Using Semantic Queries For Cohort Discovery Across Research Networks

Amanda Hicks, PhD[1], William R. Hogan, MD, MS[1], Zhe He, PhD, MS[2], Josh Hanna, MS[1], Betsy Shenkman, PhD, MSN[1], Jiawei Yuan, PhD[3], and Jiang Bian, PhD, MS[1]

[1]Univsesity of Florida, Gainesville, Florida; [2]Florida State University, Lake City, Florida; [3]Embry-Riddle Aeronautical University, Daytona Beach, Florida

Abstract

Datasets from clinical research networks (CRNs) such as the National Patient-Centered Clinical Research Network (PCORnet) and the Accrual to Clinical Trials (ACT) are rapidly growing in both number and variety. This raises the question, how can we best integrate heterogeneous datasets? Even with common data models (CDM), we still face the problem of how to query across networks using different data models. Thus, we propose a design of a framework that uses Semantic Web technology (e.g., ontology and semantic query) to query across different CDMs and demonstrate its potentials using realistic use cases for cohort discovery.

Background

The last few years have witnessed an increasing number of CRNs building immense collections of electronic health records (EHRs), claims, and patient-reported outcomes (PROs).

Even when data are transformed and integrated into a structured format using a CDM, we still face the problem of how to query across networks using different data models. One approach is to develop a superset CDM, but Extract-Transform-Load (ETL) data from different networks into a superset CDM is not cost-effective.

- Creating a superset CDM costs time and effort.
- Keeping a superset CDM up-to-date with individual CDMs is difficult.
- ETL processes introduce data and information quality issues.
- Duplicating the same data increases data privacy risks.

Even though many CDMs leverage ontologies and standard terminologies to give the data model some semantics, very little work has focused on using the Semantic Web technology to query across different CDMs.

Methods

The framework we propose is modeled in Figure 1. In essence, we align data elements in each individual CDM to existing biomedical ontologies, develop a cohort discovery ontology (CDO) framework reusing those ontologies as modules, and include domain specific classes and relations where necessary for individual studies as study specific application ontologies. We then use semantic queries (e.g., using the SPARQL Protocol and RDF Query Language) instead of conventional SQL queries for cohort discovery.

As a demonstration, we modeled demographic data from OMOP and PCORNet CDM using the Ontology of Medically Related Social Entities (OMRSE), analyzed the classes required for cohort discovery, and added new classes to OMRSE where necessary. We constructed SPARQL queries that utilize these demographic data.

