Development of a Graph Model for the OMOP Common Data Model

Mengjia Kang, MS[1], Jose A. Alvarado-Guzman, MS[2], Luke V. Rasmussen, MS[1], Justin B. Starren, MD, PhD[1]

[1]Northwestern University, Feinberg School of Medicine, Chicago, Illinois; [2]Neo4j, Inc., San Mateo, California

Background

- Current phenotyping and systems biology research requires not only integration of large volumes of Electronic Health Record (EHR) and multi-omics data, but also capturing the multitudes of relations among the concepts.
- Graph databases have emerged as a promising technology for such tasks, supporting not only local analysis but also global analysis leveraging graph algorithms like Centrality, Community Detection, Path Finding or Node Embeddings.
- EHR data is rarely available in a graph format unfortunately. While a naïve row-to-node conversion is possible, the resulting graph is typically attribute-heavy, resulting in suboptimal performance.
- To address this limitation, we developed a modelling method to convert data form the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) to the Neo4j [www.neo4j.com] graph property model.

Research Objectives

The goal is to build the graph knowledge database and use the network topology to find out global patterns in EMR to support clinical decisions, in our case a Pneumonia study - The Successful Clinical Response in Pneumonia Therapy (SCRIPT). The database will be further enlarged to include survey data and multi-omics (genomics, metagenomics, transcriptomics, etc.) dataset.

Methods

Our overall strategy was to encode as much information as possible in the edge topology to take advantage of the intrinsic strengths of the graph database. In general, nominal and categorical data were converted to nodes; foreign keys to edges; and numerical values to node or edge properties as appropriate. We also implemented self-directed relationships RELATED_TO and NEXT on the Concept and VisitOccurrence node separately. The former defines the nature and type of direct relationships between any two Concepts and the later builds up the patient journey.

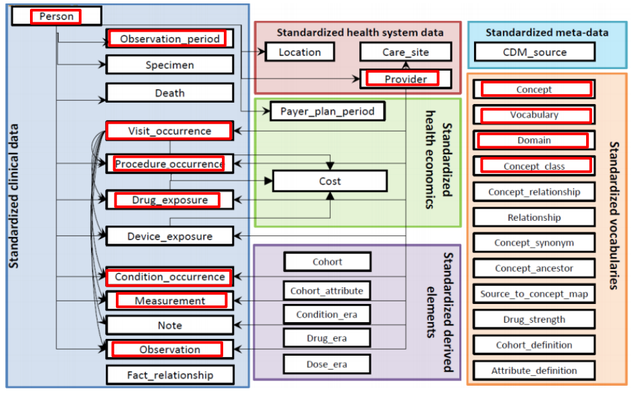


Figure1. OMOP CDMv5.0.1 The highlighted boxes are the tables included in the SCRIPT study (https://github.com/OHDSI/CommonDataModel/wiki)

