Toward a Cognitive Task Analysis for Biomedical Query Mediation
Gregory W. Hruby, MA1, James J. Cimino, MD1,2, Vimla Patel1,3 and Chunhua Weng, PhD1
1. Department of Biomedical Informatics, Columbia University, New York City;
2. NIH Clinical Center, Bethesda, Maryland;
3. New York Academy of Medicine, New York City

Abstract
In many institutions, data analysts use a Biomedical Query Mediation (BQM) process to facilitate data access for medical researchers. However, understanding of the BQM process is limited in the literature. To bridge this gap, we performed the initial steps of a cognitive task analysis using 31 BQM instances conducted between one analyst and 22 researchers in one academic department. We identified five top-level tasks, i.e., clarify research statement, explain clinical process, identify related data elements, locate EHR data element, and end BQM with either a database query or unmet, infeasible information needs, and 10 sub-tasks. We evaluated the BQM task model with seven data analysts from different clinical research institutions. Evaluators found all the tasks completely or semi-valid. This study contributes initial knowledge towards the development of a generalizable cognitive task representation for BQM.

Introduction
Helping researchers access “Big Data” in the electronic health record (EHR) is essential for both public health initiatives and comparative effectiveness research (CER) in many academic medical centers, but remains a costly endeavor. In reality, CER involves complex, ultra-granular information needs that necessitate assistance from data analysts to extract representative data from the EHR. To do this, the medical researcher’s information need must be transferred to the data analyst, who may translate that information need into a precise and specific data query. The transfer of the information need from the medical researcher to the data analyst occurs through an iterative question-answering process. From this point on, we will refer to the transfer of the information need as the Biomedical Query Mediation (BQM) and a medical researcher may be any type of researcher seeking EHR data. During BQM, the data analyst may explain to the medical researcher relevant information of data restrictions and contextual data constraints, e.g., laboratory results may be more accurate than ICD-9 codes for identifying diabetes patients. Such information may guide medical researchers to reconsider and revise their queries.

Analogous to reference interview or interactive information retrieval in the field of Library and Information Science, the success of BQM depends on the effectiveness of the iterative negotiations between the data analyst and the medical researcher. Reference interview elicits a clear and well-defined statement from the patron detailing the information need. However, literature provides neither rich insights into opaque BQM processes nor differences between locating data elements in the massive EHR information space and searching for books in libraries.

Prior studies have shown that modeling interactive retrieval processes can lead to better designs of information retrieval systems. As these studies suggest that and obscure processes, such as BQM, are observed by the few performers, and the resulting understanding of the process is limited. We believe that modeling the knowledge of the series of tasks performed by data analysts during BQM is important for providing standard, user-centered support for data analysts and medical researchers.

We previously reported a content analysis of the BQM between a data analyst and medical researchers. As a natural extension to that study, this paper presents a cognitive task analysis of BQM to illustrate the BQM tasks and knowledge required to perform each task. The purpose of a cognitive task analysis is twofold, first to outline the specific tasks used to accomplish a goal and second, to detail both the controlled and automated knowledge needed to perform each task identified. A cognitive task analysis contains five core steps, i.e., collect preliminary information, identify task knowledge representations, apply focused knowledge elicitation methods, analyze and verify data acquired, and format the cognitive task analysis results for the intended application. This study focuses on the first two steps. We modeled task activities and sequences, and knowledge needed to perform each task. Our task model underwent a face and content validation by external reviewers. Columbia University Medical Center Institutional Review Board approved this study. The rest of this paper first reports the methods and results and then discusses the implications of these findings for enhancing the BQM process.
Methods

Data Collection: Between July 2011 and January 2012, 31 discussions between one data analyst and 22 medical researchers were recorded and transcribed.

Data Analysis: Our analysis focused on the tasks occurring during BQM to accomplish transfer of the medical researcher’s information needs to the data analyst. We extended our previous work’s description of BQM constructs to seed BQM task identification. The seeding content used was (1) the research question, (2) the clinical process, and (3) EHR data elements locations. Through a random selection of BQM transcripts and e-mails we initially identified tasks related to the seeded content and extrapolated sub-tasks related for each of these tasks. Through an iterative task and sub-task identification and refinement process, a final task and sub-task list emerged. Next, we ordered the task list temporarily. As prescribed by cognitive task analysis, we elaborate on the individual task attributes by identifying the task goals and knowledge required to complete each task. Additionally, we constructed a BQM knowledge representation in the form of hierarchical task complexity.

