Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.ijmijournal.com

A data-driven concept schema for defining clinical research data needs

Gregory W. Hruby[a], Julia Hoxha[a], Praveen Chandar Ravichandran[a], Eneida A. Mendonça[b,c], David A Hanauer[d,e], Chunhua Weng[a,*]

[a] Department of Biomedical Informatics, Columbia University, NY, New York, USA
[b] Department of Pediatrics, University of Wisconsin, Madison, WI, USA
[c] Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI, USA
[d] Department of Pediatrics, University of Michigan, Ann Arbor, MI, USA
[e] School of Information, University of Michigan, Ann Arbor, MI, USA

ARTICLE INFO

Article history:
Received 29 November 2015
Received in revised form 4 March 2016
Accepted 31 March 2016

Keywords:
Medical informatics
Comparative effectiveness research
Needs assessment
Data collection
Models
Theoretical

ABSTRACT

Objectives: The Patient, Intervention, Control/Comparison, and Outcome (PICO) framework is an effective technique for framing a clinical question. We aim to develop the counterpart of PICO to structure clinical research data needs.
Methods: We use a data-driven approach to abstracting key concepts representing clinical research data needs by adapting and extending an expert-derived framework originally developed for defining cancer research data needs. We annotated clinical trial eligibility criteria, EHR data request logs, and data queries to electronic health records (EHR), to extract and harmonize concept classes representing clinical research data needs. We evaluated the class coverage, class preservation from the original framework, schema generalizability, schema understandability, and schema structural correctness through a semi-structured interview with eight multidisciplinary domain experts. We iteratively refined the schema based on the evaluations.
Results: Our data-driven schema preserved 68% of the 63 classes from the original framework and covered 88% (73/82) of the classes proposed by evaluators. Class coverage for participants of different backgrounds ranged from 60% to 100% with a median value of 95% agreement among the individual evaluators. The schema was found understandable and structurally sound.
Conclusions: Our proposed schema may serve as the counterpart to PICO for improving the research data needs communication between researchers and informaticians.

1. Introduction

The rich data made available by electronic health records (EHRs) represents a promising resource for accelerating clinical and translational research [1]. However, medical researchers face significant barriers to accessing EHR data, including the articulation of their often abstract and vague data needs without knowing data details and to mapping these needs to fine-grained, contextual lower-level data representations. Two mechanisms for overcoming the barrier to mapping the data need to EHR data representations are self-service query tools [2–4] and common data elements (CDE) [5–7]. The latter are developed for standardizing research data collection and retrieval. However, complex data needs often cannot be specified in the current generation of self-service query tools [8]. At the same time, CDEs have not been widely adopted and suffer from their limited coverage, which is a common problem in clinical terminologies. As such, many medical researchers find existing query formulation solutions inadequate to help them resolve their data needs and hence have to ask an informatician to aid their data retrieval using a process called biomedical query mediation (BQM) process [8,9]. A big part of the BQM process involves mapping abstract medical concepts to local heterogeneous data representations, while most of these data are not defined using CDEs. Moreover, it is impractical to validate the structural and content comprehensiveness of a research data query using the large number of CDEs. A preferred and more practical approach would be an abstracted concept schema that summarizes key concept classes representing clinical research data needs at a higher level. An unorganized list of many CDEs may be overwhelming to a researcher. In contrast, a concept schema can organize medical concepts com-

* Corresponding author at: Department of Biomedical Informatics, Columbia University, 622 West 168 Street, PH-20, New York, NY 10032, USA.
E-mail addresses: chunhua@columbia.edu, cw2384@cumc.columbia.edu (C. Weng).

http://dx.doi.org/10.1016/j.ijmedinf.2016.03.008
1386-5056/© 2016 Elsevier Ireland Ltd. All rights reserved.

mensurate with the way in which medical researchers organize those concepts. This will allow researchers to refer to the concept classes to ensure the comprehensiveness of their data requests without reviewing the extensive lists of all medical concepts.

Information needs assessment is an established research field. For any information-seeking endeavor, users are required to specify their information needs upfront [10]. In the realm of EHR data requests, task-oriented static online query forms have been explored to enable medical researchers to specify their research data needs [11]. Templates, which guide users to specify their information needs with increased specificity, have been shown effective at structuring an information need request and improving the precision and recall of information needs [12]. The best template example in the medical domain is the PICO framework [13], where P standards for population, I for intervention, C for control or comparison, and O for outcome. PICO is an effective technique for expressing information needs free of ambiguity [14] and improves information retrieval accuracy [15,16]. The PICO framework has been shown to be effective at improving the resolution of information needs for medical literature [12,17]. The success of PICO inspired us to develop its counterpart for articulating clinical research data needs.

