Ontology-enabled Analysis of Study Populations

Shruthi Chari[1](charis@rpi.edu), Miao Qi[1](qim@rpi.edu), Nkechinyere N. Agu[1](agun@rpi.edu), Oshani Seneviratne[1](senevo@rpi.edu), James McCusker[1](mccusj2@rpi.edu), Kristin P. Bennett[1](bennek@rpi.edu), Amar Das[2](amardas@us.ibm.com), Deborah L. McGuinness[1](dlm@cs.rpi.edu)

[1] Rensselaer Polytechnic Institute 110 8th St., Troy, NY, USA 12180 | [2] IBM Research, Cambridge, MA, USA

Annotated Example of a "Table 1"

Key: Study Arm; Subject Characteristic; Descriptive Statistic

Table 1. Baseline Characteristics of the Patients.*

Variable	Data Available in Final Year of Post-Trial Monitoring: Yes (N=1525)	No (N=705)	P Value†	Sulfonylurea–Insulin Group: Conventional Therapy (N=880)	Intensive Therapy (N=2118)	P Value†	Metformin Group: Conventional Therapy (N=309)	Intensive Therapy (N=279)	P Value†
Age — yr	62±8	60±9	0.002	63±9	63±9	0.78	63±9	64±9	0.56
Male sex — no. (%)	892 (58.5)	391 (55.5)	0.22	532 (60.5)	1248 (58.9)	0.44	142 (46.0)	127 (45.5)	0.92
Race or ethnic group — no. (%)‡			<0.001			0.59			0.42
White	1161 (76.1)	564 (80.0)		710 (80.7)	1717 (81.1)		262 (84.8)	235 (84.2)	
Afro-Caribbean	143 (9.4)	51 (7.2)		58 (6.6)	159 (7.5)		23 (7.4)	28 (10.0)	
Asian Indian	209 (13.7)	84 (11.9)		105 (11.9)	230 (10.9)		21 (6.8)	12 (4.3)	
Other	12 (0.8)	6 (0.8)		7 (0.8)	12 (0.6)		3 (1.0)	4 (1.4)	
Weight — kg			0.97			0.01			0.42
Median	81.0	81.00		79.0	80.0		87.0	86.0	
Interquartile range	71.00–92.00	70.00–91.00		69.0–90.0	71.0–92.0		76.0–97.0	75.0–95.8	
Body-mass index	29.4±5.5	29.4±5.4	0.86	28.7±5.6	29.3±5.5	0.005	32.2±5.7	31.7±5.4	0.34
Blood pressure — mm Hg									
Systolic	137±19	137±19	0.98	138±21	139±20	0.52	139±22	141±18	0.43
Diastolic	77±10	78±10	0.22	77±10	77±10	0.06	77±10	78±10	0.22
Fasting plasma glucose — mg/dl	164±59	168±61	0.34	178±58	161±61	<0.001	182±55	177±64	0.12
Glycated hemoglobin — %			0.25			<0.001			0.12
Median	8.0	8.1		8.5	7.9		8.9	8.4	
Interquartile range	6.9–9.4	7.0–9.6		7.3–9.7	6.8–9.2		7.5–10.0	7.2–9.7	
Cholesterol — mg/dl									
Total	198±39	198±37	0.66	197±37	197±39	0.63	200±37	204±41	0.37
Low-density lipoprotein	127±34	127±32	0.81	126±32	126±34	0.92	129±32	130±36	0.98
High-density lipoprotein	42±12	43±13	0.87	43±12	42±13	0.23	40±12	42±13	0.08
Triglycerides — mg/dl			0.60			0.97			0.10
Median	127	132		128	127		143	157	
Interquartile range	84–184	88–190		88–180	85–182		103–203	107–231	
Plasma creatinine — mg/dl			0.87			0.61			0.03
Median	1.00	0.98		1.02	1.02		0.96	1.03	
Interquartile range	0.87–1.15	0.89–1.14		0.89–1.17	0.90–1.17		0.83–1.11	0.85–1.23	
Ratio of albumin to creatinine§			0.99			0.42			0.48
Median	12.3	12.7		14.9	14.3		19.9	19.8	
Interquartile range	6.2–33.9	5.8–36.7		6.5–49.7	6.8–43.8		8.1–82.8	8.0–61.2	

Fig. 1: A Table 1 example from the "10-Year Follow-up of Intensive Glucose Control in Type 2 Diabetes" [1] clinical trial, cited in the Pharmacologic Interventions Chapter (Chapter 8) [2] of the ADA Standards of Medical Care 2018 Clinical Practice Guideline (CPG)

Representing population description of research studies in a knowledge graph using the Study Cohort Ontology (SCO) helps physicians visually determine their applicability to patients.