Face and Content Validity Evaluation: We presented our BQM task model to seven external data analysts from medical centers that are known to engage in BQM between data analysts and medical researchers (Northwestern University and Columbia University). We asked the data analysts to fill out a 14-item questionnaire for the derived BQM task list. Two items asked the data analyst to provide their experience with and frequency performing BQM. We used two items to assess face validity on a scale from 1-10 for the dimensions of representativeness of and usefulness for BQM. If the median score was greater than or equal to 7 for representativeness and usefulness, we considered the BQM task list to have face validity. Content validity is a metric determining whether a representation is capable of performing its intended task. Ten items were used to measure content validity; each data analyst judged the 10 sub-tasks as essential, useful, or non-useful. Inter-rater agreement in the form of content validity ratio was applied to assess content validity. Task content validity were achieved if the content validity ratio reaches the minimum critical value of 0.620. Tasks were deemed semi-valid if at least half of the evaluators rated the task as essential.

Results

BQM task complexity: Figure 1 presents the hierarchical complexity representation of the BQM tasks. We identified five tasks, i.e., define research statement, illustrate clinical process, identify related data elements, locate EHR data elements, and end mediation, with ten corresponding sub-tasks. A typical BQM contains iterative topic switching; therefore, the BQM process is not a linear progression among these tasks. Descending the pyramid, each task is broken down to relatively simpler though still quite complicated tasks. This hierarchy also highlights the iterative process between both the clinical process with EHR data element location and other data elements with EHR data element location. As the medical researcher becomes aware of what they do not know about what they think they may know as it pertains to available data elements within the EHR, one of two things may happen: i.e., the BQM may stop or the medical researcher may revise the research statement to accommodate the new knowledge gained through the BQM.

![Figure 1: The complexity hierarchy and task flow for BQM](image-url)
### BQM task process and dimensions

**Table 1: BQM tasks and activities performed by the data analyst**

| Task | Sub-task | Goal | Knowledge Required | Example |
|------|----------|------|--------------------|---------|
| 1. Define research statement | 1.1 Elicit the clinical research scenario | To introduce core data elements of the information need | Study types | *What is the research question?* |
|     | 1.2 Understand the design of the proposed research | To establish the relationships among data elements | Study types | *Are you looking at pre-treatment factors that affect the outcome measure?* |
| 2. Illustrate clinical process | 2.1 Elicit the clinical progression related to the information need | To establish the temporal order of abstract data elements | Medical domain knowledge | *Patients with disease x that undergo treatment y, can you describe the diagnosis, treatment and follow-up timeline?* |
|     | 2.2 Gather specific details and data representations of the ordered abstract data elements | To establish EHR data definitions for abstract data elements | Medical domain knowledge | *Do all doctors refer to treatment X as x? What billing codes/image studies/lab tests are used for that type of visit?* |
|     | 2.3 Create list of unknown data elements | To provide inputs for task 3 | Heuristics | *What is the data element X? Please describe.* |
|     | 2.4 Understand how to calculate derived variables from EHR data elements | To provide calculation parameters for derived variables | Heuristics | *The Duke University risk score takes into account variables x and y using this formula, x/y + 5.* |
| 3. Identify related data elements | 3.1 Elicit relevant abstract data elements not represented in the clinical process | To establish static variables required for the study | Medical domain knowledge | *What demographic information do you need? Any specific comorbidities?* |
| 4. Locate EHR data elements | 4.1 Show or request to see the location of the EHR data element | To establish location of data element within the data model of the EHR | EHR data model; EHR graphical user interface | *I’m unfamiliar with the data element X, where is it recorded in the EHR?* |
|     | 4.2 Describe availability and consistency of data elements | To educate the medical researcher on data quality, accessibility and reliability | EHR data model; Data quality, accessibility, and reliability | *Data element X is not collected in the EHR; Data element Y is available sporadically from patient to patient.* |
| 5. End mediation | 5.1 Inform the medical researcher whether or not the information need can be satisfied | To allow the medical researcher to reformulate their information need or end the BQM | EHR data model; Data quality, accessibility, and reliability | *That data element is contained in a scanned image and can’t be extracted from the EHR.* |

**Evaluator Characteristics:** Of the seven evaluators, 29% (2/7), 43% (3/7), and 29% (2/7) have been facilitating data access for >10, 3-5, and 1-2 years, respectively; 29% (2/7), 43% (3/7), 14% (1/7) and 14% (1/7) facilitate >10, 5-10, 3-5, and 1-2 BQM per month.

**Face and content evaluation results:** Table 2 shows the score distribution and content validity ratio of the 10 items for the 10 sub-tasks from Table 1.

