

Evaluating Semantic Loss in EHR Terminology Mapping in Common and Rare Diseases

Melanie Philofsky¹, Bryan J Laraway², Michael G Kahn³, Melissa A Haendel²

¹Epam Systems, ²University of North Carolina, ³University of Colorado

Corresponding author: Philofsky@ohdsi.com

Background:

The increasing reliance on real-world data from electronic health records (EHRs) for observational research and evidence generation underscores the critical importance of accurate terminology mapping.[1] Semantic loss—defined as the reduction or alteration of clinical meaning when mapping between terminologies—poses a significant challenge to the comprehensive understanding of health and disease, particularly when source and target terminologies differ in granularity and structure.

We previously demonstrated semantic loss in mapping Intelligent Medical Object's (IMO) interface terminology to SNOMED CT via both direct and indirect (ICD-10-CM intermediary) strategies[2] (Figure 1). This study assesses the generalizability and applicability of these findings to a broader spectrum of diseases, including both common and rare conditions. By expanding the scope to include distinct healthcare settings and diagnoses, we seek to inform best practices for terminology mapping that support robust, high-fidelity observational research.

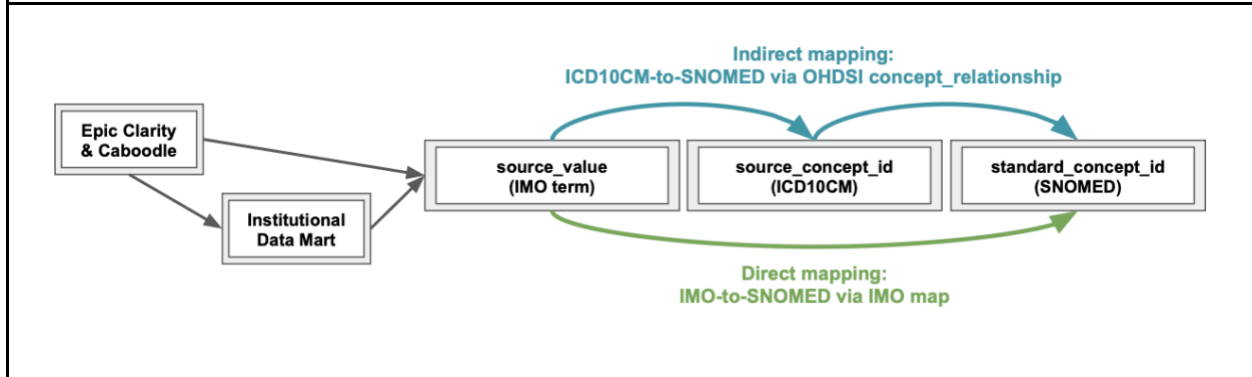
Methods:

We evaluated two widely used mapping strategies [2,3]:

1. **Direct mapping:** IMO terms mapped directly to SNOMED CT using IMO-provided mappings.
2. **Indirect mapping:** IMO terms mapped to SNOMED CT via ICD-10-CM using OHDSI-provided concept_relationship tables.

The study includes the 200 most frequent IMO terms at two academic medical centers for common diseases based on unique patient counts from January 1, 2020, to December 31, 2025, and 200 randomly selected rare diseases identified by matching the Mondo ontology to IMO terms in the EHR. Only current leaf mappings (terms with no subclasses) were used. Two blinded annotators assessed each mapping for concordant (exact lexical or semantic matches) and discordant (preference for direct, indirect, or neither mapping) mapping to the original diagnosis description. Annotators selected the term with the closest semantic meaning, without mapping to a more granular term or losing specificity. Disagreements were resolved by a third annotator. All annotators were blinded to the source institution.

