Ontology-enabled Analysis of Study Populations

Shruthi Chari[1][0000−0003−2946−7870], Miao Qi[1][0000−0002−2917−0965], Nkechinyere N. Agu[1][0000−0003−1386−8602], Oshani Seneviratne[1][0000−0001−8518−917X], James P. McCusker[1][0000−0003−1085−6059], Kristin P. Bennett[1][0000−0002−8782−105X], Amar K. Das[2][0000−0003−3556−0844], and Deborah L. McGuinness[1][0000−0001−7037−4567]

[1] Rensselaer Polytechnic Institute, Troy, NY 12180, USA
[2] IBM Research, Cambridge, MA

Abstract. We address the problem of modeling study populations in research studies in a declarative manner. Research studies often have a great degree of variability in the reporting of population descriptions. To make study populations easily accessible for decision making related to study applicability, we will show the usage of our ontology-enabled prototype system in different applications. Our system leverages our Study Cohort Ontology and the related cohort Knowledge Graph (as described in our accepted resource track paper). We aim to address three retrospective population analysis scenarios, designed to specifically determine the study match, study limitations, and evaluate the study quality. We also provide visualizations of a patient (or patient population) to a treatment arm. In addition, for each guideline recommendation that depends upon a study, we provide a summary of the relevant study's cohort description. We describe some of our applications and their potential impacts.
Resource Website: `https://tetherless-world.github.io/study-cohort-ontology/`

Keywords: Ontology Development · Analytics supported by Knowledge Graphs · Determination of Study Applicability

1 Introduction

[3]Treatment recommendations in Clinical Practice Guidelines (CPG) are often supported by evidence from clinical trials and observational case studies (collectively referred to as research studies). When medical practitioners are determining whether a study applies to their patient, they may consider the similarity of the study population to their patient. Characteristics of population descriptions are reported in tabular formats, often in the first table of research studies, more popularly called Table 1s. These Table 1s contain summarized descriptive statistics of characteristics (e.g., demographics, and anthropometric properties) for a set of study subjects belonging to treatment arms in the study. We developed the Study Cohort Ontology (SCO), reusing terms from existing biomedical

ontologies to build a conceptualization of the study subject collections, the associated characteristics and the descriptive statistics on them. Through RDF knowledge graphs (KGs) modeled on SCO, we represented the Table 1s of 20 research studies that are cited in the Pharmacologic Interventions (Chapter 8) and the Cardiovascular Complications (Chapter 9) of the American Diabetes Association (ADA) Standards of Medical Care 2018 CPG.[4] Our ontology development and KG modeling approach are described in greater detail in our resource track paper, and in this poster we elaborate on the applications enabled by our semantic modeling.

Through the suggestions and validation of the medical practitioner on our team, we design and address three scenarios of clinical relevance by our ontology-enabled system: (1) study match - determine if a study population is similar to a given patient, (2) study limitation - expose population underrepresentations, and (3) study quality evaluation - analyze Table 1s to check for conformance to required best practices. Additionally, through KGs, we support cohort similarity visualizations, in which we overlay patient records against the treatment arms, that serve as quick comparisons for a study match. SCO also contributes to the larger goal of the Health Empowerment by Analytics, Learning, and Semantics (HEALS) project.[5] In HEALS, we are also developing a guideline provenance (G-Prov) ontology to model the provenance behind guideline recommendations. We depict a use case integrating the two ontologies to summarize population descriptions of research studies backing a guideline recommendation.

2 Related Work

Existing ontologies (e.g., [3,2]) for scientific literature (medical in particular) have been largely focused on addressing study design methods and do not specifically address cohort modeling scenarios. ProvCaRe, an "Ontology for provenance + healthcare research"[3], has some level of support for study data, which is limited to the study inclusion and exclusion criteria. However, their vocabulary doesn't handle the granularity and associations necessary to model the characteristics and descriptive statistics recorded on study populations. Through SCO, we support the modeling of aggregations on study populations at a disease-agnostic level. Additionally, our semantic modeling of baseline characteristics of populations enable retrospective population analyses and other patient matching capabilities, as described in section 3.