**Table 2: Task Content Validation Results**

| Sub-task | Essential (%) | Useful (%) | Non-Useful (%) | Content Validity Ratio |
|----------|---------------|------------|----------------|-----------------------|
| 1.1      | 71            | 29         | 0              | 0.43                  |
| 1.2      | 71            | 29         | 0              | 0.43                  |
| 2.1      | 57            | 43         | 0              | 0.14                  |
| 2.2      | 100           | 0          | 0              | 1                     |
| 2.3      | 71            | 29         | 0              | 0.43                  |
| 2.4      | 86            | 14         | 0              | 0.71                  |
| 3.1      | 71            | 14         | 14             | 0.43                  |
| 4.1      | 71            | 14         | 14             | 0.43                  |
| 4.2      | 100           | 0          | 0              | 1                     |
| 5.1      | 100           | 0          | 0              | 1                     |

220
The first item addressing face validity, on a scale from 1 (not at all)-10 (completely), rating ‘to what extent does the task model simulate BQM’, the median score was 8 (7-10). The second item addressing face validity, on a scale from 1 (not useful)-10 (very useful), rating ‘how useful is this representation for novice data analysts conducting BQM’ the median score was 8 (6-10).

Discussion

We identified five tasks and 10 sub-tasks used to elicit the medical researcher’s information needs to the data analyst. Additionally, the BQM tasks were categorized into a complexity hierarchy. This representation serves as an initial theoretical framework for BQM. A closer look at similar frameworks used for interactive information retrieval and mediated searching shows some similarities. Specifically, the ASK hypothesis and Berrypicking model share similar task progressions to arrive at a clearer definition of an information seeker’s need. Additionally, Spink’s proposed theoretical framework details seven levels for mediated searching: Problem solving process, information seeking episodes, uncertainty, cognitive styles, interactive search sessions, successive search behavior and sets of situated actions. The proposed BQM does not necessarily overlap with these levels, but provides further depth for the interactive search session, or the dialogue between and a user and a system, in this case the channel to the system, the data analyst. We postulate the level of granularity present in EHR databases contributes significantly to the complexity of BQM. Unlike document databases, EHR database complexity is present in both the breadth of data elements and the features used to describe those data elements. Our theory resonates with other interactive information retrieval experts. Ford et al states, “The deeper and more structured are the knowledge representation formalism adopted, the more difficult it is to develop systems able to accommodate wide-ranging subject content.” System development in this context may be used as a surrogate for the model of the proposed work flow the system is attempting to improve. As such, it can be inferred that increased information breadth and knowledge representation granularity will increase the complexity of the interaction with an information retrieval system.

Complexity hierarchy: Our top-level concept explains the ultimate goal of BQM, querying the EHR database. However, the concepts below it break this high-level concept into simpler atomic concepts. BQM, a critical, but yet complex component of EHR database queries is broken further down into the critical components, first the intent of the information need, what is the research the medical researcher wants to conduct. This is then followed by more granular components of assigning clinical data elements to a clinical time line and explicitly defining abstract elements into representations within the EHR. Additionally, other data characteristics that may not be directly tied to a clinical time are explored and the EHR representation is also defined. This model suggests two feedback loops and several potential BQM stopping points. As the data analyst becomes aware of the medical researcher’s information needs the data analyst has a clearer understanding of EHR information space’s ability to contain a representative dataset. Similarly, as the medical researcher’s awareness of the unknown and what is thought to be known may affect the initial intent thereby augmenting the research statement to accommodate what EHR data are suitable to represent an information need. Of note, knowledge required to perform all tasks come from different sources. These sources aid the data analyst to locate and map data elements to the internal data repository for the medical researcher. Both medical researchers and data analysts can contribute medical knowledge; that is why the conversation between the two is crucial.

BQM task process model: Of particular interest, task 1, define research statement, mirrors a similar process to the reference interview to understand the context for which the informant seeker is working from. Understanding the intent of the researcher provides a scaffolding of core information elements and the relationships of the information elements. Likewise, the initial task of BQM enables the data analyst to develop an internal information model similar to the one being used by the medical researcher MR. This mental model is a semantic relationship of the key medical data elements for which finer details related to those elements can be explored.

Tasks 2, 3, and 4 are used to facilitate the focus of the information need. It has been shown that the focusing of the information seeker’s need provides an increase in precision of the results. While this study did not assess the precision of resulting datasets, tasks 2-4 support the notion of an iterative refinement and understanding of the medical researcher’s information need. Each task builds from the initial task, and the subsequently these task may augment the initial task or end the BQM. The final task represents the data analyst understanding of the medical researcher’s information need and whether or not that need can be met by EHR data. This task can either end the mediation with no results or move the mediation into a formal EHR database query.

