A Collaborative, Realism-Based, Electronic Healthcare Graph: Public Data, Common Data Models, and Practical Instantiation

Mark A. Miller[a], Christian J. Stoeckert Jr.[a,b]

[a]Institute for Biomedical Informatics
Perelman School of Medicine, University of Pennsylvania
Philadelphia, PA, USA,
[b] Dept. of Genetics, Institute for Biomedical Informatics
Perelman School of Medicine, University of Pennsylvania
Philadelphia, PA, USA

Abstract

There is ample literature on the semantic modeling of biomedical data in general, but less has been published on realism-based, semantic instantiation of electronic health records (EHR). Reasons include difficult design choices and issues of data governance. A collaborative approach can address design and technology utilization issues, but is especially constrained by limited access to the data at hand: protected health information.

Effective collaboration can be facilitated with public, EHR-like data sets, which would ideally include a large variety of datatypes mirroring actual EHRs and enough records to drive a performance assessment. An investment into reading public EHR-like data from a popular common data model (CDM) is preferable over reading each public data set's native format.

In addition to identifying suitable public EHR-like data sets and CDMs, this paper addresses instantiation via relational-to-RDF mapping. The completed instantiation is available for download, and a competency question demonstrates fidelity across all discussed formats.

Keywords:

RDF triples, realism, EHR, OMOP

Introduction

This paper describes the reification and semantic instantiation of selected columns from public data sets that are largely representative of a prototypical electronic health record system. Because the original data come from public sources, the output (a downloadable RDF data set) is available for scrutiny by anyone who has an interest in disciplines like healthcare informatics, linked data, or ontological realism (1).

There is evidence that, while even experienced users of sophisticated upper ontologies like the Basic Formal Ontology (BFO) (2) may have some difficulty in tasks such as properly classifying individuals when working in isolation (3), those same ontologists are likely to reach a consensus as a group after reviewing one another's positions. Participation in the weekly web meeting (4) held by the Ontology for Biomedical Investigations (OBI) community confirms the value of this kind of collaborative approach. Even semi-anonymous resources like Stack Overflow can be a source of useful collaboration, given the submission of a well written question, including sample data.

Fundamentally, the collaborative approach acknowledges that no one individual or group is likely to be an authority on the content and structure of an EHR, semantic web technologies in general, and specifically, the Web Ontology Language (OWL), the BFO, and mid-level ontologies from the Open Biomedical and Biological Ontologies Foundry (OBO).

There is no reason to believe that these difficulties or the need for collaboration are unique to a semantic approach. Large-scale initiatives to harmonize, integrate, or transfer health care data with relational technologies, including CDMs to be discussed later, have benefited from precisely the large, diverse input that is recommended for similar semantic initiatives.

Despite a 50-year history, resulting in a plethora of commercial product and support offerings, the relational database world still struggles to cope with complex, heterogeneous data. Therefore, we find this to be an ideal time for innovative application of semantic web approaches to health care data. Specifically, we advocate working collaboratively with synthetic healthcare data stored as RDF triples, and using terms from OWL ontologies that follow the ontological realism method. This can be thought of a chain of progressive and cumulative commitments:

- The use of any graph format will support assertions about (and visualization of) chained or branching relations like temporal precedence, or the inputs and output of processes, without requiring self-joins that are characteristic of relational database solutions.
- Use of the W3C's RDF standard supports a linked data approach, where statements about patients can use terms that are defined in some external, public data set. We believe that the flexibility of property graphs is valuable in a standalone data integration effort, but that the subject-predicate-object structure imposed by RDF is more supportive of broader data interoperability, sharing and linking. Likewise, the use of the SPARQL query language and software libraries like RDF4J serves a protection against vendor lock-in.(5)
- We limit ourselves to using class and property/predicate terms from an ontology, which could theoretically use the RDFS, SKOS or OWL schemas. Among other things, this is an initial step in making the database self-documenting, or free from dependency on an external data dictionary. RDFS and OWL both define subclass and subproperty relations that can be used for reasoning, and OWL provides support for richer axioms (which may or may not be supported by default reasoning levels in RDF triplestores).

- At the highest level of rigor, we limit our upper ontologies to those that adhere to the principles of the OBO foundry and therefore ontological realism

The term "ontological realism" is used here to specifically mean using the BFO as an upper ontology, and more generally following the methodology advocated by Smith and Ceusters in 2010 (1) as best as possible. At a minimum, this means instantiating universal and generalizable classes of things, and resisting the temptation to structure knowledge as topical "concepts". This is a top-down approach that emphasizes computability and consistency between ontology artifacts.

While similar, the terms reification and semantic instantiation are used here to describe two different processes, both of which can be performed in an automated fashion, after some initial configuration by one or more people with domain knowledge and ontological training:

1. Seeing values like "M", "3/11/1969" and "123456" in one row from a data table; inspecting contextual information like the table and column names and a data dictionary; then coming to the following conclusions:
    a. "M" itself means 'male gender identity datum'
    b. While "3/11/1969" itself doesn't mean anything, it is associated with a datum about someone's birth. Likewise, "123456" is associated with some thing that can denote the person.
2. Writing this knowledge, in the form of RDF triples, into a semantic triplestore database. In part, an approximation of this would look like
    a. :X a 'male gender identity datum' .
    b. :Y a 'Homo sapiens' .
    c. :X 'is about' :Y .

