Making Study Populations Visible Through Knowledge Graphs

Shruthi Chari[1(✉)], Miao Qi[1], Nkechinyere N. Agu[1],
Oshani Seneviratne[1], Jamie P. McCusker[1], Kristin P. Bennett[1],
Amar K. Das[2], and Deborah L. McGuinness[1]

[1] Rensselaer Polytechnic Institute, Troy, NY 12180, USA
{charis,qim,agun,senevo,mccusj2,bennek}@rpi.edu, dlm@cs.rpi.edu
[2] IBM Research, Cambridge, MA, USA
amardas@us.ibm.com

Abstract. Treatment recommendations within Clinical Practice Guidelines (CPGs) are largely based on findings from clinical trials and case studies, referred to here as research studies, that are often based on highly selective clinical populations, referred to here as study cohorts. When medical practitioners apply CPG recommendations, they need to understand how well their patient population matches the characteristics of those in the study cohort, and thus are confronted with the challenges of locating the study cohort information and making an analytic comparison. To address these challenges, we develop an ontology-enabled prototype system, which exposes the population descriptions in research studies in a declarative manner, with the ultimate goal of allowing medical practitioners to better understand the applicability and generalizability of treatment recommendations. We build a Study Cohort Ontology (SCO) to encode the vocabulary of study population descriptions, that are often reported in the first table in the published work, thus they are often referred to as Table 1. We leverage the well-used Semanticscience Integrated Ontology (SIO) for defining property associations between classes. Further, we model the key components of Table 1s, i.e., collections of study subjects, subject characteristics, and statistical measures in RDF knowledge graphs. We design scenarios for medical practitioners to perform population analysis, and generate cohort similarity visualizations to determine the applicability of a study population to the clinical population of interest. Our semantic approach to make study populations visible, by standardized representations of Table 1s, allows users to quickly derive clinically relevant inferences about study populations.

Resource Website: https://tetherless-world.github.io/study-cohort-ontology/.

Keywords: Scientific Study Data Analysis · Knowledge graphs · Modeling Aggregations and Summary Statistics · Ontology Development

C. Ghidini et al. (Eds.): ISWC 2019, LNCS 11779, pp. 53–68, 2019.
https://doi.org/10.1007/978-3-030-30796-7_4

1 Introduction

Our goal is to build a semantic solution to model the descriptions of study populations and to assist medical practitioners in determining the applicability of a study to their clinical population. Through Fig. 1, we describe the components of a prototype system, that utilizes knowledge representation (KR) techniques to model tabular representations of study population descriptions, often captured in the first table of the scientific publication. We build a Study Cohort Ontology (SCO) (Sect. 4) to support the vocabulary in these Table 1s (plural form) and to model their structure. Further, we encode Table 1s as Resource Description Framework (RDF) knowledge graphs (KGs) [3] (Sect. 5) to expose in a declarative manner[1] these study populations. We demonstrate our ontology and the use of our knowledge graphs with two applications (Sect. 6): one aimed at helping medical practitioners determine the similarity of a patient or a clinical population to the study population, and another aimed at supporting retrospective analysis of a study to expose possible biases or population gaps, such as racial underrepresentations.

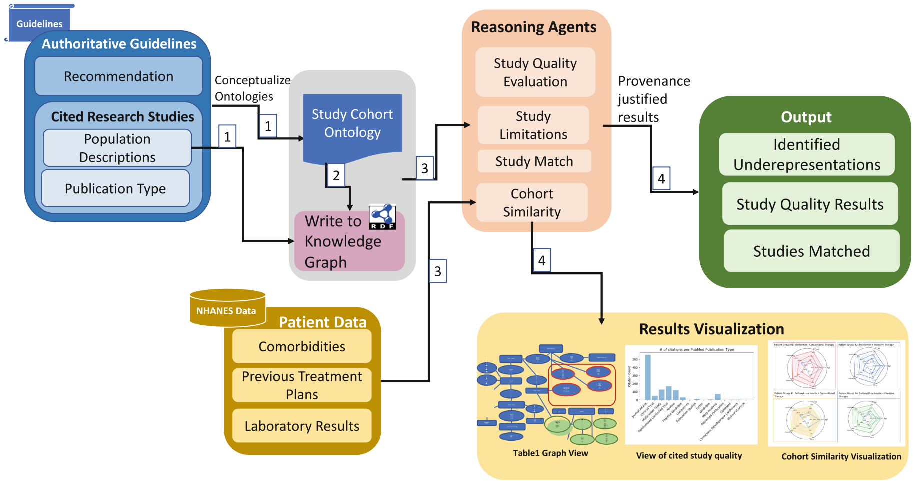


Fig. 1. An overview of the cohort analytics workflow which (1) ingests terms from population descriptions of research studies, (2) standardizes their representations via KR techniques and (3) supports study applicability applications. The numbering is in-line with the figure and is indicative of data flow.

1.1 Use Case

Evidence-based Medicine (EBM) has been gaining popularity, and medical practitioners are using it more often. However, it is challenging to design the CPGs

[1] Declarative manner: in a clear, unambiguous, and computer understandable manner.

to stay current with the growing body of clinical literature. Additionally, medical literature is continuously being revised, e.g., typically, new versions of CPGs are released annually. Treatment recommendations in CPGs are often supported by evidence from cited research studies, i.e. clinical trials and observational case studies, targeting highly selective populations with sociodemographic and comorbid characteristics. In clinical practice, it is well-known that there are biases in clinical evidence that reduce their generalizability. The widely-cited research article, "Trustworthy Clinical Practice Guidelines: Challenges and Potential," [8] states some of the problems in existing guideline practices, such as "Failure to include major population subgroups in the evidence base thwarts our ability to develop clinically relevant, valid guidelines."

