

Unlocking the latent semantic graph of the OHDSI vocabularies as a reasoning backbone for grounded clinical AI

Nico Loesch, PhD¹, Ghazaleh Niknam Shirvan^{1,2}, PhD, Richard D. Boyce³, PhD, Georgina A. Kennedy, PhD^{1,2}

¹University of New South Wales, Sydney, Australia; ²Ingham Institute for Applied Medical Research, Sydney, NSW, Australia; ³University of Pittsburgh, Pittsburgh, PA, USA

Introduction

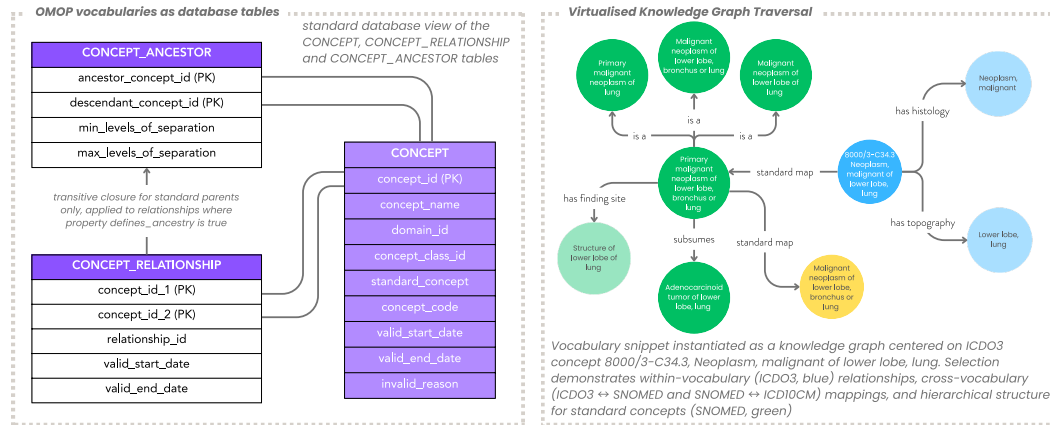
The Observational Health Data Sciences and Informatics (OHDSI) community develops open infrastructure for conducting large-scale observational research across heterogeneous healthcare databases. Central to this is the OMOP Common Data Model (CDM), which standardizes the representation of clinical data so that analyses can be executed consistently across institutions and data sources¹. Supporting this is a comprehensive system of vocabularies that integrates existing and bespoke clinical terminologies such as SNOMED CT, RxNorm, LOINC, and ICD into a unified framework.

Beyond data standardization, the vocabulary tables implicitly encode a large biomedical knowledge graph. Relationships between concepts link entities across clinical domains through hierarchical, compositional, and associative structures, despite the fact that the system was not designed as a general-purpose reasoning framework. Because the semantics of a given relationship depends on its source terminology, naïve graph traversal can produce ambiguous or incorrect inference. As a result, this latent reasoning capability remains largely unrealized. This is a gap worth closing. Deterministic, interpretable reasoning over clinical terminologies is increasingly important for grounding the outputs of LLMs and agentic AI systems operating in biomedical contexts. These systems are being applied to tasks such as cohort identification, phenotype validation and clinical decision support, where concept grounding errors carry real consequences. Embedding similarity alone, the dominant approach for concept resolution today, offers no guarantees of semantic correctness and provides no interpretable audit trail for how a concept was selected.

This work introduces *omop-graph*², a Python library that exposes a virtual knowledge graph over the OHDSI vocabulary system. This provides a framework for interpretable graph traversal and deterministic validation of concept grounding, while preserving compatibility with existing relational CDM deployments.

Methods

The OMOP vocabulary schema represents a semantic network in relational form through the **CONCEPT**, **CONCEPT_RELATIONSHIP**, and **CONCEPT_ANCESTOR** tables (Figure 1), with concepts as graph nodes, and relationships as directed edges. Rather than materializing a separate graph database, *omop-graph* implements a virtual knowledge graph layer³, operating directly on the relational vocabulary tables. This design allows reasoning workflows to be deployed within existing CDM infrastructure without requiring additional graph storage systems.