At this point in time, there is limited precedence for realism-based semantic instantiations of EHRs. This paper primarily builds upon ideas developed in the PennTURBO project (6,7). Beyond that, one especially relevant paper describes realism-based instantiation of electronic dental records (8), although the patient data has not been made public. Bona, Nolan and Brochhausen have generated realism based RDF triples from the non-image-related clinical data present in the Cancer Imaging Archive (9). Elkin and colleagues have applied natural language approaches to electronic health records, resulting in property- and RDF graphs that use terms from vocabularies such as SNOMED (10,11). Ceusters and colleagues have demonstrated the applicability of their referent tracking approach to electronic health records (12), and they have built a system that inserts referent tracking statements into an RDF triplestore (13). Research at the University of Murcia in Spain has resulted in several papers (14) describing a Semantic Web Integration Tool that can consume data from XML files and relational databases and then apply reasoning via the OWL API (15). At a minimum, they have applied an archetype-based colorectal cancer classifier to a 500 patient subset from a 20,000 patient database (16). While intriguing, it appears that their work doesn't share many of the objectives of this report: their terms were drawn broadly from the NCBO BioPortal, without emphasizing realism; the inputs into their classifier were instances of 'histopathology report', not instances of 'patient' or 'Homo sapiens'; there was little discussion of loading statements into an RDF triplestore.

In addition to qualitatively describing the experience of working with various public, relational, EHR-like data sets, this paper uses a competency question (CQ1) to ensure that the same result is obtained after any data transformation. The question is: how many white male patients, born between 1960 and 1980, have an average systolic blood pressure between 110 and 130?

DE-SynPUF data set

The United States Centers for Medicare & Medicaid Services (CMS) provides a data set entitled "Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF)" (17). Background information provided by the CMS includes the following:

"The DE-SynPUF was created with the goal of providing a realistic set of claims data in the public domain while providing the very highest degree of protection to the Medicare beneficiaries' protected health information."

The purposes of the DE-SynPUF are to:

1. allow data entrepreneurs to develop and create software and applications that may eventually be applied to actual CMS claims data;
2. train researchers on the use and complexity of conducting analyses with CMS claims data prior to initiating the process to obtain access to actual CMS data; and,
3. support safe data mining innovations that may reveal unanticipated knowledge gains while preserving beneficiary privacy.

DE-SynPUF consists of five types of data, for the years 2008, 2009 and 2010:

1. Beneficiary Summary
2. Inpatient Claims
3. Outpatient Claims
4. Carrier Claims
5. Prescription Drug Events

The DE-SynPUF page provides links to documentation, such as the data dictionary. It is noted that the synthetic data generation process may impose some limits on the usefulness of DE-SynPUF for inferential research.

MIMIC-III data set

The Medical Information Mart for Intensive Care III (MIMIC-III) data set (18) is described as "a large, freely-available database comprising de-identified health-related data associated with over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012."

Among other information, MIMIC-III contains

- patient demographics
- vital sign measurements (~1 data point per hour)
- laboratory test results
- procedures
- medications
- diagnosis codes

While the MIMIC-III website describes the data set as "freely-available", access is in fact limited and requires the completion of an application and proving completion of human subjects research training. This is not surprising, given that MIMIC-III consists of de-identified data (19) from actual patients of the Beth

Israel Deaconess Medical Center. Most significantly, MIMIC-III licenses are issued on a per-individual basics, and derived data must be shared via the limited-access PhysioNet website.

Synthea data set

From the Synthea website:

"SyntheaTM is an open-source, synthetic patient generator that models the medical history of synthetic patients. Our mission is to provide high-quality, synthetic, realistic but not real, patient data and associated health records…. The resulting data is free from cost, privacy, and security restrictions, enabling research with Health IT data that is otherwise legally or practically unavailable." (20)

A 1000 patient, pre-built Synthea data set is available for download in multiple formats, including CSV, from the Synthea homepage.

In addition to the pre-built files, users can build their own Synthea data sets by downloading Apache-licensed code from the Synthea GitHub repository (21).

Methods

Loading public EHR-like data sets into a Relational Database

DE-SynPUF

The DE-SynPUF data set is split into 20 collections of independent CSV files. All five types of data from all three years were downloaded for this paper, but only from sample 1 of 20, resulting in a data set with slightly more than 100,000 unique patient identifiers. PostgreSQL schemas for the beneficiary, inpatient claim and drug event CSV files were constructed with the csvsql application from the csvkit (22) package, and then the CSV files were imported into a PostgreSQL database.

MIMIC-III

The complete process of downloading the MIMIC-III data set and populating it into a PostgreSQL 11 database on an Ubuntu 18 server was executed without complications using a Makefile provided by the MIT Laboratory for Computational Physiology (23).

Synthea

For this project, a Synthea data set was created by first setting the 'exporter.csv.export' parameter in

- ./src/main/resources/synthea.properties

to true and then compiling the source code. Then, the data were synthesized with the following command:

- ./run_synthea -s 42 -p 1000 Pennsylvania Philadelphia

Which calls for 1000 patients, representative of people from Philadelphia. An initial seed was set to 42, so that the generation process could be replicated by collaborators. The resulting CSV files were loaded into a PostgreSQL database in tandem with a format conversion that is described below.

Conversion to OMOP common data model

The three previously mentioned EHR-like data sets each have distinct structures, in terms of how different kinds of data are segmented into separate CSV files, how the columns are named, etc. For all three to be directly instantiated into a triplestore, three different sets of SPARQL statements would need to be developed. Instead, we converted all three data sets into a single CDM. This commitment to a CDM allows the use of a single set of SPARQL statements, with minor modifications, to implement all three instantiations.

The Observational Medical Outcomes Partnership schema (OMOP) was chosen based on literature evaluations (24) and utility within the PennMedicine organization. Since the CDM functions as a staging area between tabular data sets and the semantic graph, limitations of OMOP from the perspective of ontological realism (25) and concerns about accurate counting (26) were tolerated.

OMOP conversion tools exist for all three of the public data sets under consideration and were used to load or migrate the data sets into new PostgreSQL schemas in the OMOP format. At the time this paper was written, OMOP schema version 6.0 was under development, but the conversion tools all used a recent 5.x schema.