Furthermore, when medical practitioners are faced with the treatment of complicated patients who do not wholly align with guideline recommendations, they may want to consult research studies with relevant findings to determine if the study applies to their clinical population. Hence, we are developing a semantic solution to address these challenges, by providing medical practitioners access to high-quality and applicable guideline evidence. We evaluate our solution on the American Diabetes Association's (ADA) Standards of Medical Care 2018 CPG[2] cited research studies, which we will introduce in Sect. 3.

2 Related Work

Existing ontologies for study design and clinical trials are more focused on the study design and methodology aspects of clinical trials, and their vocabulary is insufficient to support cohort descriptions. ProvCaRe [22], an "Ontology for provenance + healthcare research," was developed to assess the scientific rigor and reproducibility of scientific literature. Based on the NIH "Rigor and Reproducibility" guidelines [13], this ontology identifies three components of a study contributing to provenance: study methods (study protocol followed), study instruments (equipment and software used in the study), and study data (metadata about data collection). However, within the ProvCaRe ontology, support for study data is limited to that of inclusion and exclusion criteria, and there is no support for Table 1 terminology, such as subject characteristics and study arms. The Ontology of Clinical Research (OCRE) [20], a widely cited study design ontology used to model the study lifecycle, addresses goals similar to our study applicability scenario. They adopt an Eligibility Rule Grammar and Ontology (ERGO) [21] annotation approach for modeling study eligibility criteria to enable matching a study's phenotype against patient data.

Since we encode a provenance component of guideline evidence, we searched for ontologies for scientific publications. We found that most clinical trial ontologies, e.g., CTO-NDD [24], are domain specific and not directly reusable for a population modeling scenario. Other ontologies, such as the EPOCH suite of clinical trial ontologies [19], that was developed to track patients through their

[2] ADA 2018 CPG at: https://diabetesed.net/wp-content/uploads/2017/12/2018-ADA-Standards-of-Care.pdf.

clinical trial visits, had class hierarchies that were insufficient to represent the types of publications cited in the ADA Standards of Care CPG. Additionally, there is another cohort ontology [11] being developed. However, our modeling of the association of descriptive statistics with subject characteristics differs from their modeling decision to define new properties to represent these associations. Instead, we introduce classes to accommodate new subject characteristic terms upon Table 1 ingestion, and we limit the number of descriptive statistics to a standard set of central tendency measures and boundary values. Hence, we do not leverage their ontology. Further, their ontology is domain specific, including many sleep disorder classes. In SCO we provide a generalized and richer, domain-agnostic Table 1 vocabulary (sufficient to support research studies targeting various diseases).

Clinical trial matching has been attempted multiple times, largely as a Natural Language Processing problem, including a KR approach that improves the quality of the cohort selection process for clinical trials [17]. Clinical trial matching work [17] was carried out with the help of an ontology, and TBOX (knowledge-based) assertions were created from SNOMED-CT for supporting ABOX (real-world) assertions of patient records. However, the focus of their effort was mainly on efficient KR of patient data, and study eligibility criteria was formulated as SPARQL queries on the patient schema. We tackle the converse problem of identifying studies that are applicable to a clinical population based on the study populations reported. We address this problem from the perspective of modeling the study populations.

3 Dataset

Our evaluation dataset is comprised of research studies, cited in the ADA Standards of Medical Care 2018 CPG. We manually reviewed the entire guideline to understand the types of evidence utilized to support treatment recommendations. ADA treatment recommendations are supported through citations within the discussion, which serve as implicit evidence for the recommendation. Further, we used PubMed APIs[3] on the Medline[4] publications, cited in evidence sentences across chapters of the ADA CPG, to retain only those publications that met the qualifications for our definition of research studies. We only considered publications tagged with Pubmed Publication types[5] of: Randomized Controlled Trial, Clinical Trial, and Multicenter Study.

We focused on the pharmaceutical treatments and comorbidities associated with type-2 diabetes, and we filtered our evaluation dataset to contain cited research studies from the Pharmacologic Interventions (Chapter 8) [1] and the Cardiovascular Complications (Chapter 9) [2] of the ADA 2018 CPG. We did a thorough, manual investigation of research studies from these chapters, looking

[3] https://pypi.org/project/pubmed-lookup/.

[4] https://www.nlm.nih.gov/bsd/medline.html.

[5] Find the list of all supported publication types at https://www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T.publication_types/.

for any variance in Table 1s and identifying important study data that explained Table 1 variables. Furthermore, although we were able to gather full-text links for Medline citations through programmatic means, we had to manually follow these links to ensure they are freely available, and, if not, we checked for the availability of the study in other sources. Due to these challenges, we narrowed down the number of research studies to 20 that we list on our resources website.

4 Study Cohort Ontology

As introduced in Sect. 1, we build a Study Cohort Ontology (SCO) to serve as a vocabulary to model the components of a Table 1, the study arms (columns) and their characteristics (rows). We also ensure that the implicit associations exhibited between these components are reflected in SCO. We adopt a bottom-up approach to modeling, that follows, as a by-product of our investigative efforts, the description in Sect. 3. Further, we have attempted to keep our main SCO ontology as domain-agnostic as possible to ensure easy reuse and longevity. In Subsect. 4.1, we introduce the main concepts in our ontology to provide a contextual understanding of the descriptions of populations reported in Table 1s, and walk through our approach to ontology reuse in Subsect. 4.2.

