International Journal of Medical Informatics 93 (2016) 74–84

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.ijmijournal.com

A multi-site cognitive task analysis for biomedical query mediation

Gregory W. Hruby[a], Luke V. Rasmussen[b], David Hanauer[c,d], Vimla L. Patel[a,e], James J. Cimino[a,f], Chunhua Weng[a,*]

[a] Department of Biomedical Informatics, Columbia University, New York, NY, USA
[b] Division of Health and Biomedical Informatics, Northwestern University Feinberg School of Medicine, Chicago, IL, USA
[c] Department of Pediatrics, University of Michigan, Ann Arbor, MI, USA
[d] School of Information, University of Michigan, Ann Arbor, MI, USA
[e] The New York Academy of Medicine, New York, NY, USA
[f] Informatics Institute in School of Medicine, University of Alabama, Birmingham, AL, USA

ARTICLE INFO

Article history:
Received 22 February 2016
Received in revised form 8 June 2016
Accepted 9 June 2016

Keywords:
Electronic health records
Task performance and analysis
Information storage and retrieval

ABSTRACT

Objective: To apply cognitive task analyses of the Biomedical query mediation (BQM) processes for EHR data retrieval at multiple sites towards the development of a generic BQM process model.
Materials and methods: We conducted semi-structured interviews with eleven data analysts from five academic institutions and one government agency, and performed cognitive task analyses on their BQM processes. A coding schema was developed through iterative refinement and used to annotate the interview transcripts. The annotated dataset was used to reconstruct and verify each BQM process and to develop a harmonized BQM process model. A survey was conducted to evaluate the face and content validity of this harmonized model.
Results: The harmonized process model is hierarchical, encompassing tasks, activities, and steps. The face validity evaluation concluded the model to be representative of the BQM process. In the content validity evaluation, out of the 27 tasks for BQM, 19 meet the threshold for semi-valid, including 3 fully valid: "Identify potential index phenotype," "If needed, request EHR database access rights," and "Perform query and present output to medical researcher", and 8 are invalid.
Discussion: We aligned the goals of the tasks within the BQM model with the five components of the reference interview. The similarity between the process of BQM and the reference interview is promising and suggests the BQM tasks are powerful for eliciting implicit information needs.
Conclusions: We contribute a BQM process model based on a multi-site study. This model promises to inform the standardization of the BQM process towards improved communication efficiency and accuracy.

1. Background and significance

Rich clinical data made available by the Electronic Health Records (EHRs) are invaluable for medical knowledge discovery [1–3]. However, as such data increases in volume, velocity, and variety, biomedical researchers face significant data access barriers [4], including convoluted regulatory processes [5], inconsistent and limited data quality reporting [6], and opaque data representations [7]. To facilitate data access, data analysts have developed self-service query tools to enable biomedical researchers to navigate and query EHR data autonomously [8–11]. These self-service tools support a wide range of users with simple data needs, but are often unable to represent complex data queries or provide contextual guidance for query clarification [12,13]. They have reduced the barrier for some medical researchers but do always resolve complex queries.

Each medical condition may have multiple data representations in EHRs, structured or unstructured, which are collected for billing or clinical care purposes of discrepant data quality [6]. If structured, the coding schema can be from a broad range of clinical terminologies, such as ICD-9, ICD-10, ICD-O, SNOMED, and so on. Regardless of the terminology used, the real life clinical scenario does not necessarily match up one-to-one with the structured documentation. For example, a cohort with Crohn's Disease or ulcerative colitis can be retrieved using at least two instances of any of the five