Full utilization of OMOP includes obtaining vocabularies from their Athena system in order to understand the meaning of its concept codes, like "8507" for "MALE". While a semantic instantiation will still need to map "8507" to a term like 'male gender identity datum' (OMRSE_00000141), the use of a CDM eliminates the need to map three different codes from three different data sets.

One contributor to database population difficulties described below could be minor mismatches between the schema created by the Extract/Transform/Load (ETL) scripts and the most recent OMOP vocabulary downloads.

Instantiation via Relational-to-RDF Mapping

Realism-based statements about the Synthea data, from the previously described OMOP schema, were loaded into a Ontotext GraphDB triplestore with a relational-to-RDF (R2R) mapping approach. Because all three EHR-like datasets evaluated in this paper had been loaded into OMOP schemas, three different, source-specific variable recodings were not required.

The PennTURBO team uses Ontotext GraphDB as its primary triplestore because it has multiple text indexing solutions, a sophisticated web-based SPARQL development environment, and a visual data exploration tool. Unfortunately, its Ontorefine tool can only instantiate data from files, not database connections.

Because we couldn't find any single R2R tool that met all of our needs, the ability to instantiate triples from the contents of an OMOP database was added to PennTURBO's existing "Carnival/Drivetrain" data integration and harmonization software suite. Carnival and Drivetrain have not been completely released into the public domain yet.

In order to demonstrate the general applicability of our approach, we have also performed an instantiation with Stardog's Virtual Graph feature. Following the example of Carnival/Drivetrain, the Synthea data in the OMOP PostgreSQL database were instantiated as shallow "data models with shortcuts", using terms from the TURBO Ontology whose scope is limited to "data space", like the class 'person data model' (TURBO_0010161) and the predicate 'shortcut person data model to DOB (textual)' (TURBO_0010085). Expansion of the shallow triples into statements using OBO foundry terms, including the recoding of categorical variables, was performed by writing federated

SPARQL queries from GraphDB, against the Stardog Virtual Graph.

Insertion of statements about the precedence of a given person's healthcare encounters did not require the typical comparison of encounter dates, as OMOP populates a "preceding_visit_occurrence_id" column in the "visit_occurrence" table as part of the Synthea ETL.

Results

DE-SynPUF

DE-SynPUF's "Beneficiary Summary" and "Prescription Drug Events" files were found to have some useful overlap with the patient demographics and medication-order tables in the University of Pennsylvania's clinical data warehouse. "Beneficiary Summary" also contains some less relevant (or at least off-topic) summary phenotype and claims/ utilization data.

"Inpatient Claims", "Outpatient Claims", and "Carrier Claims" all contain provider identifiers, dates, diagnosis codes, and procedure codes. All of the claims files, especially "Carrier Claims" contain numerous columns for the financial aspects of health insurance claims, which were not examined for this report.

Each of the claims tables use multiple columns per table for diagnosis and procedure codes, since the tables were normalized to one row per claim. This is not especially appealing for input into a semantic instantiation, in which diagnoses and codes are first class citizens, just like claims. "Prescription Drug Events" refers to drugs by NDC codes, which are less desirable than RxNorm codes, due to their higher granularity, or number of codes per product/route/dose. DE-SynPUF provides values about the patients' dates of birth, genders and races, but no clinical findings or measurements like height, weight or blood pressure. See Table 1.

MIMIC-III

Compared to the DE-SynPUF data set, MIMIC-III uses similar vocabularies and contains essentially all of the same clinical datatypes, with the addition of clinical observations and measures, and free-text clinical notes. On the other hand, MIMIC-III includes a much smaller number of patients: 1,152. Finally, while the MIMIC-III data access policy isn't unreasonable for an individual investigator, it does pose a limitation for a multi-investigator, collaborative effort.

Synthea

Data generated with Synthea contain all of the clinical datatypes present in the DE-SynPUF and MIMIC-III data sets, except for the free text notes that are available in MIMIC-III alone. Synthea primarily refers to drugs with RxNorm codes, which is more directly compatible with existing PennTURBO work than the NDCs used in DE-SynPUF and MIMIC-III, yet it refers to disorders with SNOMED codes, which is a minor incompatibility with PennTURBO.

Table 1– Suitability of EHR-like data for collaborative work on semantic healthcare graphs

Data Source	Unrestricted Access?	ICD-X Diagnoses?	RxNorm Medications?	Demographic & Quantitative Clinical Data?	Free Text Notes?
DE-SynPUF	Public domain	ICD-9, with multiple columns per "claim" row.	NDC	Patient demographics are present, but not any clinical measurements.	No
MIMIC	Approval requires data-only human subjects training and a research proposal. MIMIC is only licensed to individuals and derived work can only be distributed through the PhysioNet website.	ICD-9	NDC	Demographics + numerous clinical measurements and findings.	Yes
Synthea	Freely available downloads; or generate with Apache-licensed scripts	SNOMED	RxNorm	Numerous demographic and clinical values. LOINC terms are used for labs, etc. "Hispanic" is considered a race, and ethnicities look more like nationalities (Italian, Portuguese, etc.)	No

The association of SNOMED codes with disorders is also a conflation of concepts, which can be remedied with a realism approach. According to the Ontology for General Medical Science, a disorder is a 'A material entity which is clinically abnormal and part of an extended organism. Disorders are the physical basis of disease.' In contrast, a SNOMED or ICD code is an information entity, not a material entity, although it may have an aboutness relationship with the patient, or some material anatomical entity.