4.1 Primary Classes and Property Associations

The descriptions of study populations that are reported in Table 1s follow a pattern in which columns represent study arms, a group of study subjects who receive an intervention or control regime. The subject characteristics are presented in rows, and are aggregated upon and reported via descriptive statistical measures in the cells of the table. In a conceptual model of SCO as shown in Fig. 2, we depict our modeling of these Table 1 components and the additional details that are necessary to describe a study population in the context of a research study. A more detailed version can be found on our resources website.

As will become evident from a representative Table 1 example shown in Fig. 3, the row and column headers in Table 1s contain specific medical codes and variables that can further be grouped into broad general classifications: Anthropometric Properties (chear:Anthropometry),[6] Demographics (chear:Demographic), Laboratory Results (ncit:C36292), Diseases (doid:0004), and Medical Interventions (provcare:Intervention). Further, we associate all these broad, general classifications we just identified, such as subject characteristics, diseases, interventions etc., via sio:hasAttribute and sio:hasProperty relations to the study subject. More specifically, for properties such as disease and interventions that per-

[6] We use the ontology prefixes: (1) sio: SemanticScience Integrated Ontology (2) uo: The Units of Measurement Ontology (3) chear: Children's Health Exposure Analysis Resource Ontology (4) ncit: National Cancer Institute Thesaurus (5) provcare: ProveCaRe (6) doid: Human Disease Ontology (7) sco: Study Cohort Ontology (8) hasco: Human-Aware Science Ontology (9) prov: The PROV ontology (10) dct: Dublin Core Terms (11) vann: A vocabulary for annotating vocabulary descriptions.

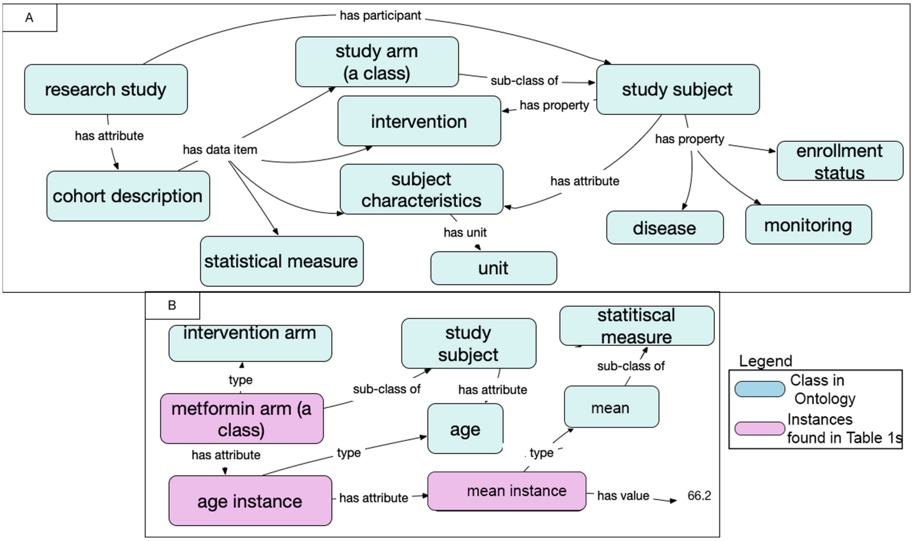


Fig. 2. (A) A high-level overview of SCO that captures the vocabulary and associations needed to model the descriptions of study populations. (B) We depict associations that cannot be realized without actual instantiation of Table 1 data.

sist over time and are characterized by the state the study subject exhibits,[7] we use a sub-property of sio:hasAttribute, i.e. sio:hasProperty, to link them to the study subject. Additionally, we do not maintain certain property associations (e.g. compositional relation between the study arm and study subject) in our ontology and only create them upon the representation of actual Table 1 content in RDF KGs. For the ease of understanding, we depict instances in pink in Fig. 2 to help visualize the realism in our modeling.

To summarize, essentially through SCO, we build a framework to model a set of study subjects, who participate (sio:isParticipantIn) in a research study and belong to a study arm and whose subject characteristics are measured (sio:hasUnit) in units, and are aggregated upon via descriptive statistics. Since we are dealing with the biomedical domain, where multiple definitions may exist for a term, through blank nodes and reification techniques we allow support for this and we maintain provenance for our definitions via prov:wasAttributedTo (person) and dct:source (online source). For example; hasco:ResearchStudy sio:hasAttribute [a skos:definition; sio:hasValue 'A scientific investigation that involves testing a hypothesis'; prov:wasAttributedTo AmarDas]. Additionally, we also provide example usages of our terms via vann:example, to help future users/contributors of our ontology get an idea of the intended usage of the class. Our main SCO ontology, and our accompanying suite of ontologies, Lab Results,

[7] View the definition of sio:hasProperty and sio:hasAttribute relations at: https://raw.githubusercontent.com/micheldumontier/semanticscience/master/ontology/sio/release/sio-subset-labels.owl.

Diseases, Drugs, and Therapies, in which we maintain diabetes specific content, are available as resources. Further, we tested our ontology with the Hermit reasoner.

4.2 Ontology Reuse

We reuse classes and properties from existing biomedical ontologies as much as possible, and only define them ourselves when they do not exist. We primarily reused ontologies available from Bioportal [16] that are regularly maintained and have significant reuse. We have tried to reuse terms from a small set of applicable ontologies to avoid enlarging the ontology when we bring in new classes and additional axioms. We categorize the ontologies, from which we reuse terms, broadly into Study Design ontologies (ProvCaRe, HASCO), Mid-Level ontologies (SIO), Medical ontologies (NCIT, CHEAR, etc.), and Statistical ontologies (STATO, UO). We present a list of our reused ontologies against their groupings on our resources website.

