# Maximizing EHR Semantic Meaning for Rare Diseases Utilizing a Direct Mapping Strategy

**Melanie Philofsky[1], Kathleen R Mullen[2], Bryan J Laraway[2], Michael G Kahn[3], Melissa A Haendel[2]**

**[1]EPAM Systems, [2]University of North Carolina at Chapel Hill, [3]University of Colorado Anschutz Medical Campus**

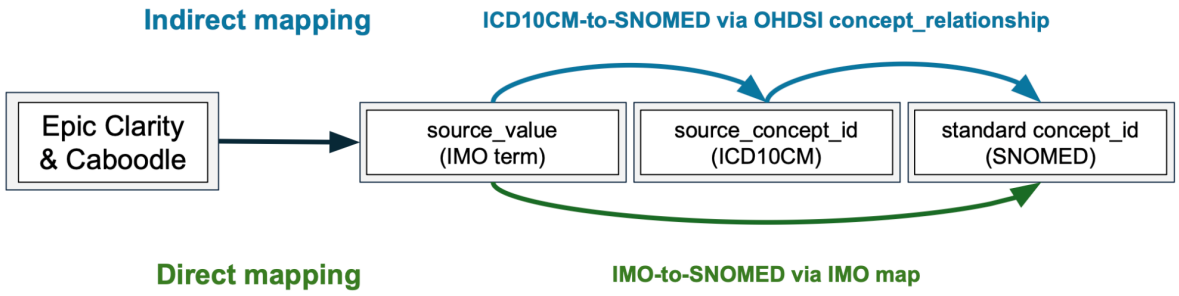
## Background

Many electronic health records (EHRs) provide a clinician-friendly interface terminology that captures the nuances of a patient's diagnosis and observations that cannot be represented in administrative coding systems. (1) In the USA, the widely used Intelligent Medical Objects (IMO) interface terminology contains nearly one million terms that express subtle distinctions in clinical observations and diagnoses. In collaboration with the Monarch Initiative, IMO has incorporated Mondo, an extensive rare disease terminology that contains over 10,000 rare disease concepts. (2,3)

The OMOP common data model (CDM), standardized terminologies, and harmonization best practices enable international-scale real-world health research (4,5). Large data networks could eliminate a major barrier to rare disease research - insufficient patient numbers. (6,7) For rare diseases, the specificity of interface terminologies is often lost when these terms are mapped to administrative or epidemiological coding systems. The potential consequences are significant: the analytic cohort may be heterogeneous and inaccurate evidence generated, or the study simply cannot be performed due to low statistical power.

Conceptually, the OHDSI CDM contains two versions of every data element: a "source" string which exactly represents the original data (source_value) and a "standard" code (condition_concept_id, drug_concept_id, etc), the result of applying OHDSI transformation and harmonization rules. An intermediate concept, the source_concept_id, represents a coded version of the original string. The source_concept_id may be a code from any of the 112 vocabularies supported by OHDSI or an institution-specific code. The OMOP concept_relationship table provides a mapping from an OHDSI concept to standard concept_ids. Figure 1 (top) shows this pipeline for the OMOP condition_occurrence table. The source_concept_id is derived from the ICD-10-CM vocabulary, and the standardized SNOMED code is derived from the OHDSI concept_relationship table. (5,8) This approach is called indirect mapping because the ICD-10-CM code, rather than the IMO term, maps to the standard OMOP concept.

**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."**




There has been much discussion about potential information loss by transforming the original clinician-entered term into a standard concept_id. (9–11) Of interest here is mapping an IMO term into an intermediate source_concept_id, followed by a second mapping into the standard concept_id using the OHDSI concept_relationship table. An alternative approach (Figure 1, bottom) is to map the term directly to the standard concept_id. This is called direct mapping because no intermediate concept is used to obtain the standard concept. One study compared these two mapping strategies using the IMO mapping to standard SNOMED concepts. (12)

In this study, we evaluate the same mapping strategies (Figure 1 top versus Figure 1 bottom) with a focus on rare disease diagnoses.

## Methods

**Data**: The study data set was created from an instance of OMOP CDM V5.4 with vocabulary v20250227. The ETL pipeline uses Epic Caboodle to implement Figure 1 (top). Only concepts with domain_id = "Condition" are mapped to the condition_occurrence table. The study used all condition_occurrence records with visit_start_dates between January 1, 2020, and December 31, 2024. IMO terms were extracted from the condition_occurrence.source_value; source ICD-10-CM codes from the condition_occurrence.source_concept_id; and standard SNOMED codes from the condition_occurrence.concept_id. IMO terms with fewer than 9 patients were

eliminated. OMOP concept_ids were converted to human-readable concept_code strings.

