

An Explainable, Query-Time Latent Semantic Graph View of the OMOP Vocabulary for Transparent Concept Reasoning

Ghazaleh Niknam-Shirvan¹² · Nico Loesch¹ · Georgina A. Kennedy¹²

THE MAIN FINDING

omop-graph transforms the rich, latent semantic structure of OMOP vocabulary tables into a query-time Virtual Knowledge Graph, enabling transparent, auditable concept grounding for phenotyping and AI-assisted workflows. A principled bridge between probabilistic AI and deterministic graph reasoning, without new databases, pre-materialisation, or schema changes.

THE PROBLEM

The OMOP vocabulary is semantically rich, but hard to reason over safely.

- CONCEPT_RELATIONSHIP aggregates thousands of relationship types from dozens of vocabularies
- Some encode real clinical meaning: subsumption, equivalence, composition
- Others exist purely for vocabulary maintenance
- Naïve traversal produces ambiguous or incorrect inference

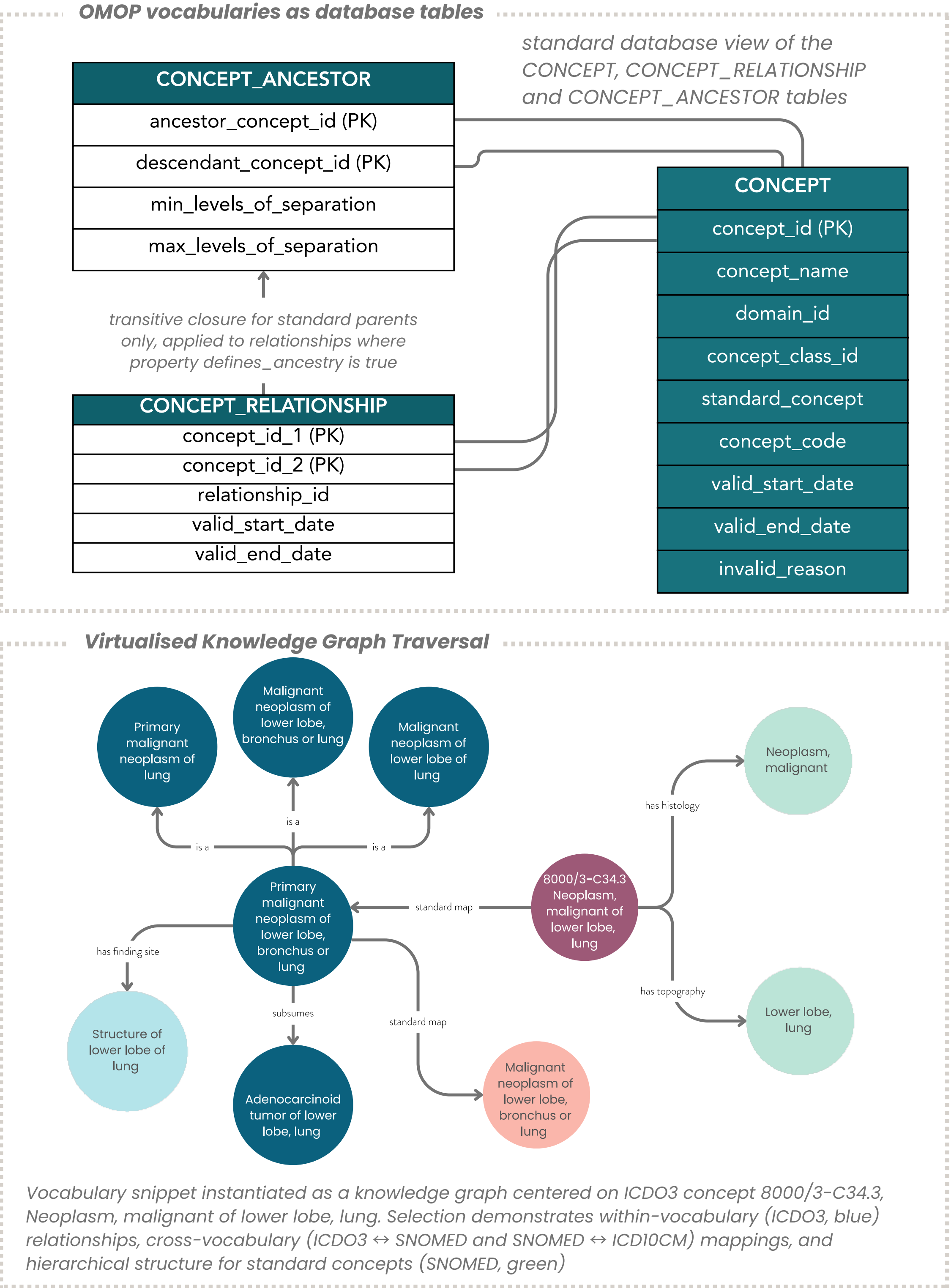
As LLMs and agentic AI enter clinical workflows, principled, auditable concept grounding becomes critical.