In our approach to ontology reuse, we include minimum information to reference a term (MIREOT) [5] for most of our reused ontologies, such as ProvCare and NCIT, unless we leverage their structure completely. However, we do import a light-weight version of the Child Health Exposure Analysis and Resource (CHEAR) ontology, by applying the MIREOT technique to extract the demographics and anthropometric branches alone. We prefer to import the CHEAR ontology, as it builds off SIO and additionally imports the HAScO human aware science ontology, that we leverage. We utilized an online tool, Ontofox [23], to apply the MIREOT technique to a few ontologies that were supported on this platform. However, for ontologies that were not available on Ontofox, we designed our own SPARQL query to gather subclass and superclass trees for a given ontology class. On our resources website, we make our MIREOT Python script available. This runs the SPARQL query against a Blazegraph endpoint and returns the RDF version of the subset class tree.

5 Knowledge Graph Modeling

We use an annotated example of a Table 1, seen in Fig. 3, to explain our approach of modeling the collections of study subjects, subject characteristics defined on collections, and the descriptive statistics used to summarize these characteristics. We present an RDF snippet in Listing 1.1, and explain smaller sub-portions of our modeling in each subsequent subsection. These snippets form the fundamental pieces of our Table 1 KG. On our resources website, we release the KG representations of the studies in our evaluation dataset, for interested readers to run their analyses.

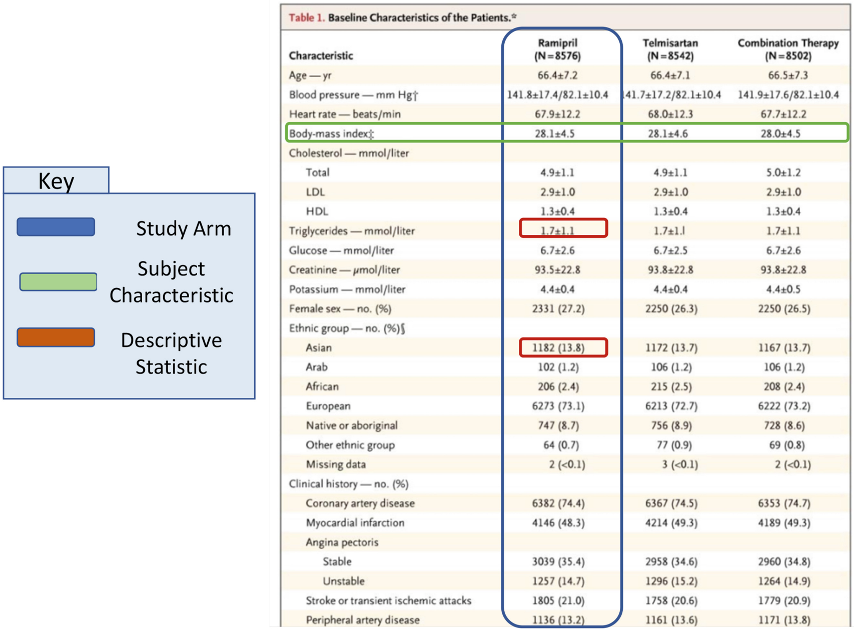

Key
🟦 Study Arm
🟩 Subject Characteristic
🟧 Descriptive Statistic

Table 1. Baseline Characteristics of the Patients.*

Characteristic	Ramipril (N=8576)	Telmisartan (N=8542)	Combination Therapy (N=8502)
Age — yr	66.4±7.2	66.4±7.1	66.5±7.3
Blood pressure — mm Hg†	141.8±17.4/82.1±10.4	141.7±17.2/82.1±10.4	141.9±17.6/82.1±10.4
Heart rate — beats/min	67.9±12.2	68.0±12.3	67.7±12.2
Body-mass index‡	28.1±4.5	28.1±4.6	28.0±4.5
Cholesterol — mmol/liter			
Total	4.9±1.1	4.9±1.1	5.0±1.2
LDL	2.9±1.0	2.9±1.0	2.9±1.0
HDL	1.3±0.4	1.3±0.4	1.3±0.4
Triglycerides — mmol/liter	1.7±1.1	1.7±1.l	1.7±1.1
Glucose — mmol/liter	6.7±2.6	6.7±2.5	6.7±2.6
Creatinine — μmol/liter	93.5±22.8	93.8±22.8	93.8±22.8
Potassium — mmol/liter	4.4±0.4	4.4±0.4	4.4±0.5
Female sex — no. (%)	2331 (27.2)	2250 (26.3)	2250 (26.5)
Ethnic group — no. (%)§			
Asian	1182 (13.8)	1172 (13.7)	1167 (13.7)
Arab	102 (1.2)	106 (1.2)	106 (1.2)
African	206 (2.4)	215 (2.5)	208 (2.4)
European	6273 (73.1)	6213 (72.7)	6222 (73.2)
Native or aboriginal	747 (8.7)	756 (8.9)	728 (8.6)
Other ethnic group	64 (0.7)	77 (0.9)	69 (0.8)
Missing data	2 (<0.1)	3 (<0.1)	2 (<0.1)
Clinical history — no. (%)			
Coronary artery disease	6382 (74.4)	6367 (74.5)	6353 (74.7)
Myocardial infarction	4146 (48.3)	4214 (49.3)	4189 (49.3)
Angina pectoris			
Stable	3039 (35.4)	2958 (34.6)	2960 (34.8)
Unstable	1257 (14.7)	1296 (15.2)	1264 (14.9)
Stroke or transient ischemic attacks	1805 (21.0)	1758 (20.6)	1779 (20.9)
Peripheral artery disease	1136 (13.2)	1161 (13.6)	1171 (13.8)

Fig. 3. An annotated example of Table 1 from a clinical trial "Telmisartan, ramipril, or both in patients at high risk for vascular events" [10] cited in the Cardiovascular Complications (Chapter 9) of the ADA CPG.