To identify rare diseases, we used the Mondo ontology (mondo-rare.obo June 3, 2025 release) created by the Monarch Initiative. (2,13) Combining class_labels and exact synonyms yielded 61,353 unique rare disease labels. Of these, 2100 rare disease labels matched the condition_source_value for 10 or more unique patients.

As a control group, we included the 200 most frequent IMO terms based on unique patient counts.

**Maps**: The OHDSI-provided concept_relationship table mapped ICD-10-CM codes to standard SNOMED codes (Figure 1, top). IMO to SNOMED mappings (Version 4/1/2025) directly mapped IMO terms into standard SNOMED codes (Figure 1, bottom). Only current leaf mappings (i.e., terms with no subclasses) were used. Historical mappings were not assessed.

**Processing**: For each IMO term, the two mapping methods illustrated in Figure 1 were applied. One-to-many mappings were concatenated to a single result.

**Analysis:** For each IMO term, both mapping results were randomly assigned to a “Map1” or “Map2” column. Two blinded annotators determined which mapping more completely captured the information in the IMO term. A third independent reviewer adjudicated discrepancies. Relative proportions were calculated. No statistical inferences were performed.

## Results

After applying both strategies, 2300 unique IMO terms were mapped (2100 rare disease diagnoses; Top 200 diagnoses). Both mapping strategies yield identical standard concept_ids in 1037 instances (45%), with a marked difference in identical mappings for rare diseases (43% identical) versus top 200 diseases (68% identical). There were 1,200 discordant mappings for rare diseases and 63 discordant mappings for Top200 diseases.

Removing concordant mappings, 1,263 discordant mappings were assessed for preferred mappings. For rare diseases, direct IMO mappings were preferred 92% (1107/1200); for Top200 diagnoses, direct IMO mappings were preferred 73% (46/63) (Table 1).

**Table 1: Preferred mappings for 1,263 discordant mappings.**

| | Indirect (ICD-10-CM) | Direct IMO mapping preferred |
|---|---|---|

| **Table 1: Preferred mappings for 1,263 discordant mappings.** | | |
|---|---|---|
| | mapping preferred (Figure 1, top) | (Figure 1, bottom) |
| Rare disease diagnoses (n=1,200) | 93 | 1107 |
| Top200 disease diagnoses (n=63) | 17 | 46 |

Various mapping outcomes are shown in Table 2:

| **Table 2: Examples of mapping outcomes. For each example, the preferred mapping is in bold font.** | | | |
|---|---|---|---|
| **Indirect mapping semantic loss due to poor intermediate (ICD-10-CM) code** | | | |
| IMO Term | Intermediate ICD-10-CM code | Standard SNOMED concept | |
| | | Indirect mapping | Direct mapping |
| Mesocardia (HC CODE) | Q24.8 | Congenital heart disease | **Mesocardia** |
| Shone syndrome (HC CODE) | Q24.8 | Congenital heart disease | **Shone complex** |
| Li-Fraumeni syndrome | Z15.01 | Genetic predisposition | **Li-Fraumeni syndrome** |
| Gardner's syndrome (HC CODE) | Q87.89 | Congenital malformation syndrome | **Gardner syndrome** |
| Noonan's syndrome (HC CODE) | Q87.19 | Congenital malformation syndromes associated with short stature | **Noonan's syndrome** |
| | | | |
| **Direct mapping semantic loss due to poor IMO mappings** | | | |
| Livedoid vasculitis | L95.0 | **Idiopathic livedo reticularis with summer ulceration** | Idiopathic livedo reticularis |

| **Table 2: Examples of mapping outcomes. For each example, the preferred mapping is in bold font.** | | | |
|---|---|---|---|
| Fructose intolerance | E74.10 | **Fructose metabolism disorder** | Intolerance to food |
| | | | |
| **Synonym mappings: No obvious "better" mapping** | | | |
| Mobitz II | I44.1 | **Second degree atrioventricular block** | **Mobitz type II atrioventricular block** |
| | | | |
| **No acceptable mappings** | | | |
| Glucose intolerance | E74.39 | Impaired intestinal carbohydrate absorption | Disorder of carbohydrate metabolism |
| Molar pregnancy (HC CODE) | O02.0 | Disorder of product of conception | Hydatidiform mole, benign |
| Bile duct adenocarcinoma (HC CODE) | C24.0 | Primary malignant neoplasm of extrahepatic bile duct | Bile duct proliferation \| Malignant adenomatous neoplasm |
| Glomus tumor | D18.00 | Hemangioma | Neuroendocrine neoplasm |
| Levocardia (HC CODE) | Q24.1 | Situs inversus with levocardia | Sinistrocardia |