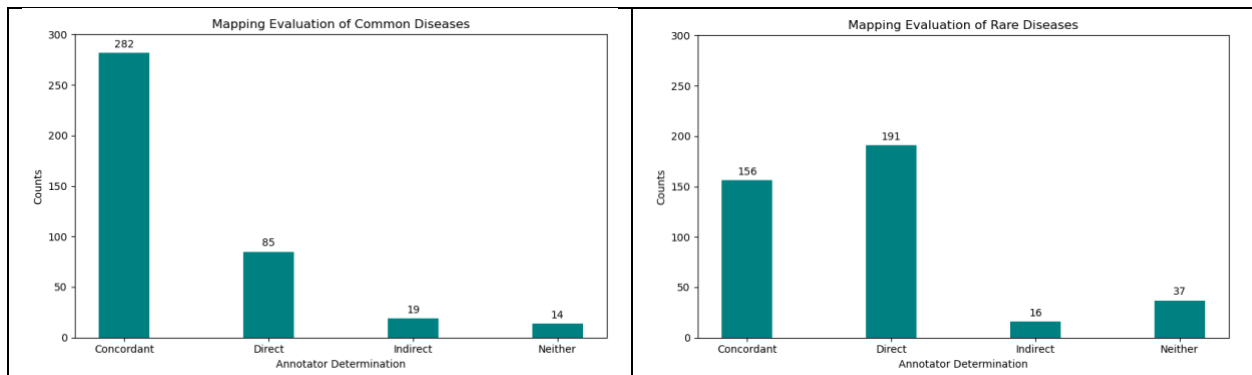
Figure 1: Alternative mapping strategies. The top pipeline uses OHDSI-provided ICD-10-CM-to-SNOMED concept_relationship mappings. This approach is called “indirect mapping” using ICD-10-CM as an intermediate concept. The bottom pipeline uses IMO-provided IMO-to-SNOMED mappings. This approach is called “direct mapping.”



Results:

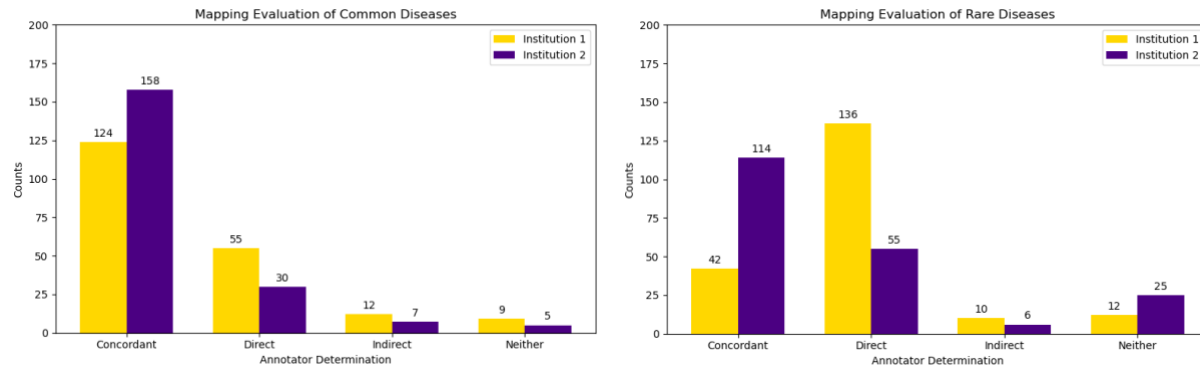
Combined analysis across all annotators revealed concordance rates of 71% (282/400) for common diseases and 39% (156/400) for rare diseases. As illustrated in figure 2, both mapping methods yielded a majority of semantically exact or equivalent target concepts for common conditions. For rare diseases, however, the direct method was consistently preferred and more pronounced in the rare diseases. Rare diseases also exhibited a higher proportion of unacceptable (“Neither”) mappings (3.5% versus 9.3%, respectively).

Figure 2: Best mapping method: common versus rare diseases.



When stratified by institution, Institution 2 demonstrated higher concordance rates for both common (79%) and rare (57%) conditions compared to Institution 1 (62% and 11%, respectively) (Figure 3). Across both sites, direct mapping was markedly preferred over indirect mapping. However, rare diseases had a higher percentage of instances where neither mapping method was acceptable.