Listing 1.1. Representation of a portion of the Ramipril Study Arm

```
sco-i:RamiprilArm
       a    owl:Class, sco:InterventionArm;
       rdfs:subClassOf sio:StudySubject;
       sio:isParticipantIn sco-i:TelmisartanRamiprilStudy;
       sio:hasAttribute
       [ a sco:PopulationSize; sio:hasValue 8576],
       [ a sio:Age; sio:hasUnit sio:Year;
          sio:hasAttribute
          [ a sio:Mean; sio:hasValue 66.4],
          [a sio:StandardDeviation; sio:hasValue 7.2 ]
        ] .
```

5.1 Modeling of Collections of Study Subjects

Study arms are specific subpopulations of study cohorts comprised of a subset of enrolled study subjects. Hence, they are a natural fit for modeling as classes in the OWL web ontology language [4], "Classes provide an abstraction mechanism for grouping resources with similar characteristics. Like RDF classes, every OWL class is associated with a set of individuals, called the class extension.", and model collections as classes.

As discussed earlier in Subsect. 4.1, study arms are represented as columns in Table 1s. Further, the RDF snippet in Listing 1.1 shows a semantic definition of a particular study arm as an instance of the sco:InterventionArm. Study arm definitions are either those of *InterventionArm* or *ControlArm* and they are gathered from the Table 1 columns themselves, if sufficient, if not we consult the study data to find relevant content that describes the arms.

In some Table 1s, there also exist subsets of study arms, created by the presence of categorical row variables (e.g. percentage of Asians), expressed in percentages[8]. Such subsets are expressed as rdfs:subClassOf the main study arm, and have an owl:Restriction defined on them for membership. An example of the representations of these subsets, can be viewed as a part of the KG creation documentation on our resources website.

5.2 Modeling of Characteristics and Descriptive Statistics

As briefly introduced in Subsect. 4.1, subject characteristics are the phenotype properties that are collected for study subjects. In our evaluation dataset, we have observed that all study arms belonging to a study share the same set of characteristics. However, the range of values for these characteristics differ across study arms depending on their composition. Borrowing from our grouping of characteristics from Sect. 4, we reemphasize that characteristics persisting over a period of time are modeled as sio:hasProperty, and the rest are modeled via sio:hasAttribute property. From this discussion it becomes apparent that our modeling of characteristics on study arms is fairly straightforward and we only utilize two SIO property associations. In Listing 1.1, we depict the association of age as a sio:hasAttribute of the *Ramipril study arm*. Further, characteristics can also be classified broadly as categorical, discreet, and continuous. Categorical characteristics are represented in subsets, and their representation is discussed in the previous subsection.

5.3 Modeling of Descriptive Statistics

Another problem we address in this paper is the KR of aggregate statistics on subject characteristics of study populations. Although aggregate statistics are reported in multiple domains, there has been little work on a convention for supporting the modeling of aggregations in RDF. The support for aggregations in Linked Data is presented in [6]. However, their process is more focused on the publishing of statistical data and the metadata than on the representation of statistical data itself.

Descriptive statistics have conventionally been defined, as statistical measures that summarize the data.[9] In Table 1s, they are used to describe summarized values of the characteristics of study subjects, who belong to a study arm.

[8] More Table 1 reporting style and composition details at https://prsinfo.clinicaltrials.gov/webinars/module6/resources/BaselineCharacteristics_Handouts.pdf.

[9] Definition adapted from: https://en.wikipedia.org/wiki/Descriptive_statistics.

From our analysis of Table 1s, we have seen a limited set of descriptive statistics measures: mean +/− standard deviation, median +/− interquartile range, and percentages. We model these aggregations and descriptive statistics, seen as reified triples on a property. Reification is an RDF technique developed to "make statements about statements" [18]. As can be seen in the RDF snippet above, we define descriptive statistics as reified triples on an age characteristic. Additionally, since we only reuse SIO object and data properties, we eliminate the need for further punning techniques, to represent these descriptive statistic properties as instances of sio:hasAttribute. In this paper, we only present an example of a mean +/− standard deviation measure. Examples of representing median +/− interquartile via sio:MinimalValue and sio:MaximalValue boundary classes, and percentage association, can be viewed on our resources website.