Keywords
Biomedical Ontology Development; Knowledge Graphs; Analysis of Study Populations

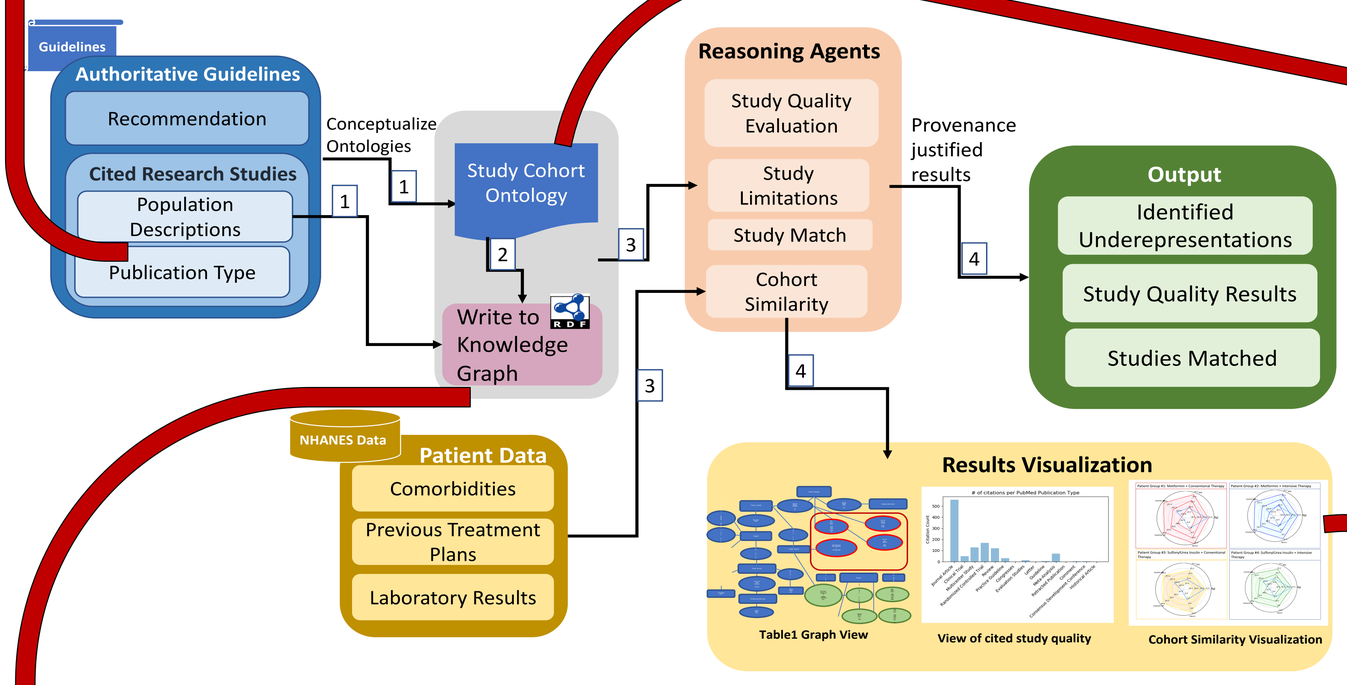

Fig. 2: An overview of the cohort analytics workflow illustrating 1). recovery of population descriptions from studies, 2). representation of them and 3). a few analyses that are powered off the cohort knowledge graphs. Numbering is in-line with the figure.

Ontology Overview

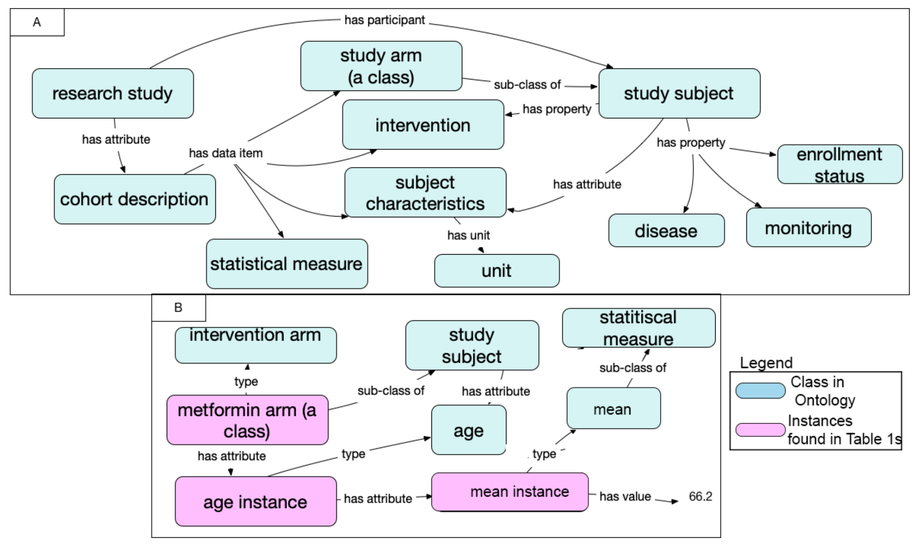

Fig. 3: A). High level overview of the model of SCO to support encoding of Table 1 data.
B). A modeling example of an aggregation reported on an age attribute for the study subjects belonging to the the Metformin study arm from the Table 1 example in Fig. 1