Since Synthea offers scripted generation, records can be created for any number of patients. The previously discussed 1000-patient Synthea dataset "from Philadelphia" was selected over DE-SynPUF and MIMIC-III for the rest of this report

CQ1.S, for the Synthea data in their native format:

```
select count(*) from (
        select
                distinct p.id
        from
                native.patients p
        join native.observations o on
                o.patient = p.id
        where
                race = 'white'
                and gender = 'M'
                and birthdate
                        between '1960-01-01'
                        and '1980-01-01'
                and o.code = '8480-6'
        group by
                p.id
        having
                avg(cast(o.value as decimal(4, 1)))
                        between 110 and 130) as included
```

Result: 45 people

OMOP CDM

The DE-SynPUF data didn't require an ETL per se, as it can be downloaded as CSV files (27) that are ready to be directly imported into a relational database using OMOP schema. The Observational Health Data Sciences and Informatics collaborative (OHDSI), which created the OMOP schema, also provides scripts (28) for doing a complete load of DE-SynPUF, as downloaded in its native CMS format, into the OMOP schema.

When running the MIMIC-III OMOP ETL (29), it appeared that a NOT NULL constraint was violated by at least one value in the vocabulary_reference column from the vocabulary table, which contains metadata about the vocabularies. Therefore, that NOT NULL constrain was removed.

OHDSI hosts a GitHub repository containing code for loading Synthea data from CSV files into a PostgreSQL database that uses the OMOP schema. A useful side effect of this process is loading the same Synthea data, in its own native format, into another PostgreSQL schema. Multiple solutions are provided for this task, in support of multiple operating systems. The Synthea OMOP ETL code is evolving over time, and minor bugs were observed in the wrapper scripts each time the GitHub software repository was fetched. In any case, the repository consistently contained all of the SQL commands necessary to build the tables, load the data and build reasonable indices.

It appeared that the ETL correctly migrated most of the Synthea observations table into the OMOP measurement table, but not the units or values columns. (see https://github.com/OHDSI/ETL-Synthea/issues/19) Sufficient keys were shared between the two tables for the "unit_source_value" and "value_source_value" columns in the measurement table to be repopulated after the fact. Population of the "unit_concept_id" column was largely automated by looking up the freshly-loaded unit source values in a map table created as part of the ETL process. Finally, OMOP's "value_as_number" column was copied from "value_source_value" for each row in which a unit concept had successfully been mapped.

Migrations were generally performed as single-threaded operations on a 64 GB Amazon Web Services server, running PostgreSQL 11 and Ubuntu 18. The MIMIC-III and Synthea ETLs each took over one hour. The RAM allocation was decreased to 16 GB after the ETLs and indexing were completed.

CQ1.O, for the Synthea data in an OMOP schema:

```
select count(*) from
        (
        select
                p.person_id,
                avg(cast(m.value_source_value as
                decimal(4, 1)))
        from
                cdm_synthea10.person p
        join cdm_synthea10.measurement m on
                p.person_id = m.person_id
        where
                race_concept_id = 8527
                and gender_concept_id = 8507
                and birth_datetime between '1960-01-01'
                and '1980-01-01'
                and m.measurement_source_value = '8480-6'
        group by
                p.person_id
        having
                avg(cast(m.value_source_value as
                decimal(4, 1))) between 110
                and 130) as included
```

Result: 45 people

Instantiation

After using Carnival to read from the 1000-patient Synthea OMOP schema into a property graph, Drivetrain can perform

realism-based instantiation, RDFS+ reasoning, and import of additional ontologies and linked data sets in roughly ten minutes. Instantiating a topical subset of the data, like patient demographics alone, can be completed with the Stardog Virtual Graph + GraphDB federation approach in roughly the same amount of time, but the subsequent steps have not been automated outside of Drivetrain and are therefore more time consuming.

The PennTURBO ontology (30) is loaded into its own named graph, as are the Monarch Disease Ontology, the Drug Ontology and the Chemicals of Biological Interest ontology. Several RDF linked data sets are also imported, in order to link clinical codes to labels and other relationships (while remaining wary of their concept orientation): RxNorm, Vaccines Administered (CVX) (31) and SNOMED. The RxNorm file is conveniently available for download from the NCBO BioPortal. Generating the SNOMED and CVX files requires a conversion from the UMLS .nlm format to the .RRF format with MetaMorphoSys, loading that into a MySQL database, and then writing that to RDF with scripts from NCBO. (32,33)

CQ1.R, for the Synthea data in realism-based graph:

```
PREFIX : <http://transformunify.org/ontologies/>
PREFIX efo: <http://www.ebi.ac.uk/efo/>
PREFIX obo: <http://purl.obolibrary.org/obo/>
PREFIX pmbb: <http://www.itmat.upenn.edu/biobank/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
select (count(distinct ?patient) as ?count)
where
{
  {
    select ?patient (avg(xsd:float(?sbpv)) as ?avgsbpv)
    where {
      graph pmbb:expanded {
        ?mgidInst a obo:OMRSE_00000141  ;
                  obo:IAO_0000136 ?patient .
        ?wridInst a obo:OMRSE_00000184  ;
                  obo:IAO_0000136 ?patient .
        ?dob a efo:EFO_0004950 ;
             obo:IAO_0000136 ?sns ;
             obo:IAO_0000004 ?dobValue .
        ?patient a obo:NCBITaxon_9606 ;
                 :TURBO_0000303 ?sns ;
                 obo:RO_0000056 ?encounter ;
                 obo:RO_0000087 ?patientRole .
        ?patientRole a obo:OBI_0000093 ;
                     obo:BFO_0000054 ?encounter .
        ?encounter a obo:OGMS_0000097 .
        ?bpassay a obo:VSO_0000006 ;
                 obo:BFO_0000050 ?encounter ;
                 obo:OBI_0000299 ?sbpdatum1 .
        ?sbpdatum1 a obo:HTN_00000001 ;
                   obo:OBI_0001938 ?svs ;
                   obo:IAO_0000221 ?bpq .
        ?bpq a obo:VSO_0000004 ;
             obo:RO_0000052 ?patient .
        ?svs a :TURBO_0010149 ;
             obo:IAO_0000039 obo:UO_0000272 ;
             obo:OBI_0002135 ?sbpv .
        filter(?dobValue > "1960-01-01"^^xsd:date &&
          ?dobValue < "1980-01-01"^^xsd:date)
      }
    }
    group by ?patient
  }
  filter(?avgsbpv > 110 && ?avgsbpv < 130)
}
```