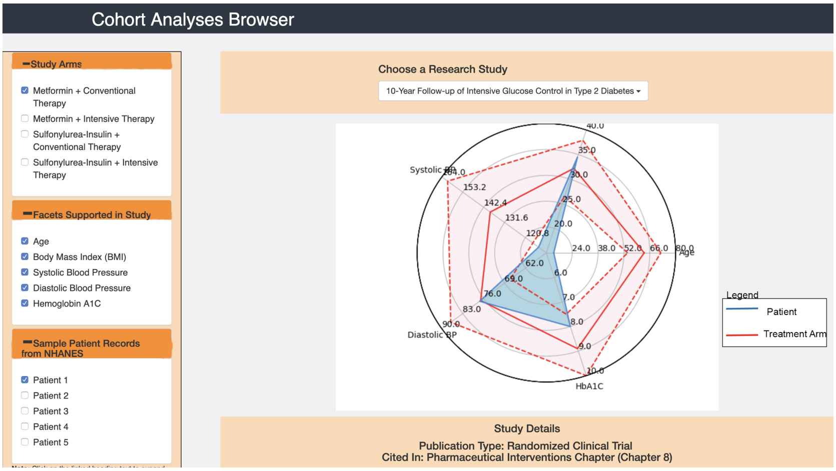
6 Applications

Our study applicability applications leverage the declarative specifications of study populations in our Table 1 KG. In Subsect. 6.1, we frame three scenarios of clinical relevance that mimic the decision-making of a medical practitioner to determine study applicability. Additionally, we present a cohort similarity visualization strategy in Subsect. 6.2. In Subsect. 6.3, we describe a faceted browser interactive visualization tool aimed at medical practitioners. Moreover, as shown in Fig. 1, we include study details in our application results. Hence, we provide medical practitioners with provenance-justified results that could be used for future analyses and investigation.

6.1 Population Analysis Scenarios

As discussed in Sect. 1.1, there exist challenges with study biases and the varying quality in research studies. Medical practitioners need to be aware of these issues when deciding on applicable studies for their clinical population. Three scenarios of clinical relevance were suggested by our medical expert on the Health Empowerment by Analytics, Learning and Semantics (HEALS) project. Through queries to our Table 1 KG we address a representative competency question for each of these scenarios: (1) Study match: Is there a study that matches this patient on a feature(s)? (2) Study limitation: Is there an absence or an under-representation of population groups in this study? (3) Study quality evaluation: Are there adequate population sizes and is there a heterogeneity of treatment effects among arms? Our declarative representations of Table 1s, allow us to trigger retrospective queries that combine subject characteristics (SPARQL AND clauses), various descriptive statistical representations (limited patterns of modeling as seen in Sect. 5), and aggregate study arms or study cohorts (leveraging SPARQL math constructs such as SUM). Our competency questions and their SPARQL queries[10] can be found on our resources website.

[10] https://tetherless-world.github.io/study-cohort-ontology/application#scenarioquery.



Cohort Analyses Browser

−Study Arms

- ☑ Metformin + Conventional Therapy
- □ Metformin + Intensive Therapy
- □ Sulfonylurea-Insulin + Conventional Therapy
- □ Sulfonylurea-Insulin + Intensive Therapy

−Facets Supported in Study

- ☑ Age
- ☑ Body Mass Index (BMI)
- ☑ Systolic Blood Pressure
- ☑ Diastolic Blood Pressure
- ☑ Hemoglobin A1C

−Sample Patient Records from NHANES

- ☑ Patient 1
- □ Patient 2
- □ Patient 3
- □ Patient 4
- □ Patient 5

Choose a Research Study: 10-Year Follow-up of Intensive Glucose Control in Type 2 Diabetes ▾

Study Details

Publication Type: Randomized Clinical Trial
Cited In: Pharmaceutical Interventions Chapter (Chapter 8)

Fig. 4. A snapshot of our faceted browser tool that provides medical practitioners with the ability to customize cohort analyses. Currently, the feature facets are limited to the patient features from NHANES, that overlap with, Table 1 data. If a study doesn't contain some of these 5 features, they will be disabled.

6.2 Cohort Similarity Visualizations

We define cohort similarity as an analytical problem to determine the similarity or closeness of a patient to a given study population. We currently support determination of cohort similarity by generating visualizations, such as a star plot (Fig. 4), by overlaying features of patient records against study arm characteristics. For the purpose of visualization, we select a few sample type-2 diabetes patient records from the National Health and Nutrition Examination Survey (NHANES)[11]. Additionally, we adopt different visual strategies for continuous and categorical variables. In this paper, on the resources website and through our faceted browser we only support star plot visualizations for continuous variables, and we are exploring visualizations such as a pie chart for categorical variables. Visualizations are generated on a per study arm, per patient basis, through results of SPARQL queries triggered to our Table 1 KG. Our visualizations are built by Python plotting modules such as Seaborn[12] and Matplotlib[13], and our visualization code is made available as a resource.

[11] Dataset Information Page. https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2015.

[12] https://seaborn.pydata.org/.

[13] https://matplotlib.org/.

Since our visualizations serve the purpose of being quick assessors, we design them with reduced complexity. Specifically, we aim for them to (1) contain sufficient detail that is not considered overwhelming and (2) carry information such as variable ranges and the extent of the patient match, to serve as indicators for future analysis.

6.3 Faceted Browser

We built a faceted browser tool for medical practitioners by utilizing a Python model-view-controller framework, Flask[14]. On the backend (model), we utilized the RDFLib[15] module to trigger SPARQL queries on the ingested ontology and KG files. Through this tool medical practitioners can interact with our Table 1 KGs, and run cohort similarity analyses on studies of their choice. They can choose from a list of studies and, subsequently, a faceted view will be rendered for the study arms of this selected study. As seen in Fig. 4, they can also choose the variables that they would like to visualize. Hence, our prototype faceted browser interface serves as a per-study inspection tool and uses NHANES patient records to illustrate the facets.

7 Results

In the Study Analysis Table 1, shown below, we present a quantitative summarization of the results of each competency question (described in Subsect. 6.1). Some interesting, medically relevant inferences that we output, and that are often spoken about in medical literature, include the lack of a representation of adults above 70,[16] and the lack of heterogeneity in treatment effects.[17] We were surprised that only 6% of the studies in our evaluation dataset were conducted on a large-scale, that their study arms were evenly divided, and all their study subjects were put on the basic, antidiabetes treatment of *guanidines*. We also find that the SCO ontology is epistemologically adequate for representing all Table 1s in our evaluation dataset. We cover 360 ($\approx$ 17 in each study on average) subject characteristics from 20 cited research studies, and 28 study arm definitions. The study arm definitions included terms belonging to classes such as medical interventions, control regimes, and, less commonly occurring, diseases, dosage, year of follow-ups, and titration targets. We found that 19 cohort variables (a term we use to collectively describe interventions and subject characteristics) commonly occur across studies.

