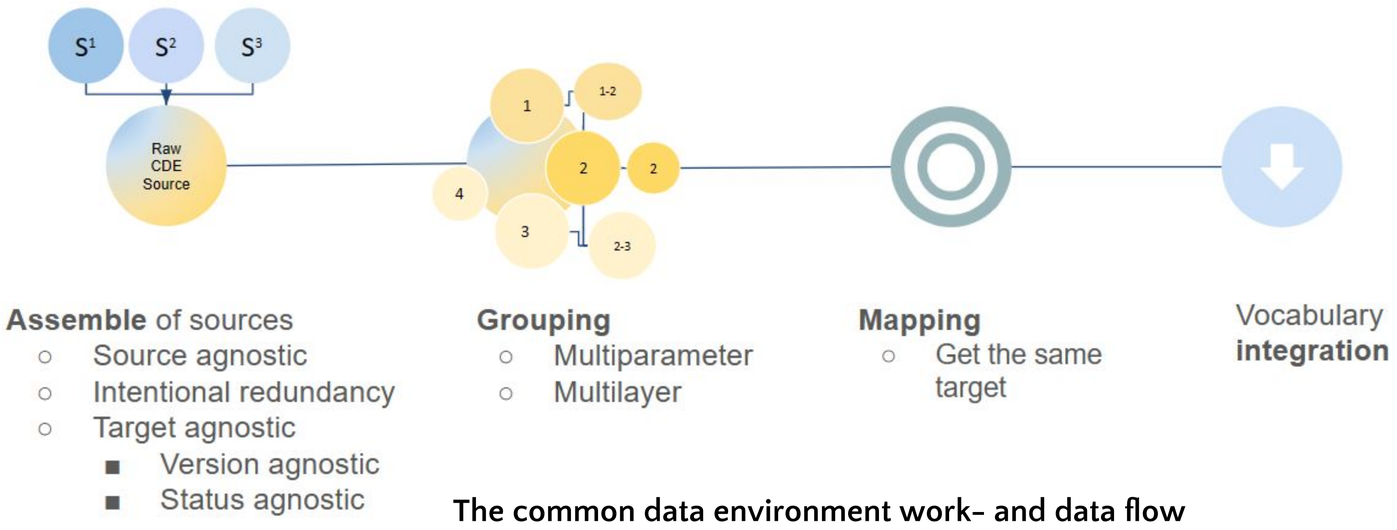
Common Data Environment and ICD family mappings implications

Irina Zherko, Oleg Zhuk, Vlad Korsik, Timur Vakhitov, Alexander Davydov

ODYSSEUS DATA SERVICES INC

Background: Observational health data often contain similar information from various sources that are controlled medical terminologies built by different sets of rules or even free-text, leading to mapping discrepancies within OMOP vocabularies and across different datasets. Mapping curation is time-consuming and requires medical and terminology expertise, highlighting the need for a more structured and semi-automated approach to maintain mappings. The OHDSI Vocabulary team has already put significant efforts to address these challenges, introducing an automated approach to updating mapping targets and developing a common data environment (CDE) for custom data sets [1]. This work takes it a step further by creating the universal CDE for all source data types.

Methods: We have created the universal CDE and implemented the approach for ICD family vocabularies refresh during the ICD overhaul (February 2024 Vocabulary release [2]). CDE is a universal way of data organization for vocabularies, groups of vocabularies (like ICD), and ETL data, irrespective of the origin.



The common data environment work- and data flow

The idea behind the CDE is to create groups of concepts shared by semantic entities (clinical facts):

- Strict group – concepts with the same or insignificantly different (according to OMOP use cases) meaning. This group name is used for mapping.
- Medium group – concepts with close but not identical meanings. Medium groups are used to align mappings for the related concepts.
- Broad group – groups by other ways of grouping (e.g. hierarchical ICD categories or other classifications).

The CDE contains concept codes and concept names, mapping with their sources and metadata. Mapping candidates originate from automated and manual mapping tables, automatic mapping replacements, external mapping sources, such as SNOMED to ICD10CM equivalence tables, and the community contribution pipeline. The format of metadata in CDE is SSSOM-compatible [3, 4] and includes predicate_id ('Maps to equivalent', 'Maps uphill', 'Maps downhill', etc.), mapping_tool and mapping_justification.