Result: 45 people

At least a partial understanding of the resulting RDF triples can be inferred from the previous SPARQL query. Additionally, Figure 1 provides a visualization of some of the data items and aboutness relationships, and Figure 2 illustrates denotation and mentioning patterns. All of the RDF triples generated for this paper are available as a compressed n-quads RDF file, which is further described in the discussion section.

A handful of design patterns are proposed in this instantiation and have already been the subject of some collaborative evaluation. We invite further feedback from those who have read this paper and/or loaded a dump of our work into their own triplestore.

Highlighted Patterns:

- What should we aspire towards in terms of succinct, consistent, and semantically clear aboutness patterns? We are currently asserting that racial and gender identity datums are about the patient, but date of birth is about the ‘start of neonate stage’ that the patient participates in. Some clinical measurements, like blood pressure, are asserted to be about a quality inhering in the patient, and supported with the instantiation of a specific assay class and a value specification with units. What then would a ‘body mass index’ datum be about?
- Perhaps we are being overconfident in translating values of “M” from the person.gender_source_value column as instances of class ‘male gender identity datum’, OMRSE_00000141. The ontology of medically relevant social entities defines gender identity datums as being the output of gender identification processes. If the "M" value is based on genotype data or a health care professional’s examination of external genitalia, is the resulting datum really about gender identity? To support this inquiry, male and female biological sex datum classes have been added to the TURBO ontology, along with defined classes for the union of male gender identity datums and male biological sex datums (along with the analogous case for females).

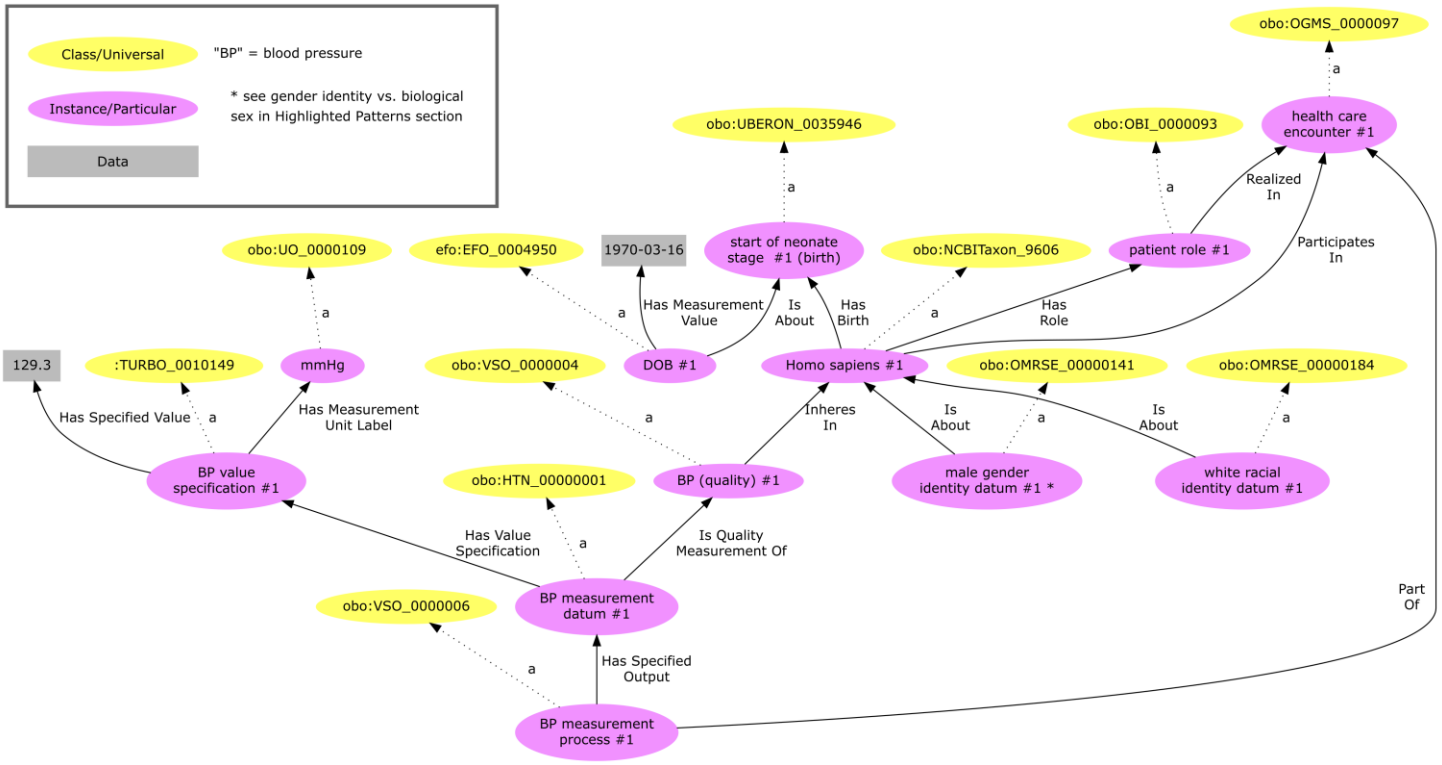

Figure 1–Aboutness Relationships, including 'Is Quality Measuremnt Of', Relevant to Competency Question CQ1.R