[14] http://flask.pocoo.org/.

[15] https://rdflib.readthedocs.io/en/stable.

[16] https://www.statnews.com/2019/01/31/nih-rule-make-clinical-research-inclusive/.

[17] NIH Collaboratory run grand-round presentation: https://www.nihcollaboratory.org/Pages/Grand-Rounds-02-28-14.aspx.

Table 1. Percentage of studies meeting the competency question criteria for the population analysis scenarios.

Question	Percentage	Population analysis type
Studies with a representation of Male African American study subjects	75%	Study match
Study Arms with adults below the age of 70	47.6%	Study limitations
Studies with cohort sizes $>$ 1000 and study arm administered drugs of the guanidines family, with sizes 1/3rd those of the cohort size	6%	Study quality evaluation

8 Resource Contributions

We expect the following publicly available artifacts, along with the applicable documentation, to be useful resources for anyone interested in performing analysis on study populations reported in research studies.

1. Ontologies:
   (a) Study Cohort Ontology (SCO)
2. Knowledge Graphs:
   (a) Table 1 Knowledge Graph
3. Source Code:
   (a) MIREOT Script
   (b) Cohort Similarity Visualization
4. Data:
   (a) NHANES Patient Records

9 Future Work

Having demonstrated our ability to apply semantic techniques to make study populations visible, we plan to incorporate interdiscplinary methods to improve on a few aspects of our solution. We have found that there exist variances in Table 1 reporting styles ranging from differences in row and column headers, table formats etc. These variances pose challenges for the scalability and automation aspects of the KG construction. Furthermore, often some subject characteristics and column headers require a contextual understanding for disambiguation, that is present in the unstructured body of the study. Hence, we are exploring a combination of natural language processing and semantic techniques to support an ontology-driven parsing and clean-up of Table 1 data and to develop a contextualized and medical standards compliant Table 1 KG. Further, to ensure longevity and easy reuse of SCO, we plan to develop a set of tools/algorithms to predict the best-fit position for a new term in our SCO suite of ontologies. We also plan to expand and refine our set of competency questions, based on feedback from medical practitioners, and to allow for partial and fuzzy matches using query relaxation [9] and semantically targeted analytics [14].

10 Discussion

We have utilized KR techniques, i.e. OWL encodings of SCO and a knowledge graph of Table 1 content to model and expose descriptions of study populations in an attempt to make scientific data more accessible. Further, we have utilized our semantic modeling to support analytical use cases to determine study applicability. Our evaluation dataset currently is solely comprised of type-2 diabetes research studies. We have kept our descriptions and examples minimally domain specific. We believe that our ontology and KG documentation can serve as resources for researchers interested in the pan-disease analysis of study populations.

Our ontology, SCO, is developed using best-practice ontology principles, some of which are listed at [7]. Specifically, we reuse SIO properties and do not define any new properties. We reuse classes from a limited yet standard set of biomedical ontologies in order to increase the interoperability of SCO.

There have been attempts at improving the reporting of Table 1s in the medical community, such as the Table 1 project [15]. However, they have been confined to the identification of desirable properties for standardization. Our semantic solution presented in this paper, that at its heart utilizes a KR approach, is a step towards achieving this standardization. This can be seen in Listing 1.1 where we have presented an RDF snippet representing fundamental building blocks of our Table 1 KG, i.e. our modeling of collections, subject characteristics, and statistical measures. These identified patterns are reused as templates to realize the association of various variables with study populations reported in Table 1s.

Our Table 1 KGs allows us to address study applicability scenarios motivated from medical literature and to support visualizations that clearly depict cohort similarity. By these capabilities, we demonstrate how our solution addresses our use case of determining study applicability. We believe there is potential for this work to be reused by researchers performing study population analyses. Also in this paper we make assumptions on the content a medical practitioner might want to see, and, from a medical practitioner user survey we are conducting, we will incorporate feedback on their additional requirements.

Our solution does not address or include support for the modeling of study eligibility criteria, i.e. inclusion and exclusion criteria. However, we reuse metadata expression terms from Dublin Core Terms (DCT) to include a link to registries such as ClinicalTrials.gov or International Standard Randomised Controlled Trial Number (ISRCTN),[18] where the criteria is made available as a part of the study data. We expect that the SCO vocabulary is sufficient to express the criteria, but since we are still investigating the merge of the criteria with the Table 1 content, we defer it to future work.

Finally, all the resources that we listed in Sect. 8, are made publicly available in a Github repository and the ontology is hosted on Bioportal. SCO is released

[18] http://www.isrctn.com/page/about.

under the Apache 2.0 license specification. Our resources will be maintained periodically by the authors.

11 Conclusion

We have presented a prototype KR system that can be used to model study populations, to aid in the assessment of study applicability. Our model is tailored around use cases aimed at assisting medical practitioners in the treatment of complex patients and who also often require “efficient-literature searching” [12] capabilities. We presented a solution to make descriptions of study populations more accessible for quick decision-making. We believe that the resources we release, especially SCO, can serve as an extensible schema to represent population descriptions across diseases. We have demonstrated the adequacy of the ontology through a set of what we believe are representative applications supporting a range of use cases contributed by our medical expert. We plan to continue our outreach and ontology reuse in additional diverse evidence-based medicine application settings.