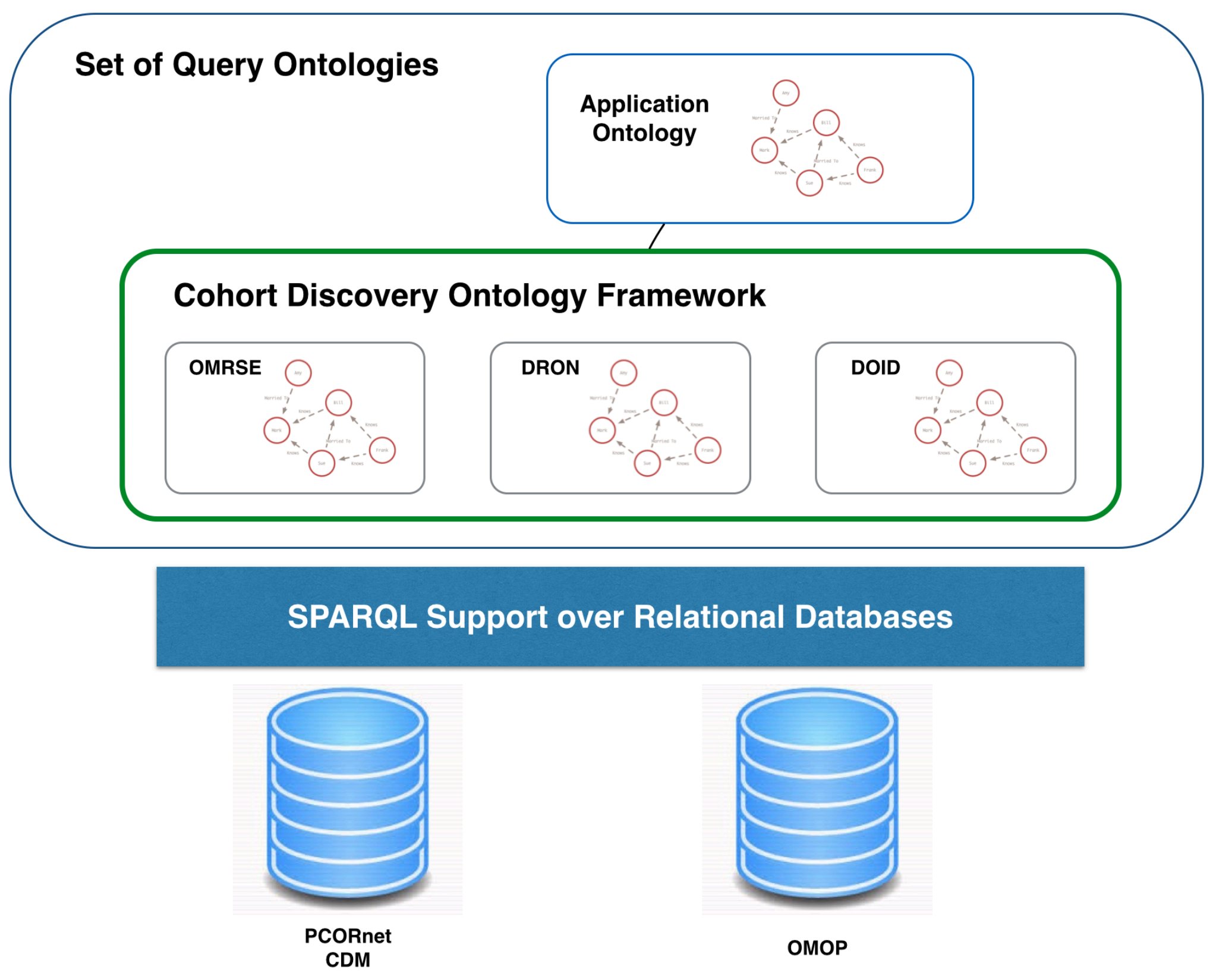

Set of Query Ontologies
Application Ontology
Cohort Discovery Ontology Framework
OMRSE
DRON
DOID
SPARQL Support over Relational Databases
PCORnet CDM
OMOP

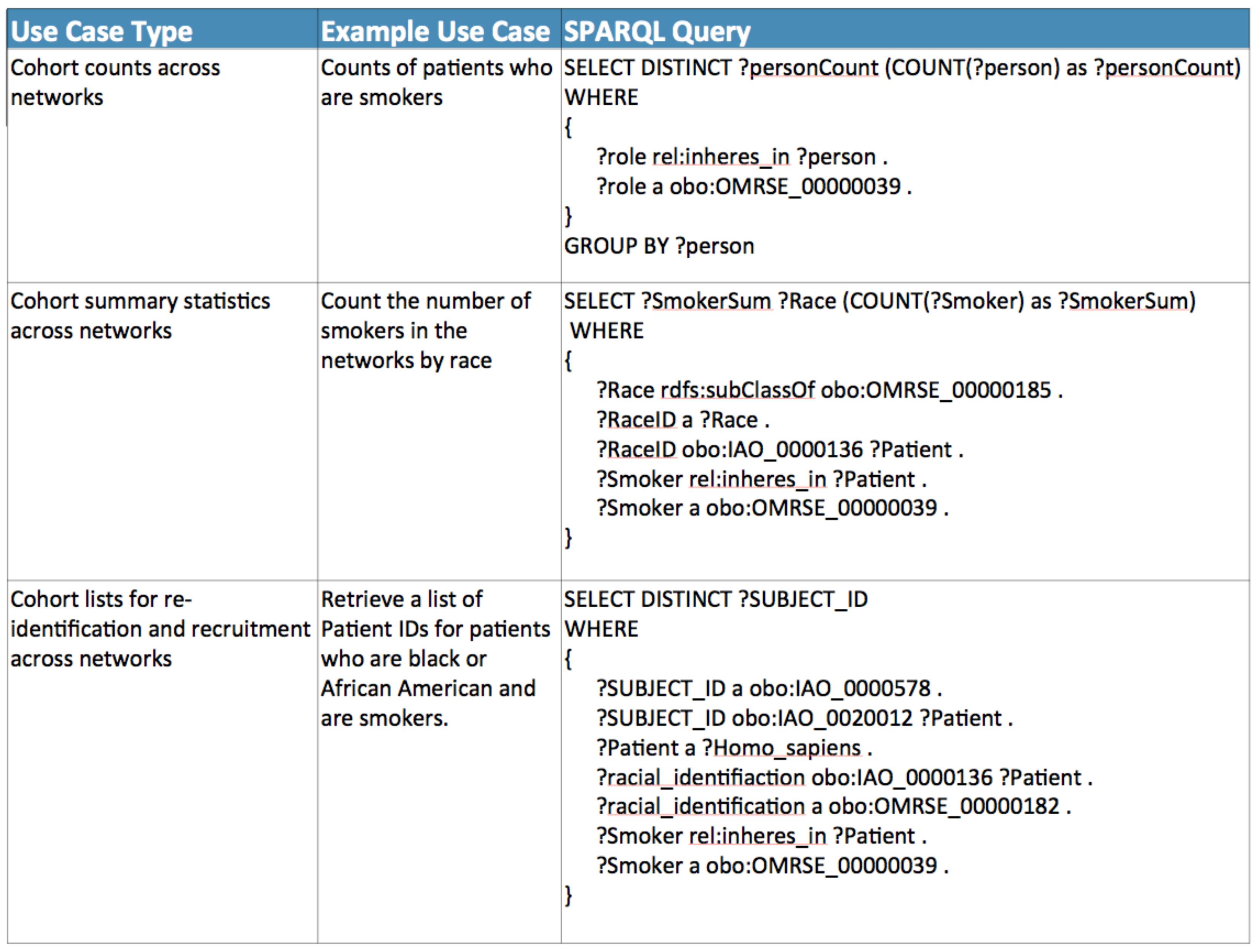
Figure 1. System architecture for cohort discovery across research networks with different CDMs.