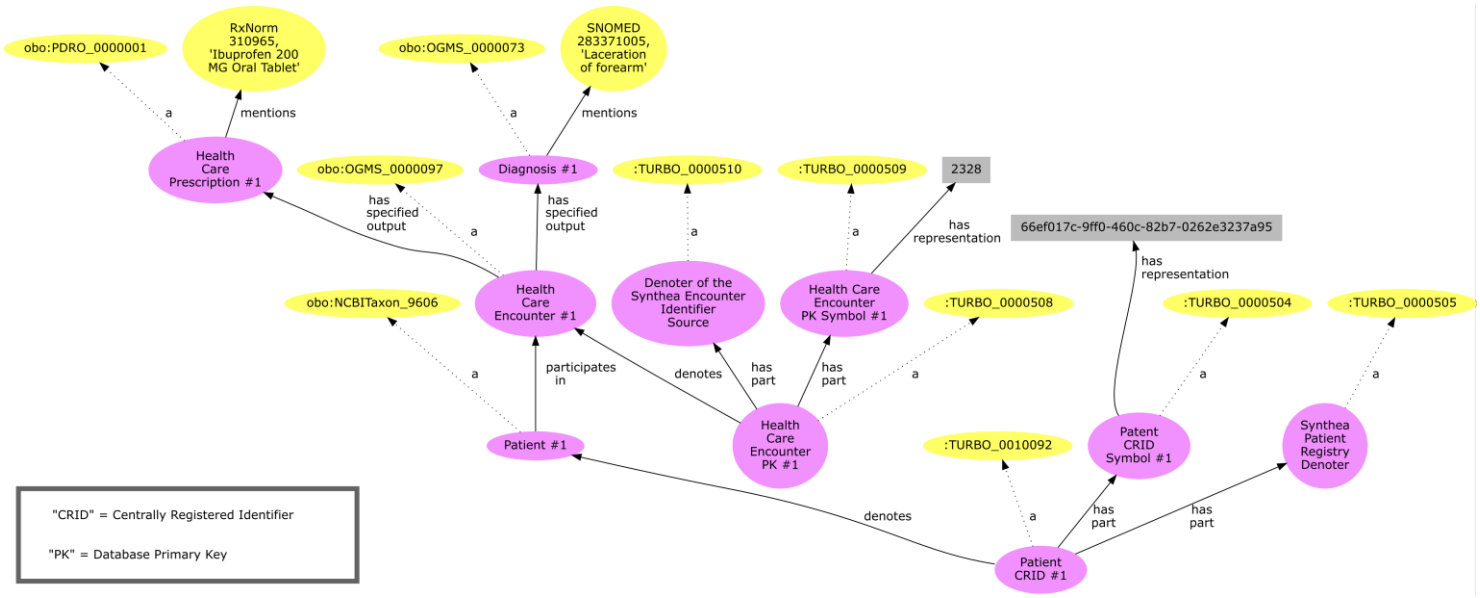

Figure 2–Denotation and Punned Mentioning Relationships

- PennTURBO's Drivetrain application has the capability of making data-driven inferences, even in the case of missing or contradictory data. The conclusion drawing process and its evidence are instantiated explicitly in the graph, and collaborators can specify what rules they want to apply. For example, the presence of three male gender identity (or biological sex?) datums and one female datum might lead to the inference that a female biological sex quality inheres in the patient. Can inferences about the population in which a patient is a member be drawn from racial identity datums? Is this an ontological question or a societal question?

- How can dates of birth be expressed succinctly and in adherence to the ontological realism methodology? Synthea/OMOP "birth_datetime" values have been modeled as instances of 'date of birth' (EFO_0004950), which is placed into the TURBO ontology as a subclass of 'time measurement datum' (IAO_0000416). The 'date of birth' instances take xsd:date literals and are about 'start of neonate stage' (UBERON_0035946) process boundary instances. A 'born on' property (TURBO_0000303) is used to link the patient to the process boundary, and has the following characteristics: domain 'Homo sapiens'; range 'start of neonate stage'; definition "'participates in' o inverse (starts)"". Supplemental class definitions and property chains are being added to the TURBO ontology and the PCORowl ontology
- How should we use information content entities for denoting, given that identifying values from a database might be rigorously maintained by some authority or could just be auto-generated primary keys? We have illustrated these cases with a 'centrally registered identifier' (CRID) for denoting the patient, and a 'database primary key' from the TURBO ontology for the encounter. (This mimics the actual situation we face with our EHR.). Also, the Information Artifact Ontology (IAO) defines a CRID as having some part that denotes some 'centrally registered identifier registry', but IAO does not include any class that is defined as denoting registries. Based on the fact that the domain of 'denotes' is 'information content entity', we have added 'identifier source', 'identifier source denoter' and 'registry denoter' classes, as subclasses of the slightly more specific 'data item' class, into the TURBO ontology. Additionally, what datatype predicate should bind the lexical representation of an identifier to the symbol part of a CRID or primary key? Because IAO does not appear to have a suitable predicate, a 'has representation' predicate has been added to the TURBO ontology and will shortly be added to OBI.
- How can we provide context for clinical codes without violating ontological realism principles or suggesting that the graph contains knowledge that was recklessly interpreted from the codes? Medication codes and "condition" codes are present in the Synthea/OMOP data set. We say that these codes, manifest as URIs for classes, are mentioned by diagnoses and prescriptions, which are in turn the outputs of 'health care encounters'. Since the object of these (instance-level) 'mentions' statements are defined in their source ontologies as classes, this is a case of OWL2 punning. When combined with additional public ontologies and clinical lined data sets (see Instantiation, above), it becomes possible to indirectly answer real collaborator requests like "count the patients with diabetes who were taking statin drugs." However, the graph doesn't truly know what disease dispositions inhere in the patients, or which drugs were actually ingested (etc.), only the codes that were assigned or recorded.