Acknowledgements. This work is partially supported by IBM Research AI through the AI Horizons Network. We thank our colleagues from IBM Research, Dan Gruen, Morgan Foreman and Ching-Hua Chen, and from RPI, John Erickson, Alexander New, and Rebecca Cowan, who greatly assisted the research.

References

1. American Diabetes Association (ADA) et al.: 8. Pharmacologic approaches to glycemic treatment: Standards of medical care in diabetes - 2018. Diabetes Care **41**(Suppl. 1), S73–S85 (2018)
2. American Diabetes Association (ADA) et al.: 9. Cardiovascular disease and risk management: standards of medical care in diabetes - 2018. Diabetes Care **41**(Suppl. 1), S86–S104 (2018)
3. Auer, S., Kovtun, V., Prinz, M., Kasprzik, A., Stocker, M., Vidal, M.E.: Towards a knowledge graph for science. In: Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics, p. 1. ACM, Novi Sad (2018)
4. Bechhofer, S., et al.: OWL web ontology language reference. OWL Reference Guide. https://www.w3.org/TR/owl-ref/
5. Courtot, M., et al.: MIREOT: The minimum information to reference an external ontology term. Appl. Ontol. **6**(1), 23–33 (2011)
6. Cyganiak, R., Field, S., Gregory, A., Halb, W., Tennison, J.: Semantic statistics: bringing together SDMX and SCOVO. In: Proceedings of the Linked Data on the Web Workshop (LDOW 2010), Raleigh, North Carolina, USA, 27 April 2010 (2010). http://ceur-ws.org/Vol-628/. Accessed 26 Mar 2019
7. Garijo, D., Poveda-VillalÃşn, M.: A checklist for complete vocabulary metadata. List of Desirable Ontology Best-Practices. http://dgarijo.github.io/Widoco/doc/bestPractices/index-en.html
8. Graham, R., et al.: Trustworthy clinical practice guidelines: challenges and potential. In: Clinical Practice Guidelines We Can Trust, pp. 53–75. National Academies Press (US), Washington D.C. (2011)

9. Hurtado, C.A., Poulovassilis, A., Wood, P.T.: Query relaxation in RDF. J. Data Semant. X **4900**, 31–61 (2008)
10. Ontarget Investigators: Telmisartan, ramipril, or both in patients at high risk for vascular events. N. Engl. J. Med. **358**(15), 1547–1559 (2008)
11. Jang, M., Jahanshad, N., Espiritu, R.: The cohort ontology. Enigma Knowledge Capture and Discovery Project. https://knowledgecaptureanddiscovery.github.io/EnigmaOntology/release/cohort/1.0.0/index-en.html
12. Masic, I., Miokovic, M., Muhamedagic, B.: Evidence based medicine-new approaches and challenges. Acta Inform. Med. **16**(4), 219 (2008)
13. National Institute of Health (NIH): Rigor and Reproducibility. Introduction and need for principles. https://www.nih.gov/research-training/rigor-reproducibility
14. New, A., Rashid, S.M., Erickson, J.S., McGuinness, D.L., Bennett, K.P.: Semantically-aware population health risk analyses. Presented as a Poster at Machine Learning for Health (ML4H) Workshop, NeurIPS, Montreal, Canada (2018). https://arxiv.org/abs/1811.11190. Accessed 20 Mar 2019
15. NIH Colloboratory: Table 1 project. Rethinking Clinical Trials. https://sites.duke.edu/rethinkingclinicaltrials/ehr-phenotyping/table-1-project/
16. Noy, N.F., et al.: BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res. **37**(suppl$_2$), W170–W173 (2009)
17. Patel, C., et al.: Matching patient records to clinical trials using ontologies. In: Aberer, K., et al. (eds.) ISWC 2007. LNCS, vol. 4825, pp. 816–829. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-76298-0_59
18. Reinhardt, S.: Property reification vocabulary. A Strawman Draft. https://www.w3.org/wiki/PropertyReificationVocabulary
19. Shankar, R.D., Martins, S.B., O'Connor, M.J., Parrish, D.B., Das, A.K.: Epoch: an ontological framework to support clinical trials management. In: Proceedings of the International Workshop on Healthcare Information and Knowledge Management, pp. 25–32. ACM, Arlington (2006)
20. Sim, I., et al.: The ontology of clinical research (OCRe): an informatics foundation for the science of clinical research. J. Biomed. Inform. **52**, 78–91 (2014)
21. Tu, S.W., et al.: A practical method for transforming free-text eligibility criteria into computable criteria. J. Biomed. Inform. **44**(2), 239–250 (2011)
22. Valdez, J., Kim, M., Rueschman, M., Socrates, V., Redline, S., Sahoo, S.S.: Provcare semantic provenance knowledgebase: evaluating scientific reproducibility of research studies. In: AMIA Annual Symposium Proceedings, vol. 2017, p. 1705. American Medical Informatics Association, Washington D.C., USA (2017)
23. Xiang, Z., Courtot, M., Brinkman, R.R., Ruttenberg, A., He, Y.: OntoFox: web-based support for ontology reuse. BMC Res. Notes **3**(1), 175 (2010)
24. Younesi, E.: A knowledge-based integrative modeling approach for in-silico identification of mechanistic targets in neurodegeneration with focus on Alzheimer's disease. Ph.D. thesis, Department of Mathematics and Natural Sciences, Universitäts-und Landesbibliothek Bonn, Bonn, Germany (2014)