3 Applications

3.1 Population Analysis Scenarios

Each of our population analysis scenarios (introduced in section 1) are implemented by SPARQL queries to our cohort KGs. On our resources website, we

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

[5] HEALS: `https://idea.tw.rpi.edu/projects/heals`

have an example SPARQL query for a competency question for each of the scenarios. During our bottom-up approach to construct SCO and the cohort KGs, we found that we can broadly model Table 1s' content as the modeling of collections (i.e., study arms or categorical variables such as race) of study subjects, their characteristics, and the descriptive statistics associated with characteristics recorded on these collections. These templates, that are elaborated on in our resource track paper, provide the flexibility to easily frame SPARQL queries to identify the study bias, ascertain study match and determine study quality, etc. We were able to find a general underrepresentation of older adults above 70 ($\approx 50\%$) and a lack of large, clinical trial studies (with population size ≥ 1000). Our current set of competency questions serve as a proof of concept in our ability to draw interesting and medically relevant conclusions about study populations that can assist medical practitioner. We are working on adding more competency questions to each class of population analysis scenario.

3.2 Cohort Similarity Visualizations

Another use case includes our visualizations, star plots that serve as quick determiners of patient fit to a study and are generated on a per patient, and per study arm basis. These plots are generated by a Python script that triggers a SPARQL query to a Blazegraph endpoint to retrieve continuous variables of treatment arms from the cohort KG. Additionally, patient data for these same sets of variables are also retrieved in the script. In the absence of actual EHR data, we evaluated using representative diabetic patients from the National Health and Nutrition Examination Survey (NHANES) dataset.[6] We are exploring other visualization strategies for categorical variables. As seen from the star plot in Fig. 1, we map the distributional spread of the variables from the treatment arms (i.e. mean +/- standard deviation, median and interquartile ranges) against patient values for each of these variables. For example, on the age axis, we see that participants in the Metformin arm had an average age of 53 +/- 14 and the patient's age (≈ 65) fell within this spread.

3.3 Visualizing the application of G-Prov and SCO

In general, population health managers (PHM) aim to improve the overall health outcomes of a patient population by monitoring the features that affect their health. To assist the PHM in monitoring the health of their patient population, we integrate SCO with the G-PROV ontology. G-PROV is used to capture the provenance of CPG recommendations and link the provenance to the cited research studies that back the recommendations. We represented the Table 1s of studies in cohort KGs. Finally, we displayed the information using graphs, charts, and tables. This view (as seen in Fig. 1) provides a PHM with a quick way of visualizing the population descriptions within research studies backing CPG recommendations.

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

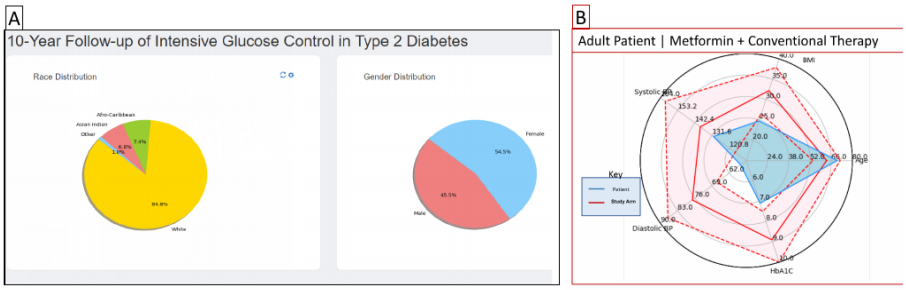

Fig. 1. A) Screenshot of the PHM view visualizing the statistical spread of two categorical variables, i.e., race and gender. B) Star plot overlaying a patient record against that of the Metformin study arm from the same "10-Year Follow-up of Intensive Glucose Control in Type 2 Diabetes"[1] study.

4 Conclusion

We have introduced some applications of our ontology-enabled prototype system that supports analyses via standardized representations of cohort descriptions in KGs. Our applications address the use case of making evidence-based medicine resources accessible. In our case, study populations are made more accessible for medical practitioners dealing with the treatment of complex patients. We have released SCO as an open-source resource along with our documented use case applications. Additionally, we are expanding upon our ontology-enabled system to make it more scalable, including techniques for automatic extraction of cohort descriptions from studies and support for a larger array of applications.

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. Holman, R.R., Paul, S.K., Bethel, M.A., Matthews, D.R., Neil, H.A.W.: 10-year follow-up of intensive glucose control in type 2 diabetes. New England J. Medicine **359**(15), 1577–1589 (2008)
2. Sim, I., Tu, S.W., Carini, S., Lehmann, H.P., Pollock, B.H., Peleg, M., Wittkowski, K.M.: The ontology of clinical research (ocre): an informatics foundation for the science of clinical research. J. Biomed. Informatics **52**, 78–91 (2014)
3. 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 Annu. Symp. Proc. vol. 2017, p. 1705. Amer. Med. Inform. Assoc., Washington D.C., USA (2017)