Face and content validation: Our expert evaluators deemed the preliminary process model to have face validity. The ratings suggest our initial representation of BQM is both representative of and useful for BQM. Additionally, 40% (4/10) of the sub-tasks, sub-tasks 2.2, 2.4, 4.2, and 5.1, were judged to have content validity. All tasks were
judged either semi-valid or valid. Given these positive results, we believe this initial knowledge representation of BQM is acceptable to continue with our cognitive task analysis by applying focused knowledge elicitation methods with data analysts representing diverse approaches to BQM.

**Limitations:** Our study contains several limitations. This representation is based on a specific data set covering the BQM process for just one data analyst and one medical research domain, urologic oncology. Our findings may not be representative of information needs from other medical domains, nor may it be inclusive of other expert data analyst tasks used to transfer a medical researcher’s information need to them. Additionally, the EHR data being accessed was represented by an integrated research data repository, which contained highly granular data. The type of EHR information source may affect the BQM process. Regardless of these limitations, the initial components of the cognitive task analysis will allow us to move forward with informed semi-structure interviews of other data analyst experts with the sole purpose of extracting additional knowledge of BQM.

**Conclusions**

This study contributes preliminary knowledge of BQM task sequence and a task complexity knowledge representation. This knowledge can guide future work on cognitive task analysis for acquisition of additional information from a diverse group of data analysts on the tasks used to accomplish a generalizable BQM.

**Acknowledgments**

This research was funded under NLM grant R01LM009886, R01LM010815, and 5T15LM007079, and CTSA award UL1 TR000040. Its contents are solely the responsibility of the authors and do not necessarily represent the official view of NIH. The authors would like to thank Luke Rasmussen (Northwestern University) for helping with recruiting evaluators for our study.

**References**

1. D’Avolio LW, Farwell WR, Fiore LD. Comparative effectiveness research and medical informatics. *The American Journal of Medicine.* 2010;123(12):32-37.
2. Hoffman S, Podgurski A. Big Bad Data: Law, Public Health, and Biomedical Databases. *Public Health, and Biomedical Databases (October 30, 2012). Journal of Law, Medicine and Ethics.* Forthcoming, 2012.
3. Rein A. Finding Value in Volume: An Exploration of Data Access and Quality Challenges. *AcademyHealth: Briefs and Reports.* 2012:9.
4. Taylor RS. *QUESTION-NEGOTIATION AN INFORMATION-SEEKING IN LIBRARIES*: DTIC Document; 1967.
5. Kuhlthau CC. Inside the search process: Information seeking from the user's perspective. *JASIS.* 1991;42(5):361-371.
6. Spink A. Study of interactive feedback during mediated information retrieval. *Journal of the American Society for Information Science.* 1997;48(5):382-394.
7. Vakkari P. Task-based information searching. *Annual review of information science and technology.* 2005;37(1):413-464.
8. Suchman L. Making work visible. *Commun. ACM.* 1995;38(9):56-ff.
9. Weir CR, Nebeker JJ, Hicken BL, Campo R, Drews F, LeBar B. A cognitive task analysis of information management strategies in a computerized provider order entry environment. *Journal of the American Medical Informatics Association.* 2007;14(1):65-75.
10. Hruby GW W, A, Weng C. Analysis of Query Negotiation between a Researcher and a Query Expert. Paper presented at: AMIA; 1780, 2012; Chicago.
11. Hruby GW BM, Cimino JJ, Gao J, Wilcox AB, Hirschberg J, Weng C. Characterization of the Biomedical Query Mediation Process. *AMIA Summits on Translational Science Proceedings.* San Francisco2013:5.
12. Schraagen JM, Chipman SF, Shalin VL. *Cognitive task analysis:* Lawrence Erlbaum; 2000.
13. Clark RE, Estes F. Cognitive task analysis for training. *International Journal of Educational Research.* 1996;25(5):403-417.
14. Wilson FR, Pan W, Schumsky DA. Recalculation of the critical values for Lawshe’s content validity ratio. *Measurement and Evaluation in Counseling and Development.* 2012;45(3):197-210.
15. Belkin NJ, Oddy RN, Brooks HM. ASK for information retrieval: Part I. Background and theory. *Journal of documentation.* 1982;38(2):61-71.
16. Bates MJ. The design of browsing and berrypicking techniques for the online search interface. *Online Information Review.* 1989;13(5):407-424.
17. Spink A, Wilson TD, Ford N, Foster A, Ellis D. Information- seeking and mediated searching. Part 1. Theoretical framework and research design. *Journal of the American Society for Information Science and Technology.* 2002;53(9):695-703.
18. Ford N, Ford R. Towards a cognitive theory of information accessing: an empirical study. *Information Processing & Management.* 1993;29(5):569-585.
19. Ross CS, Nilsen K, Dewdney P. *Conducting the Reference Interview: A How-to-do-it Manual:* Neal-Schuman; 2002.