Sample Knowledge Graph

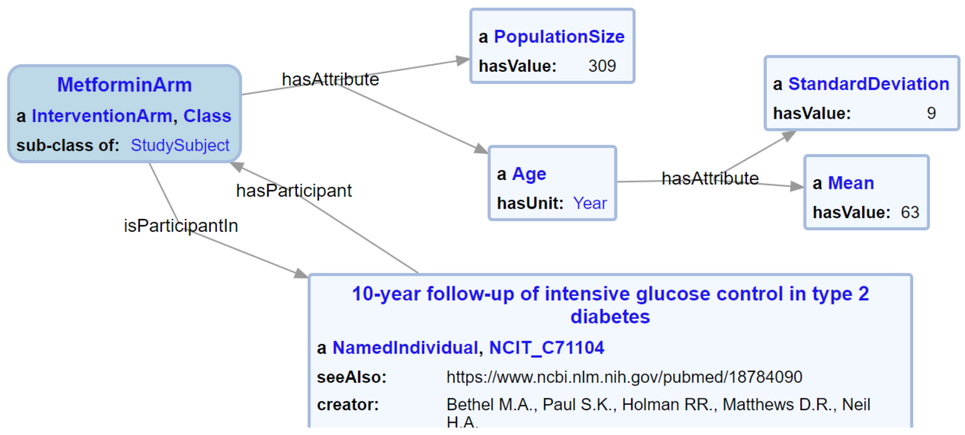

Fig. 4: Visual overview of a KG snippet capturing the modeling of a group of study subjects from the Table 1 example in Fig. 1

Applications

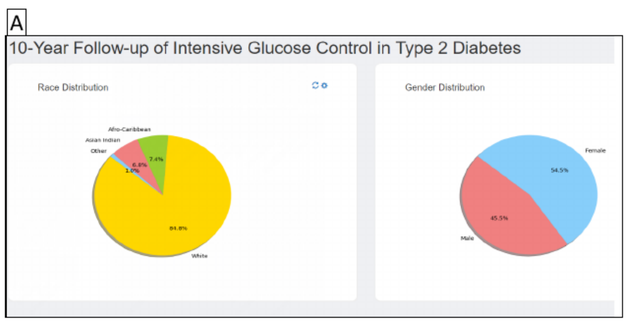

Fig. 5: A). Screenshot of the Population Health Manager view visualizing the statistical spread of two categorical variables

Fig 5: B). Star plot overlaying a patient record against that of a study

Results

We evaluated our workflow on **20 research studies** cited in the pharmaceutical (Chapter 8) and cardiovascular complications (Chapter 9) chapters of the ADA 2018 CPG.

Question	Percentage	Population Analysis Type
Studies with a representation of Male African American study subjects	75%	Study Match
Study Arms without adults above the age of 70	20%	Study Limitations
Studies with cohort sizes > 1000 and study arm administered drugs of the guanidines family, with sizes 1/3 of those cohort sizes	6%	Study Quality Evaluation

Table 1: Summary of results for the competency questions that were designed for each of the population analysis scenarios

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. American Diabetes Association, 2018. 8. Pharmacologic approaches to glycemic treatment: Standards of Medical Care in Diabetes-2018. *Diabetes care*, *41*(Suppl 1), p.S73
3. Agu, N. N., Keshan, N., Chari, S., Seneviratne, O., McCusker, J. P., Das, A., McGuinness, D. L., G-PROV: Provenance management for clinical practice guidelines. in Proceedings of the Semantic Web solutions for large-scale biomedical data analytics Workshop. CEUR, p. to appear (2019)

Glossary

1. **Research Study**: A scientific investigation that involves testing a hypothesis. Normally include clinical trials, case studies and meta-analysis
2. **Table 1s**: Population descriptions are often reported in the first table of research studies, hence referred to as Table 1s
3. **Study Arm:** An arm in a clinical trial is a group of subjects receiving the same therapeutic intervention (or none)

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.

View more at: https://tetherless-world.github.io/study-cohort-ontology/

Our ISWC 2019 Resource Track Paper