Semantic Modeling of Cohort Descriptions in Research Studies

Shruthi Chari¹, Rukmal Weerawarana², Oshani Seneviratne PhD¹, James P. McCusker PhD¹, Deborah L. McGuinness PhD¹, Amar Das MD PhD³

¹ Rensselaer Polytechnic Institute, Troy, New York, ² Stevens Institute of Technology, Hoboken, New Jersey, ³ IBM Research, Cambridge, Massachusetts

1 Introduction

Authors of clinical practice guidelines (CPGs) regularly cite clinical trial results and observational case studies (referred here as research studies) as scientific justification for treatment recommendations. Often research studies are based on highly selective populations with restricted sociodemographic or comorbid characteristics. As a result, one of the challenges that physicians face in using CPG recommendations is determining how well the study evidence applies to a general clinical population. Previous computer-based CPG models[1] have been built in a depth-first manner, primarily modeling the semantics and structural design of CPGs. These efforts did not capture the nature of the evidence used as provenance of CPG recommendations. We are addressing this limitation by modeling the patient attributes in cohort descriptions used in the underlying research studies, with the ultimate goal of allowing physicians to better target clinical recommendations. We have begun efforts on a prototype system rooted in Semantic Technologies to extract and model the baseline characteristics of cohorts utilized for cited research studies in the American Diabetes Association (ADA)'s Standards of Medical Care in Diabetes 2018² CPG. Our system ultimately aims to parse publications broadly identified as research studies, and extract the cohort variables and exposure or intervention groups defined within the structured population descriptions, and append them to the Cohort Ontology (CO). Individual research studies will be instantiated as RDF Knowledge Graphs, modeled on CO, and, utilizing similarity techniques along with a query language (currently SPARQL), we plan to align patients to research studies.

2 Methodology

We are exploring a breadth-first approach to enrich the ADA CPG recommendations with cohort information from cited research studies to allow both physicians and computers to assess similarity between the study population and a particular patient in a clinical population.

2.1 Guideline and Other Knowledge Sources

Information in the ADA 2018 CPG is in an unstructured format and is freely available on the web. We used web scraping techniques to convert the CPG into Computer Interpretable Guidelines (CIG) - structured JSON format. We designed intermediate evidence models around recommendation blocks, encapsulating recommendations with their associated evidence sentences (contain citations) and their constituent citations. We leveraged Pubmed metadata of cited Medline publications (Publication Types³ in particular), to retain citations that are either observational case studies, clinical trials, or meta analyses. For the purpose of evaluation, we use a clinician’s expert knowledge to map candidate evidence sentences of, for example, “Pharmacologic Approaches to Glycemic Treatment” Chapter (Chapter 8), Section 3⁴ to their corresponding recommendation.

2.2 Ontology Construction

We considered 8 research studies from citations embedded within evidence sentences of Chapter 8, and used the population characteristics of the study often represented in the first table or ‘Table 1’ of the scientific article, to manually construct CO using OWL. On examining the various Table 1’s we observed a pattern in that the findings of research studies are backed by a combination of patient groups differentiated by various interventions, each group characterized

¹This work is partially supported by IBM Research through the AI Horizons Network
³See https://www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T.publication_types/
⁴Visit http://care.diabetesjournals.org/content/41/Supplement_1/S73
by a set of ‘patient group characteristic’s. We adopted a bottom-up approach to model CO to contain ‘study’ as the main class, with ‘patient group’ as study participants (sio:hasParticipant). The interventions (‘study intervention’ a ‘sio:MedicalIntervention’) covering treatment or monitoring options, administered to each ‘patient group’ are modeled as properties (sio:hasProperty), and group characteristics like demographics (‘chear:Demographic’), ‘lab reading’ and ‘physical parameter’, which are either measurable (have values (sio:hasValue) and units (sio:hasUnit)) or categorical variables, are treated as attributes (sio:hasAttribute).

3 Results

A snapshot of a portion of CO and an instantiation of a population description of a Diabetes Mellitus Type 2 case study\(^5\) cited (#32) in the discussion section of Chapter 8, is shown in Figure 1. We find that cohort descriptions for citations mapped to the same recommendation share similar patient characteristics and patient groups. We extracted 75 cohort variables from 8 cited research studies, and we find an average of 6 terms recurring in studies.

3 Discussion

Our preliminary efforts on CO have shown that it is possible to find broad patterns amongst population descriptions of cited diabetes research studies. Although we have observed that across research studies, the granularity and terminology of cohort descriptions vary, we plan to use UMLS mappings to link terms across research studies. In the initial phase we have adopted an incremental approach to ontology modeling, initially leveraging biomedical ontologies including CHEAR\(^6\), Disease Ontology, and foundational ontologies including SIO, we borrow the structure and hierarchy from these ontologies. Existing study and cohort ontologies are often either focused on broad study characteristics and design considerations, such as OCRe\(^2\), or they are domain specific, like CTO-ND\(^7\). Our work augmenting CIGs with cohort descriptions, can assist physicians in keeping up with the evolving guidelines - their recommendations, and cited literature. The ADA releases a new version of their CPG frequently, typically annually. The ADA 2018 CPG contains 15 chapters covering recommendations for diagnosis, treatment options and co-morbidities, with an average of \(\approx 85\) citations per chapter. Our system can ultimately assist physicians to summarize and assimilate the fine details, such as population descriptions and inclusion-exclusion criteria captured in the cited literature, and the augmented guideline representations contribute to the advancement of precision medicine.

References


\(^6\)View at: https://bioportal.bioontology.org/ontologies/CHEAR
\(^7\)View at: https://bioportal.bioontology.org/ontologies/CTO