Figure 2a. Graph schema in whiteboard view

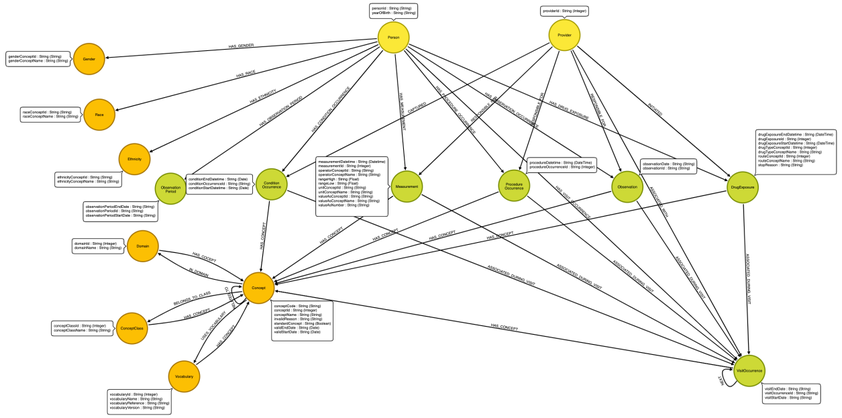


Figure 2b. Graph schema in hierarchical view

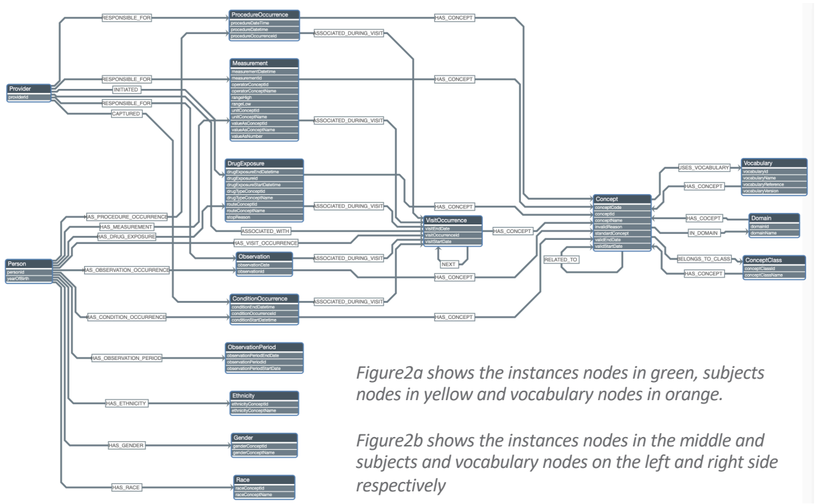


Figure2a shows the instances nodes in green, subjects nodes in yellow and vocabulary nodes in orange.

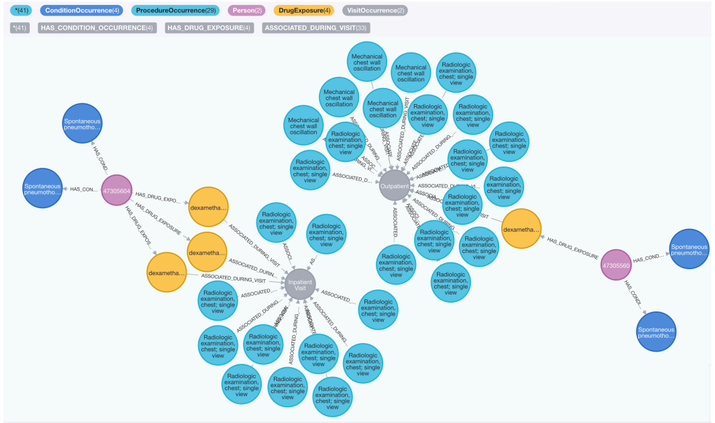
Figure2b shows the instances nodes in the middle and subjects and vocabulary nodes on the left and right side respectively

Acknowledgements

This work is sponsored by NIAID grant 5U19AI135964-04, with support from NU SCRIPT investigators

Figure 3. Local analysis example

```
match path = (c:ConditionOccurrence)←[r0:HAS_CONDITION_OCCURRENCE]-(p:Person)-[r:HAS_DRUG_EXPOSURE]→
 (d:DrugExposure)-[r1:ASSOCIATED_DURING_VISIT]→(v)-[r2:ASSOCIATED_DURING_VISIT]-(p2:ProcedureOccurrence)
where d.drug_concept_name = 'dexamethasone' and p2.procedure_concept_name contains 'chest' and
 c.condition_concept_name ='Spontaneous pneumothorax'  return path limit 100;
```



Results

Our finalized graph property model was implemented using a local installation of Neo4j 4.0.2 Community Edition. It includes 16 types of nodes (entities) and 22 types of edges (relationships) as well as 55 node properties. This model contains on average 3.44 attributes per node. This work is available in both a markdown and Cypher query language format in our GitHub repository, [https://github.com/NUSCRIPT/OMOP_to_Graph].

Limitations

- The current graph schema works perfect for local pattern matching but not necessarily in applying graph algorithms like community detection and node similarity.
- Our current graph database is lacking medication administrations, but we are actively expanding the dimensions of data.

Conclusions

Although more data preprocessing is required to load the data into our graph property model than the naïve row-to-node conversion method, previous work has demonstrated that the analytics performance will be greatly improved. This model also reduces redundancy by eliminating the denormalization (Foreign Keys) that is often added to relational databases (e.g., person_id occurs in all other OMOP tables). The model was developed for the SCRIPT project, but the transforms can be applied to other OMOP CDM v5.x databases.