Representing Study Populations in Scientific Literature in Knowledge Graphs

Shruthi Chari[1], Miao Qi[1], Nkechinyere N. Agu[1], Oshani Seneviratne[1], James P. McCusker[1], Kristin P. Bennett[1], Amar K. Das[2], Deborah L. McGuinness[1]
[1]Rensselaer Polytechnic Institute, Troy, NY
[2]IBM Research, Cambridge
{charis, qim, agun, senevo, mccusj2, bennek}@rpi.edu, amardas@us.ibm.com, dlm@cs.rpi.edu

***Abstract*—Treatment recommendations in clinical practice guidelines (CPG) are supported by evidence from research studiesthat utilize populations with highly selective sociodemographic and comorbid characteristics. When physicians are treating complicated patients, who do not wholly align with guideline recommendations, they need to determine the applicability of a study to their clinical population. We have designed the Study Cohort Ontology (SCO) and used it to build a knowledge graph (KG) exposing study populations in the CPGs published by the American Diabetes Association (ADA).**

I. From Table to Knowledge Graph

CPGs exhibit an implicit evidence model since guideline recommendations are based on evidence from clinical trials and observational case studies, referred to here as research studies.

Our knowledge representation approach exposes descriptions of study populations, which are often reported in the first table of research studies, hence referred to as Table 1s. We analyzed research studies from the Pharmacologic Interventions and Cardiovascular Complications chapters of the ADA Standards of Care 2018 CPG[1], for patterns across Table 1s. SCO is designed to be an extensible, domain-agnostic ontology, and we reuse terms from existing biomedical ontologies, as much as possible. We do not introduce any new properties, instead we utilize the properties provided by the mid-level SemanticScience Integrated Ontology (SIO).

As seen in figure 1, Table 1s are comprised of study arms, a group of study subjects who receive an intervention or are put on a control regime, whose subject characteristics are aggregated and reported via descriptive statistic measures. Our KGs model and support these Table 1 components, and do so while mitigating the variance in Table 1 structures.

Through the declarative specification of study populations in our KG using the Resource Description Framework (RDF), we address three scenarios of clinical relevance (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. Further, we generate cohort similarity visualizations, to overlay patient records against those of study arms, to depict cohort similarity, at a glance.

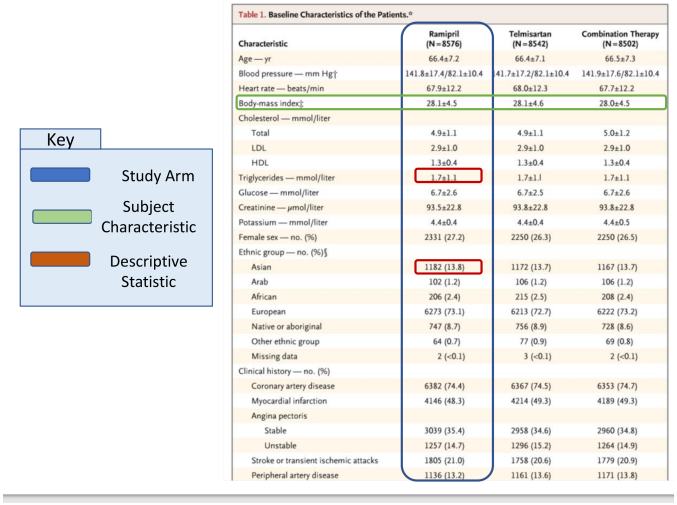

Table 1. Baseline Characteristics of the Patients.[a]

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±1.1	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. 1. An annotated example of Table 1 from a clinical trial "Telmisartan, ramipril, or both in patients at high risk for vascular events"[3]cited in the cardiovascular complications (Chapter 9) of the ADA CPG

II. Conclusion and Future Work

The reporting styles of population descriptions vary on a per study basis, and we have seen variances in table formats, row and column headers, etc. Moreover, often the content in Table 1s requires contextual understanding for disambiguation, which is present in other sections in research studies, such as study methods and study design in the respective paper. To address these scalability and automation challenges, we plan to combine Natural Language Processing and Semantic techniques to build an ontology-driven parsing and clean-up of extracted Table 1 content and to identify study data of relevance to incorporate into our Table 1 KGs.

Our semantic solution ultimately supports physicians in their decision-making of determining study applicability, and serves as an attempt to make scientific study data more accessible.

Acknowledgements

This work is partially supported by IBM Research AI through the AI Horizons Network.

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

[3]https://www.nejm.org/doi/full/10.1056/nejmoa0801317