- Phenotype development, cohort curation, and the outputs of LLMs and agentic AI systems all depend on constrained, auditable concept grounding.
- In biomedical contexts, grounding errors propagate directly into cohort definitions and clinical evidence.

WHAT WE PROPOSE & HOW IT WORKS

A virtual graph layer — zero infrastructure changes

CONCEPT, CONCEPT_RELATIONSHIP, and CONCEPT_ANCESTOR tables form a semantic network: concepts as nodes, relationships as directed edges.



RELATIONSHIP CLASSIFICATION LAYER

- Identity** → equivalence mappings (e.g. Maps to, Replaced by)
- Hierarchical** → generalisation / specialization (e.g. Is a, Subsumes)
- Compositional** → part-whole structure (e.g. drug → ingredient)
- Associative** → clinical relationships (e.g. treats, causes)
- Attribute** → concept modifiers (e.g. stage, severity)

GROUNDING PIPELINE

- 1 Resolver pipeline**
Exact label, partial match, embeddings. All candidates collected and passed to ranking.
- 2 Identity traversal**
Non-standard or source-specific candidates resolved to canonical standard concepts.
- 3 Constraint filtering**
Enforce domain, vocabulary, standardness, and hierarchy constraints. Each rejection documented.
- 4 Path scoring + ranking**
Each candidate scored as follows:
Relevance (embedding similarity × textual match score) – Parsimony (penalty increases with graph depth - separation) + Broadness (bonus for concepts with more ancestors -> Best candidate returned with full reasoning trace.

End-to-End GROUNDING EXAMPLE

1 INPUT QUERY

CPT CODE

"Radiation treatment delivery, >=1 MeV; complex"

Free-text CPT billing description submitted as a natural language query to the grounding pipeline. No OMOP vocabulary knowledge required on the user side.

2 EMBEDDING HITS

EMBEDDING: 5 HITS

- Delivery of a fraction of complex radiotherapy on megavoltage machine
- Megavoltage treatment for complex radiotherapy
- Radiotherapy delivery
- Delivery of a fraction of simple radiotherapy on megavoltage machine
- Delivery of a fraction of interstitial radiotherapy

3 GRAPH TRAVERSAL

omop_graph walks the OMOP vocabulary graph from the top embedding hit to its canonical standard form.

Embedding top hit

44790405 Delivery of a fraction of complex radiotherapy on a megavoltage machine

Concept same_as

Ground concept

44793226 Megavoltage treatment for complex radiotherapy

Resolver: PARTIAL · Vocab: SNOMED · Standard: ✓

The graph overrides the embedding's top hit, returning the semantically correct canonical concept instead of the literal match.

4 SCORED OUTPUT

WINNER . RANK 1

Megavoltage treatment for complex radiotherapy

Embedding score	0.812
Text similarity	0.160
Relevance	0.715
Parsimony penalty	-0.030
Broadness bonus	+0.020
TOTAL SCORE	0.7042

AI-native. Designed to work with modern AI tooling out of the box.

CLINICAL TERM

"heart attack"

without structured reasoning

RESULT

ambiguous / wrong / unexplained

omop-graph

STANDARD OMOP CONCEPT

Myocardial infarction

SNOMED · 4329847 · Condition

"Heart attack" matches dozens of concepts across domains and vocabularies. Without structured reasoning over the vocabulary, there is no principled way to select the right one, or explain why.

The vocabulary knows the answer, if you ask it right.



Github repo

QUESTION | COLABORATION

georgina.kennedy@unsw.edu.au
n.loesch@unsw.edu.au
g.niknam_shirvan@unsw.edu.au

AFFILIATIONS

¹University of New South Wales, Sydney, Australia
²Ingham Institute for Applied Medical Research, Sydney, NSW, Australia
³University of Technology, Sydney, Australia