Results

We identified three use cases that fit three of the four use case types described in the 2014 Workshop on Data harmonization for Patient-Centered Clinical Research. Our use cases focused on demographic information that we have represented in OMRSE. They are:

- Count unique individuals who are smokers
- Count the number of smokers in the network by race
- Retrieve a list of Patient IDs for patients who are African American and are smokers.

We then wrote SPARQL queries (Figure 2) for each use case and verified them over a small knowledge base.



Use Case Type	Example Use Case	SPARQL Query

Cohort counts across networks	Counts of patients who are smokers	SELECT DISTINCT ?personCount (COUNT(?person) as ?personCount)
		WHERE
		{
		?role rel:inheres_in ?person .
		?role a obo:OMRSE_00000039 .
		}
		GROUP BY ?person

Cohort summary statistics across networks	Count the number of smokers in the networks by race	SELECT ?SmokerSum ?Race (COUNT(?Smoker) as ?SmokerSum)
		WHERE
		{
		?Race rdfs:subClassOf obo:OMRSE_00000185 .
		?RaceID a ?Race .
		?RaceID obo:IAO_0000136 ?Patient .
		?Smoker rel:inheres_in ?Patient .
		?Smoker a obo:OMRSE_00000039 .
		}

Cohort lists for re-identification and recruitment across networks	Retrieve a list of Patient IDs for patients who are black or African American and are smokers.	SELECT DISTINCT ?SUBJECT_ID
		WHERE
		{
		?SUBJECT_ID a obo:IAO_0000578 .
		?SUBJECT_ID obo:IAO_0020012 ?Patient .
		?Patient a ?Homo_sapiens .
		?racial_identifiaction obo:IAO_0000136 ?Patient .
		?racial_identification a obo:OMRSE_00000182 .
		?Smoker rel:inheres_in ?Patient .
		?Smoker a obo:OMRSE_00000039 .
		}

Figure 2. SPARQL Queries for Cohort Discovery; Use case types come from 2014 Institute of Medicine data harmonization for patient-centered clinical research workshop.

Conclusions

Our approach can be extended to other types of patient data using other ontologies that are developed according to OBO best practices. Some examples of other existing ontologies include:

- The Drug Ontology
- Human Disease Ontology
- Vital Sign Ontology

We currently have a working group that focuses on ontologically representing the PCORnet CDM in OBO-compliant ontologies. In addition to this work, future steps for implementing this framework include ensuring that other CDMs are fully represented in the ontologies. Further work also needs to be done on mapping CDMs to a set of query ontologies and developing application ontologies with enriched classes and relations.

Acknowledgement

This work supported in part by the NIH/NCATS Clinical and Translational Science Awards to the University of Florida UL1 TR000064 and the OneFlorida Clinical Research Consortium.

Conflict of Interest:

We have no conflict of interest to declare.