Discussion

Besides this current work, we are not aware of any other realism-based instantiation of a data set that is representative of an EHR and also available for all to see. The previously mentioned instantiation of an EDR comes closest, but is not available for public review as it contains PHI.

The data set that was ultimately instantiated in this paper represents 1000 synthetic patients. While the PennTURBO team has good experience instantiating tens of thousands of patients, more work is required to determine how this method will scale to hundreds of thousands or millions of patients. It's possible that enterprise versions of the triplestore applications may be required, along with hardware and operating system optimization.

There are some differences between the data that are available in Synthea, that can fit in an OMOP schema, and that we are already routinely instantiating into PennTURBO (independent of the work described in this paper.) Synthetic genomic data is not available and has no table in the OMOP schema, but PennTURBO does make statements about predicted loss of function calls when available for Penn Medicine patients. As part of that, we instantiate the specimens that were collected from patients as part of health care encounter, and which went on to serve as the input into a chain of sequencing and bioinformatics processes. There is an OMOP specimens table, but synthesis of specimen data with Synthea might require writing a plugin. Synthea and OMOP support data about health care procedures, and we have future plans to instantiate it, as well as the procedure data in our EHR.

The PennTURBO team routinely runs its instantiations through RDFS+ reasoning, but neither PennTURBO nor this Synthea/OMOP work has been run through higher levels of reasoning, like OWL-Horst.

Conclusions

We believe that graph models of health care data will enable faster question answering and cohort building, compared to what can be done in existing relational EHRs or clinical data warehouses. Because we wish to develop this approach collaboratively, in a way that fosters interoperability and peer review, we have constructed an RDF graph model of synthetic health care data, synthea_graph_exportable.nq and have shared it at http://doi.org/10.5281/zenodo.2641233

One of our requirements for this project was identifying a source of sharable healthcare data whose contents are as similar as possible to the clinical data warehouse that provides the majority of our information. DE-SynPUF and MIMIC-III were considered, but Synthea was chosen as having both the most relevant data and a suitable redistribution policy. Specifically, Synthea allows the generation of any number of observations and includes numerical and qualitative clinical findings. MIMIC-III would be a good choice in a setting where the value of free text clinical notes outweighed the inconvenience of the more restrictive license.

The three data sources were staged in the OMOP common data model in order to minimize the effort required to become familiar with each source's structure, and also because we anticipate using the OMOP model for both ingesting data sources complementary to our clinical warehouse, and as a format for sharing portions of the clinical warehouse with other medical research institutions. Scripts for transforming each of the three "sharable" data sources into an OMOP model are available at OHDSI's GitHub software

repository. While these scripts dramatically decrease the effort required to perform the transformations, users should be prepared to do a small amount of debugging.

We have briefly demonstrated the ability to migrate the Synthea data from an OMOP-formatted PostgreSQL relational database with two methods: our internal "Carnival/Drivetrain" software suite, and the Virtual Graph feature from the Stardog triplestore, in federation with the GraphDB triplestore (which also serves as the final destination.)

A competency question, representative of a cohort-building query, was applied to Synthea data in its native format, the same data in an OMOP schema, and corresponding RDF triples. The same answer was obtained in all three cases.

We encourage readers to download our Synthea triples from the address above and load them into any RDF triplestore. The TURBO ontology is included, but not the supporting RxNorm, CVX and SNOMED clinical knowledgebases. An RDF representation of RxNorm can be obtained from the NCBO BioPortal, but obtaining RDF models of CVX and SNOMED requires performing a multi-step conversion from the Unified Medical Language System.(32,33)

We are especially interested in hearing feedback about our 'is about' relations, the way we 'mention' diagnosis and medication codes via OWL2 punning, and the way we denote entities with either centrally registered identifiers or database primary keys, depending on our confidence that the identifier is truly centrally registered. Several of these issues are already the subjects of active GitHub issues such as https://github.com/obi-ontology/obi/issues/985.

Acknowledgements

We benefitted from valuable collaborations with Amanda Hicks, our PennTURBO colleagues (David Birtwell, Hayden Freedman, Heather Williams), ontologists from the Eukaryotic Pathogen database (Jie Zheng and John Judkins), and numerous members of the OBI development community.

This work was done as part of the PennTURBO project, which is supported by the Institute for Biomedical Informatics and by the Institute for Translational Medicine and Therapeutics at the University of Pennsylvania.

Address for correspondence

Mark A. Miller, markampa@pennmedicine.upenn.edu

References

1. Smith B, Ceusters W. Ontological realism: A methodology for coordinated evolution of scientific ontologies. Appl Ontol. 2010 Nov 15;5(3–4):139–88.

2. Arp R, Smith B, Spear AD. Building Ontologies with Basic Formal Ontology. The MIT Press; 2015.

3. Stevens R, Lord P, Malone J, Matentzoglu N. Measuring expert performance at manually classifying domain entities under upper ontology classes. J Web Semant. 2018 Sep 6;

4. OBI weekly conference call [Internet]. OBI Home. Available from: http://obi-ontology.org/#contact-us

5. Alocci D, Mariethoz J, Horlacher O, Bolleman JT, Campbell MP, Lisacek F. Property Graph vs RDF Triple Store: A Comparison on Glycan Substructure Search. PLoS ONE [Internet]. 2015 Dec 14 [cited 2019 Jul 12];10(12). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4684231/

6. Stoeckert C, Birtwell D, Freedman H, Miller M, Williams H. ICBO_2018_12: Transforming and Unifying Research with Biomedical Ontologies: The Penn TURBO project. 2018.