Carpenter et al. developed a conceptual framework to define data needs for cancer research [18] based on semi-structured interviews and focus groups with over 76 stakeholders, including providers, researchers, industry representatives and journal editors. The framework defines data types, such as patient characteristics, diagnosis, treatment, and outcomes, as well as their temporal and association relations. The framework also represents the iterative nature of the cancer care continuum [18]. The framework provides a semi-granular representation of data needs yet remains compact enough to achieve an efficient representation of a complex information space. If able to extend beyond cancer, this framework may serve as a template for defining data requests for medical research in general.

Therefore, this study aims to use a data-driven approach to adapt and extend the Carpenter framework to achieve an enriched concept schema for defining clinical research data needs beyond the cancer domain. Our study validated and extended the Carpenter framework utilizing three data sources that represent researchers' data needs in various medical domains.

2. Methods

The study design is illustrated in Fig. 1. Three data sources were processed and analyzed to identify discrete variables for specifying research data needs. We used the Carpenter framework as the starting point for data annotation and iterative schema enrichment. We performed an evaluation with eight multidisciplinary medical researchers and refined the resulting class schema for representing generic clinical research data needs accordingly. This study has received the approval from Columbia University Institutional Review Board.

2.1. Data sources and characteristics

Our three data sources include the public clinical trial inclusion/exclusion criteria obtained from ClinicalTrials.gov, EHR data requests submitted to our institutional clinical data warehouse, and EHR SQL queries obtained from the Department of Urology at Columbia University. The data sources represent a diverse set of values across the attributes of (1) data request type, (2) representativeness of all data needs, and (3) granularity of EHR data needs. For example, clinical research eligibility criteria represent high-level research cohort requests that are independent of the knowledge about what is retrievable from the EHR. Therefore, they tend to be vague, ambiguous, or non-granular representations of a researcher's need. In contrast, EHR data requests are expressed by a mixture of narrative descriptions of medical concepts or various terminologies frequently used in EHRs, such as ICD-9 or 10 codes or CPT codes. Finally, SQL queries are translations of EHR data requests into executable database queries. They reflect the needs of researchers based on not only what is retrievable from the EHR but also how these available data elements are encoded. Therefore, they represent the data needs at the lowest level of concept granularity (*e.g.*, a specific representation such as "A1c" or "HbA1c" in discharge summaries or a local code for A1c in lab test results tables). We assumed these three data sources provide a rich and complementary representation for the data needs of medical researchers. Table 1 provides a detailed description of the datasets used for this project. The next section will discuss our sampling strategy for each data source.

2.2. Data sampling

To obtain a representative sample of sentences from the clinical trial eligibility criteria, we extracted 2,729,525 sentences from 181,356Clinical Trials downloaded from the public Clinicaltrials.gov on 2/12/2015. We annotated the concepts in these sentences with UMLS sematic types using a previously published method [19]. Using the K-means clustering algorithm [20], we divided all the enriched sentences into 27 classes. To cover sentences from these classes evenly, we sampled 1000 sentences evenly from these clusters for further annotation. For the EHR data requests logs, we randomly sampled 432/1200 data requests submitted to our data request service at Columbia University in the 2014 calendar year. A total of 897 sentences were extracted from these request logs. For the SQL queries, we used the SQL transact code associated with the 204 research projects performed at our institution's Department of Urology over the course of five years (2008–2012). For each project SQL code, we selected the "SELECT* FROM* WHERE*" statements and isolated the "SELECT *" clause for annotation.

2.3. Dataset annotation and analysis

Author GH annotated the datasets. This coder has 10 years of experience conducting research and 6 years of experience resolving medical researchers' data requests. We did not ask two independent annotators to annotate the datasets and measure inter-rater agreement for the following reasons. First, our goal was not to evaluate the Carpenter framework as an annotation tool, nor the process used to annotate the datasets, but to assess the portability of this framework beyond cancer and its coverage of concepts in other disease domains. Therefore, annotation is a means to achieve our goal, not the end. Second, the purpose of employing two independent annotators followed by a measurement of the inter-rater agreement is to ensure reproducible annotations generated manually. However, previous studies have reported limitations in employing inter-rater agreement for ensuring the reliability of human annotations. An example paper is provided at [21]. In this paper, the authors reported the complexities involved in reporting inter-rater reliability and some simplified inter-rater agreement calculation and reporting methods may not necessarily be reliable. Given such concerns about the limitations in the inter-rater ability assessment itself, we were more inclined to utilize a data-driven approach rather than a human-driven approach to achieve our goal. Therefore, our annotation was a semi-automatic process, which uses NLP-assisted concept recognition followed by manual mapping of each sentence represented by a set of terminology-encoded concepts into a class defined in the Carpenter model. The