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

http://dx.doi.org/10.1016/j.ijmedinf.2016.06.006
1386-5056/

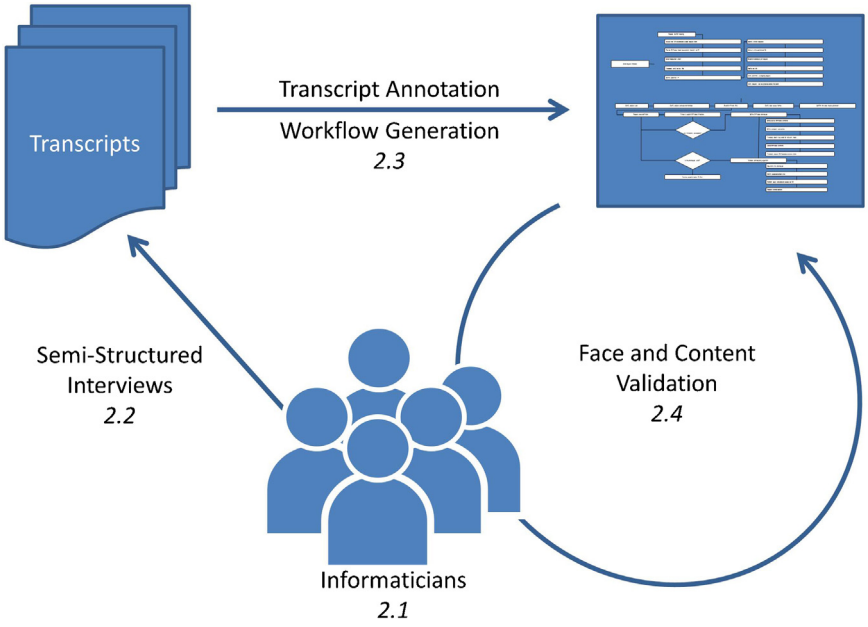

Fig. 1. The high-level overview of our research process with numbers indicating section headers in the manuscript, which was initiated with semi-structure interviews. We annotated the transcripts from these interviews to generate individual and general task flow representations. Final, we produced and evaluated the harmonized task model.

related ICD-9 diagnosis codes within a two-year time window [14]. Computable queries of a medical condition may vary across institutions due to phenotype differences in heterogeneous population subgroups and variation in EHR documentation or EHR data representation so that the identification of a cohort with the condition is non-trivial [15,16], which often makes data analysts indispensable for assisting with the data extraction process [17].

We refer to this process for understanding data needs and mapping medical concepts to data representations that are informed by knowledge of data availability and data quality as the biomedical query mediation (BQM) process [7]. Barriers to communication are often due to a lack of technical knowledge about data structures, terminology definitions, and database query languages among medical researchers and a lack of medical knowledge about medical care, medical definitions, and clinical care documentation processes among database technicians, respectively. We know little about the complexities of this process, or specifically, the sequence of tasks as well as knowledge and skills needed to communicate effectively between medical researchers and database technicians.

Meanwhile, information needs negotiation has been studied extensively in library science [18–20]. Librarians use a well-established process called "reference interview" to facilitate skilled needs negotiation between a librarian and an information seeker to translate a vague information need into an unambiguous, computable query [21] by iteratively eliciting tacit user needs, verifying implied assumptions and improving the specificity of data queries.

2. Objective

This study aims to characterize the BQM process and align it with the reference interview approach. Previously, we established a preliminary understanding of BQM processes for one institution [7,22]. Our initial understandings of BQM are narrow in scope as they are based on a study of a single institution and one data analyst's approach to BQM. To gain a deeper and more generalizable understanding of the task complexity of the BQM process, we conducted a multi-site cognitive task analysis of the BQM processes to construct a harmonized representation for the BQM process and its common tasks. Different from previous studies on one institution, here we utilized the cognitive task analysis (CTA) protocol described by Clark et al. [23] to yield information about the knowledge, thought processes, and steps for each task [24].

3. Methods

An overview of our research process is given in Fig. 1. The following sections provide details for each step of our methodology.

3.1. Participants

To identify potential participants, we asked central data warehouse managers at the participating sites to invite their data query analysts for participation via an email. We enrolled analysts on a first come first serve basis. Between May 2013 and May 2014, we recruited a convenience sample of 11 data analysts from five academic institutions (i.e., Columbia University, University of Colorado at Denver, University of Wisconsin at Madison, Northwestern University, and Kansas University) and one governmental institution (New York City Department of Health and Mental Hygiene). Table 1 provides details about the data analysts interviewed for this project. All the participants consented to be recorded. We used the interview transcripts for the analysis. This study has received the approval from Columbia University Institutional Review Board (#AAAJ8850).

3.2. Semi-structured interview

We conducted a semi-structured interview to elicit the details of the BQM process used at each institution. The interview questions were organized into three parts. In part one, to establish a general understating of the data analyst's process for BQM, we asked each participant to elaborate on their tasks, the expected outcomes of those tasks, and the knowledge required to perform those tasks and the knowledge source. This part also prepares the participants for performing a hypothetical BQM in the second part, in which we presented three information need scenarios from published comparative effectiveness research studies [25–29]. We asked each participant to randomly select a scenario, which we decomposed