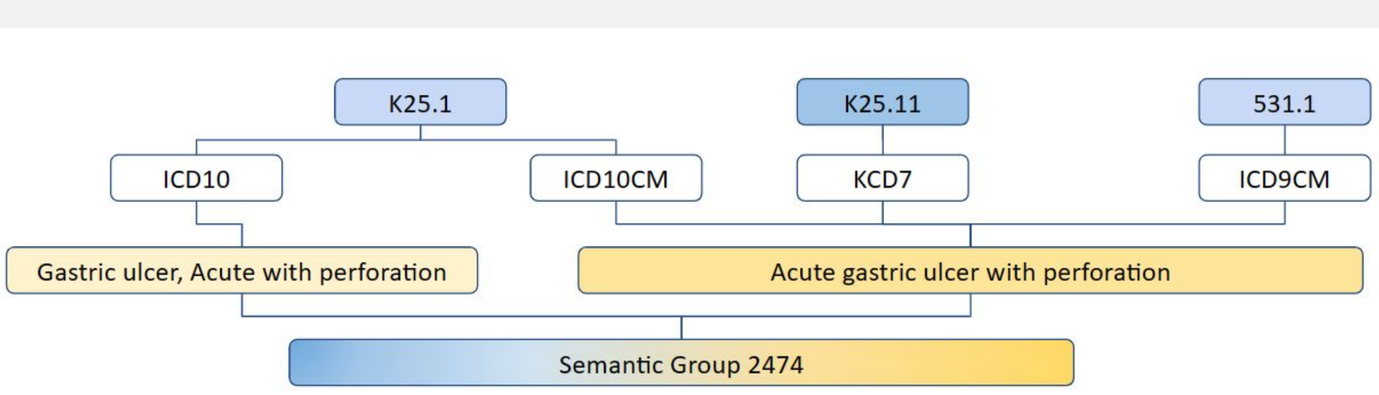
Group_name	Group_code	Mappings_origin	Relationship_id	Relationship_id_predicate	Decision	Target_concept_id	Target_concept_code	Target_concept_name
Other and unspecified abnormal findings in urine	{ICD10CM:R82,ICD10CM:R82.9,ICD10:R82.9}	crs,functions_updated	Maps to	equivalent	1	4099313	27171005	Urinalysis
Other and unspecified abnormal findings in urine	{ICD10CM:R82,ICD10CM:R82.9,ICD10:R82.9}	crs,functions_updated	Maps to value	equivalent	1	4135493	263654008	Abnormal
Other and unspecified cirrhosis of liver, Child-Pugh A	{ICD10CM:K74.60,KCD7:K74.60}	crm	Maps to	equivalent	1	46271811	710065009	Child-Pugh score class A
Other and unspecified cirrhosis of liver, Child-Pugh A	{ICD10CM:K74.60,KCD7:K74.60}	crm, crs	Maps to	equivalent	1	4064161	19943007	Cirrhosis of liver

Example of the CDE at the manual curation step

Results: The developed CDE was implemented for the mapping refresh and harmonization across ICD family vocabularies.

- Data was gathered by inserting all ICD10 (18333 codes), ICD10CM (138708 codes), and ICD9CM (23833 codes) data with mappings tables. For local versions such as ICD10GM (326 codes), ICD10CN (2579 codes), KCD7 (1170 codes), and CIM10 (1117 codes), only data from manual curated mappings were inserted. Unique pairs of source_code and target_concept_id were inserted, with the mapping origin preserved.
- Potential replacement mappings for source codes mapped to non-standard or invalid concepts were inserted.
- The concepts were then grouped into strict groups based on criteria, such as identical source_code_description or identical mappings for unprocessed source codes.
- The longest concept_name from the ICD10 vocabulary was considered the group name if a group contained multiple members. If there were no concepts from ICD10 in a group, the longest concept_name served as the group_name.
- Once the concepts were grouped, the group_name, group_code, and mappings were transferred to table editor software for mapping curation and decision-making.
- After the grouping and mapping curation, all members of a group were assigned the same target_concept_id, which was then implemented through each vocabulary ETL (load_stage scripts).

This process resulted in the creation of **116,705 semantic groups** across ICD vocabularies and **3,032** mapping **conflicts resolution**. 79 groups contained mapping incorporated from the community contributions, resulting in the harmonization of mappings of 350 concepts.



Representation of common data environment implementation result

Conclusion: Application of the Common Data Environment approach to data processing can greatly improve efficiency in process of data mapping and its iterative improvement. This approach can be recommended for both sets of controlled vocabularies or custom data, enhancing the accuracy of data processing, ultimately leading to the creation of higher-quality OMOP or other CDM instances and offering greater consistency and reliability in clinical informatics and research domains.

Scan the QR code for CDE table DDL

References:

1. I. Zherko et al Common data environment for source vocabularies mapping, Conference: OHDSI European Symposium 2022
2. https://github.com/OHDSI/Vocabulary-v5.0/releases/tag/v20240229_1709217174.000000
3. O. Zhuk et al Contribution to the OHDSI Vocabularies, User-Level QC and a New Entity Mapping System SSSOM, Conference: OHDSI European Symposium 2023
4. N. Matentzoglu et all A Simple Standard for Sharing Ontological Mappings (SSSOM), Database, Volume 2022, 2022, baac035, https://doi.org/10.1093/database/baac035