 (62%)	182	2,782,647 (3.2%)	10,927 (41%)
Concepts with chained mappings	639 (94%)	162,585 (0.19%)	16,464 (62%)	143 (78%)	624,890 (0.19%)	7,634 (28%)
Concepts with domain changes	343	85,775 (0.1%)	17,104 (64%)	224	3,473,293 (4%)	13,869 (52%)

Table 1. Table summarizing the number of concepts, CDM records, and individual patients impacted by the vocabulary update

Changes to standardized vocabularies for data in this patient population did have a wide impact over the population (>50%) but impacted a relatively small proportion of observations in the CDMs. Notably, the impact on records associated with custom concepts is much higher than those from standardized vocabularies.

Conclusion

This process has shown to be an effective method for streamlining the review and manual curation of concepts after a change to the Standardized Vocabularies. The process has the potential to be scaled for more focused impact assessments for new versions of the vocabularies to inform the potential impact on ETL logic, additional manual curation, and previous analyses. It may also be possible to extend this framework to additional vocabulary content such as concept sets.

Future work is still needed to further evaluate more complex patterns such as multi-domain mappings or changes in associated “Maps to value” relationships. This also does not account for the possibility that updates to the source terminologies may have introduced new, better targets for our custom relationships. We will continue to evaluate opportunities for improved review of relationship changes and new content that may be more appropriate.

References

1. Peshansky A. Vocabulary changes between versions [Internet]. OHDSI Forums. 2016 Sep 14 [cited 2025 Jul 1]; Available from: https://forums.ohdsi.org/t/vocabulary-changes-between-versions
2. Kostka K. “Egg on face” - a cautionary tale about ongoing OMOP Vocab change management problems [Internet]. OHDSI Forums. 2021 Apr 14 [cited 2025 Jul 1]; Available from: https://forums.ohdsi.org/t/egg-on-face-a-cautionary-tale-about-ongoing-omop-vocab-change-management-problems/
3. Ostropolets A. Release planning [Internet]. GitHub. 2023 Apr 25 [cited 2025 Jul 1]; Available from: https://github.com/OHDSI/Vocabulary-v5.0/wiki/Release-planning
4. Philofsky M. Mapping Custom Source Codes to Standard Concepts: A Comparison of Two Approaches [Internet] OHDSI. 2020 Oct. [cited 2025 Jul 1]; Available from: https://www.ohdsi.org/2020-global-symposium-showcase-18/