7. TURBO|PennTURBO Documentation [Internet]. TURBO|PennTURBO Documentation. Available from: https://pennturbo.github.io/Turbo-Documentation/

8. Schleyer TK, Ruttenberg A, Duncan W, Haendel M, Torniai C, Acharya A, et al. An ontology-based method for secondary use of electronic dental record data. AMIA Summits Transl Sci Proc. 2013 18;2013:234–8.

9. Bona JP, Nolan TS, Brochhausen M. Ontology-Enhanced Representations of Non-image Data in The Cancer Imaging Archive. 2018;6.

10. Schlegel DR, Crowner C, Lehoullier F, Elkin PL. HTP-NLP: A New NLP System for High Throughput Phenotyping. Stud Health Technol Inform. 2017;235:276–80.

11. Schlegel DR, Bona JP, Elkin PL. Comparing Small Graph Retrieval Performance for Ontology Concepts in Medical Texts. In: Wang F, Luo G, Weng C, Khan A, Mitra P, Yu C, editors. Biomedical Data Management and Graph Online Querying. Springer International Publishing; 2016. p. 32–44. (Lecture Notes in Computer Science).

12. Ceusters W, Hsu CY, Smith B. Clinical Data Wrangling using Ontological Realism and Referent Tracking. :6.

13. Manzorr S, Ceusters W, Rudnicki R. Implementation of a Referent Tracking System: Int J Healthc Inf Syst Inform. 2007 Oct;2(4):41–58.

14. Semantic Web Integration Tool (SWIT) [Internet]. [cited 2019 Mar 30]. Available from: http://sele.inf.um.es/swit/publications.html

15. OWL API by owlcs [Internet]. [cited 2019 Mar 30]. Available from: http://owlcs.github.io/owlapi/

16. Fernández-Breis JT, Maldonado JA, Marcos M, Legaz-García M del C, Moner D, Torres-Sospedra J, et al. Leveraging electronic healthcare record standards and semantic web technologies for the identification of patient cohorts. J Am Med Inform Assoc JAMIA. 2013 Dec;20(e2):e288-296.

17. CMS 2008-2010 Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF) [Internet]. 2014 [cited 2019 Mar 26]. Available from: https://www.cms.gov/research-statistics-

data-and-systems/downloadable-public-use-files/syn-pufs/de_syn_puf.html

18. MIMIC [Internet]. [cited 2019 Mar 26]. Available from: https://mimic.physionet.org/about/mimic/

19. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016 May 24;3:6.

20. Synthea by the Standard Health Record Collaborative [Internet]. [cited 2019 Mar 26]. Available from: https://synthetichealth.github.io/synthea/#about-landing

21. Synthetic Patient Population Simulator. Contribute to synthetichealth/synthea development by creating an account on GitHub [Internet]. synthetichealth; 2019 [cited 2019 Mar 26]. Available from: https://github.com/synthetichealth/synthea

22. A suite of utilities for converting to and working with CSV, the king of tabular file formats.: wireservice/csvkit [Internet]. wireservice; 2019 [cited 2019 Mar 26]. Available from: https://github.com/wireservice/csvkit

23. MIMIC Code Repository: Code shared by the research community for the MIMIC-III database: MIT-LCP/mimic-code [Internet]. MIT Laboratory for Computational Physiology; 2019 [cited 2019 Mar 26]. Available from: https://github.com/MIT-LCP/mimic-code

24. Garza M, Del Fiol G, Tenenbaum J, Walden A, Zozus MN. Evaluating common data models for use with a longitudinal community registry. J Biomed Inform. 2016;64:333–41.

25. Blaisure JC, Ceusters WM. Improving the 'Fitness for Purpose' of Common Data Models through Realism Based Ontology. AMIA Annu Symp Proc. 2018 Apr 16;2017:440–7.

26. Ceusters W, Blaisure J. A Realism-Based View on Counts in OMOP's Common Data Model. Stud Health Technol Inform. 2017;237:55–62.

27. LTS Computing Downloads [Internet]. LTS Computing Downloads. Available from: http://www.ltscomputing-llc.com/downloads/

28. Workproducts to ETL CMS datasets into OMOP Common Data Model: OHDSI/ETL-CMS [Internet]. Observational Health Data Sciences and Informatics; 2019 [cited 2019 Mar 26]. Available from: https://github.com/OHDSI/ETL-CMS

29. Mapping the MIMIC-III database to the OMOP schema. Contribute to MIT-LCP/mimic-omop development by creating an account on GitHub [Internet]. MIT Laboratory for Computational Physiology; 2019 [cited 2019 Mar 27]. Available from: https://github.com/MIT-LCP/mimic-omop

30. The TURBO ontology [Internet]. The TURBO ontology. Available from: https://raw.githubusercontent.com/PennTURBO/Turbo-Ontology/master/ontologies/turbo_merged.owl

31. IIS | Code Sets | CVX | Vaccines | CDC [Internet]. 2018 [cited 2019 Jul 15]. Available from: https://wcms-wp-test-br.cdc.gov/php-app-template/index.php

32. These python scripts connect to the Unified Medical Language System (UMLS) database and translate the ontologies into RDF/OWL files. This is part of the BioPortal project.: ncbo/umls2rdf [Internet]. National Center for Biomedical Ontology; 2019 [cited 2019 Jul 15]. Available from: https://github.com/ncbo/umls2rdf

33. UMLS - Rich Release Format MySQL Load Script [Internet]. [cited 2019 Jul 15]. Available from: https://www.nlm.nih.gov/research/umls/implementation_resources/scripts/README_RRF_MySQL_Output_Stream.html