Relationship Classification: *omop-graph* introduces a classification layer that groups relationship types according to their reasoning behavior. Relationships are organized into functional classes including:

- **Identity relationships**, representing equivalence mappings between concepts;
- **Hierarchical relationships**, subsumption/inheritance traversal for concept generalization/specialization;
- **Compositional relationships**, representing structural part-whole relationships such as drug ingredients;
- **Associative relationships**, capturing clinically meaningful links between biomedical entities;
- **Attribute relationships**, representing concept modifiers such as stage or severity.

Relationships not in these classes are excluded from reasoning workflows by default, converting the heterogeneous vocabulary relationship space into a controlled reasoning graph with explicitly defined traversal semantics.

Concept Grounding Pipeline: To resolve natural language inputs to OMOP concepts, *omop-graph* implements a configurable resolver pipeline. Candidate concepts are retrieved via an ordered sequence of matching strategies including exact label, partial matching, and embedding-based semantic similarity. Once identified, traversal across identity relationships resolves non-standard or source-specific concepts to canonical standard OMOP concepts.

Graph Traversal and Path Scoring: Reasoning queries are executed through SQL query construction over the relational vocabulary tables. Traversal paths are evaluated using configurable scoring constraints that can prioritize specific relationship classes, vocabulary sources, domains, or standardization states. This enables reasoning behavior to be tuned for specific analytic tasks while preserving full traceability of inference paths.

Results

omop-graph is implemented as an open-source Python library operating directly over existing CDM vocabulary tables, requiring no graph-specific infrastructure. The framework enables two primary classes of operations over the OMOP vocabulary system. The concept grounding pipeline resolves natural language or semi-structured model outputs to canonical standard OMOP concepts via the resolver pipeline and identity relationship traversal, and constrained graph reasoning, which enables controlled traversal across hierarchical, compositional, and associative relationships relationship types with fully traceable inference paths.

A key design outcome is the separation of reasoning behavior from the underlying vocabulary schema. Because relationship classification is maintained independently of the CDM tables, the framework can be incrementally extended as additional relationship types are evaluated. This allows distributed contribution from domain experts without coordinated schema changes. The same vocabulary graph can support different reasoning profiles depending on task requirements, with path-scoring constraints tuning traversal behavior between tight concept grounding and broader associative reasoning.

Discussion

The OMOP vocabulary system has historically functioned as a passive reference terminology for standardization and hierarchical expansion, but not as an active reasoning substrate. *omop-graph* addresses this by introducing a controlled interpretation layer, making the latent semantic structure of the vocabularies available for principled computational reasoning. This is particularly relevant as analytic pipelines in learning health system contexts increasingly incorporate LLMs and agentic AI systems whose outputs require grounding to auditable, standardized concepts. Existing approaches to this problem tend to rely on embedding similarity alone, which offers no guarantees of semantic correctness and does not provide an interpretable justification for concept selection. By contrast, *omop-graph* produces traversal paths that are scored against explicit criteria and traceable to the underlying vocabulary relationships, which is a prerequisite for trustworthy deployment in clinical research workflows. The current implementation reflects an initial classification of relationship types across core vocabularies. Scaling coverage across the full breadth of the OHDSI vocabulary system will require community contribution, and evaluating reasoning quality across diverse analytic use cases remains the immediate priority for future work.

References

1. Reich C, Ostropolets A, Ryan P, Rijnbeek P, Schuemie M, Davydov A, Dymshyts D, Hripcsak G. OHDSI Standardized Vocabularies—a large-scale centralized reference ontology for international data harmonization. *Journal of the American Medical Informatics Association*. 2024 Mar 1;31(3):583-90.
2. Australian Cancer Data Network. *omop-graph* [Internet]. GitHub repository; [cited 2026 Mar 9]. Available from: <https://github.com/AustralianCancerDataNetwork/omop-graph>
3. Xiao G, Ding L, Cogrel B, Calvanese D. Virtual knowledge graphs: An overview of systems and use cases. *Data Intelligence*. 2019 Jun 1;1(3):201-23.