## Conclusion

Both indirect and direct mapping resulted in the same standard concept for 45% IMO terms. However, there was a marked difference in concordant mappings between common diseases (Top200: 68%) versus rare diseases (43%). This stark difference underscores the urgent need to develop more nuanced mapping strategies for rare diseases that retain unique diagnoses.

With discordant mappings, there was a strong bias for blinded annotators to prefer direct mappings. The bias was stronger for rare disease diagnoses (92% direct mapping preference) than for common diagnoses (61%). This finding is not surprising given the smaller size (and therefore semantic breadth) of ICD-10-CM codes used in indirect mapping.

Based on these findings, we recommend OHDSI sites that use IMO as the EHR interface terminology also leverage IMO-provided direct mappings into SNOMED rather than the indirect mapping approach supported by the OHDSI concept_relationship table. Directly aligning mappings to the terminology selected by front-line clinicians ensures that the resulting standard concept captures the intended meaning of the patient-provider interaction. Creating concept sets using ICD-10-CM codes that are mapped to standard SNOMED concepts results in the same significant loss in cohort specificity that can be avoided using standard SNOMED concepts directly. Limitations include not evaluating non-leaf nodes, historical mappings, and other interface terminologies.

## References


1. Rosenbloom ST, Miller RA, Johnson KB, Elkin PL, Brown SH. Interface terminologies: facilitating direct entry of clinical data into electronic health record systems. J Am Med Inform Assoc. 2006 May;13:277–88.

2. Monarch Initiative [Internet]. [cited 2025 Jun 23]. Available from: https://github.com/monarch-initiative

3. Monarch Initiative. Mondo Disease Ontology: Summary statistics [Internet]. 2025 [cited 2025 Jun 28]. Available from: http://mondo.monarchinitiative.org/#stats

4. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc. 2012 Jan 1;19(1):54–60.

5. Voss EA, Blacketer C, Van Sandijk S, Moinat M, Kallfelz M, Van Speybroeck M, et al. European Health Data & Evidence Network—learnings from building out a standardized international health data network. J Am Med Inform Assoc. 2023 Dec 22;31(1):209-219.

6. Reinecke I, Zoch M, Reich C, Sedlmayr M, Bathelt F. The usage of OHDSI OMOP – A scoping review. In: Röhrig R, Beißbarth T, König J, Ose C, Rauch G, Sax U, et al., editors. Studies in Health Technology and Informatics [Internet]. IOS Press; 2021 [cited 2023 Apr 23]. Available from: https://ebooks.iospress.nl/doi/10.3233/SHTI210546

7. Observational Health Data Sciences and Informatics. The Book of OHDSI [Internet].

[cited 2025 Jun 23]. Available from: https://ohdsi.github.io/TheBookOfOhdsi/

8. 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 Mar;13(1):58–67.

9. Zhou X, Murugesan S, Bhullar H, Liu Q, Cai B, Wentworth C, et al. An Evaluation of the THIN Database in the OMOP Common Data Model for Active Drug Safety Surveillance. Drug Saf. 2013 Feb;36(2):119–34.

10. Matcho A, Ryan P, Fife D, Reich C. Fidelity assessment of a clinical practice research datalink conversion to the OMOP common data model. Drug Saf [Internet]. 2014 Sep 4 [cited 2014 Sep 4]; Available from: http://link.springer.com/10.1007/s40264-014-0214-3

11. 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 [Internet]. 2014 Sep 4 [cited 2014 Sep 4]; Available from: http://link.springer.com/10.1007/s40264-014-0221-4

12. 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.

13. GitHub: Monarch-Initiative [Internet]. [cited 2025 Jun 27]. Mondo Release v2025-06-03. Available from: https://github.com/monarch-initiative/mondo/releases/tag/v2025-06-03