Figure 3: Best mapping method: common versus rare diseases by institution



Conclusions:

Understanding the impact of terminology mapping decisions made at Extraction-Transformation-Load (ETL) processes that transform health record data into the OMOP Common Data Model is critical to creating trustworthy and reproducible analytic results[4–7]. This study shows that IMO-to-SNOMED mapping (direct mapping) is preferred over indirect mapping for common and rare diseases regardless of institution. The findings confirm and extend previous work by explicitly examining mapping challenges for common versus rare diseases.

Stratification by institution revealed that Institution 2 achieved higher concordance rates for both common and rare conditions than Institution 1. Although direct mapping was markedly preferred across both sites, Institution 2's ETL process enforced a strict requirement that all mappings correspond to a Mondo rare disease, which possibly introduced selection bias and contributed to its higher observed mapping concordance.

Rare diseases have a larger range of synonyms in active use that reflect the rapidly evolving understanding of the underlying pathologies. A rare disease may have one or more eponyms named after the person(s) who noted the original description, multiple names describing the various associated pathologies, the molecular origin, and more recently, a specific genetic variant. In addition, rare disease terminologies tend to be much “bushier” (more nodes per hierarchy) as more granular molecular understanding evolves over time. As demonstrated by the markedly higher number of terms where neither mapping was acceptable, terminology mapping of rare diseases is highly susceptible to semantic loss, especially if the target terminology has limited rare disease coverage. The incorporation of specialized curated rare disease terminologies, such as the MONDO ontology,[8,9] could reduce semantic loss in these devastating diseases.

References

1. Rijnbeek PR. Converting to a Common Data Model: What is Lost in Translation?: Commentary on “Fidelity Assessment of a Clinical Practice Research Datalink Conversion to the OMOP Common Data Model.” *Drug Saf.* 2014. doi:10.1007/s40264-014-0221-4

2. Philofsky M, Mullen KR, Laraway BJ, Kahn MG, Haendel MA. Maximizing EHR Semantic Meaning for Rare Diseases Utilizing a Direct Mapping Strategy – OHDSI. In. 2025 [cited 2026 Jun 11]. Available from: <https://www.ohdsi.org/2025showcase-141/>
3. Burrows EK, Razzaghi H, Utidjian L, Bailey LC. Standardizing Clinical Diagnoses: Evaluating Alternate Terminology Selection. AMIA Jt Summits Transl Sci Proc AMIA Jt Summits Transl Sci. 2020;2020:71–9. PubMed PMID: 32477625; PubMed Central PMCID: PMC7233070.
4. DeFalco FJ, Ryan PB, Soledad Cepeda M. Applying standardized drug terminologies to observational healthcare databases: a case study on opioid exposure. Health Serv Outcomes Res Methodol. 2013;13:58–67. doi:10.1007/s10742-012-0102-1
5. Voss EA, Makadia R, Matcho A, Ma Q, Knoll C, Schuemie M, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. J Am Med Inform Assoc. 2015;22:553–64. doi:10.1093/jamia/ocu023 PubMed PMID: 25670757.
6. Denney MJ, Long DM, Armistead MG, Anderson JL, Conway BN. Validating the extract, transform, load process used to populate a large clinical research database. Int J Med Inf. 2016;94:271–4. doi:10.1016/j.ijmedinf.2016.07.009 PubMed PMID: 27506144.
7. Hripcsak G, Levine ME, Shang N, Ryan PB. Effect of vocabulary mapping for conditions on phenotype cohorts. J Am Med Inform Assoc. 2018;0:8.
8. Monarch Initiative. Mondo Disease Ontology [Internet]. [cited 2026 Jun 11]. <http://mondo.monarchinitiative.org/>. Available from: <http://mondo.monarchinitiative.org/>
9. Vasilevsky NA, Toro S, Matentzoglou N, Flack JE, Mullen KR, Hegde H, et al. Mondo: integrating disease terminology across communities. Genetics. 2026;232:iyaf215. doi:10.1093/genetics/iyaf215

