Published in final edited form as:

Int J Med Inform. 2016 September ; 93: 74–84. doi:10.1016/j.ijmedinf.2016.06.006.

A Multi-Site Cognitive Task Analysis for Biomedical Query Mediation

Gregory W. Hruby, MA¹, Luke V. Rasmussen, MS², David Hanauer, MD^{3,4}, Vimla Patel, PhD, DSc^{1,5}, James J. Cimino, MD^{1,6}, and Chunhua Weng, PhD^{1,*}

Chunhua Weng: chunhua@columbia.edu

¹Department of Biomedical Informatics, Columbia University, New York, NY, USA

²Division of Health and Biomedical Informatics, Northwestern University Feinberg School of Medicine, Chicago, IL, USA

³Department of Pediatrics, University of Michigan, Ann Arbor, MI, USA

⁴School of Information, University of Michigan, Ann Arbor, MI, USA

⁵The New York Academy of Medicine, New York, NY, USA

⁶Informatics Institute in School of Medicine, University of Alabama, Birmingham, AL, USA

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

*Corresponding author: Chunhua Weng, Department of Biomedical Informatics, Columbia University, 622 West 168 Street, PH-20, New York, NY 10032.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Author Contributions:

GWH collected data, conducted data analyses and wrote the initial draft of the paper. CW initiated and supervised GWH's research. All authors made substantial contributions to evaluation designs, results interpretation, and refinement of the paper and approved the final draft.

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.

Keywords

Electronic Health Records; Task Performance and Analysis; Information Storage and Retrieval

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 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.

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].

Methods

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

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).

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 into its information components using the PICO (Patient, Intervention, Control/Comparison, and Outcomes) framework [30]. Next, we played the role of a biomedical researcher and simulated the BQM process with the participants in part two. The third part was designed to compare and contrast the data analyst's process with our earlier findings [22]. We first identified new tasks not present in prior model performed by each interviewee. Then, we addressed tasks that were mentioned in the material presented to the data analyst but not identified as an action by the data analyst. We asked the participant if these tasks represented a part of the process that they follow, if the task required additional steps, if the goal of the tasks were accurate and what, if any, additional knowledge was needed to perform the task. Finally, we investigated whether the presented material served as a reminder of additional tasks the data analyst used for BQM. If so, we asked them to elaborate on additional tasks and activities involved to complete the task, the steps to perform each activity, the goals of the tasks, and the knowledge and skills needed to perform the tasks. Appendix 1 provides the interview instruments and example participant input regarding the BQM process and task list.

To provide a validity check for our interpretations of the interviews, we implemented two member checks [31]. First, during the interviews we reiterated concepts presented by the data analysts to ensure the clarity and completeness of information presented during the interview. Second, after completing transcript annotation, we constructed a process workflow representation of the data analyst's BQM process and contacted the participants to verify if the organization and concepts within reflected their view of the BQM process.

Transcript Annotation and Analysis

To identify a comprehensive list of tasks used to conduct a BQM, we performed a thematic analysis by iteratively annotating the interview transcripts [32]. One author (GH) annotated the eleven transcripts using previously developed task representation [22]. After all transcripts were annotated, new tasks identified were assigned a general description and were grouped based on that description. The new codes were used to perform the next round of annotations on the transcripts. This iterative process continued until no new code was generated. After the final annotation round, we constructed the workflow process for each data analyst. We used these workflow process representations to perform our second round of member checking. Through e-mail communication, we presented each participant with his/her process model and asked (1) “does this process model represent your task flow?”, (2) “do you have a problem with any of the language used to describe a particular task?”, and (3) “what task(s), if any, would you remove or add to improve this representation of your workflow?” We augmented the individual process flow models according to the data analysts' input.

After all the data analysts verified their individual workflows, we constructed a hierarchal task list containing tasks, activity(s) to perform a task, and step(s) to complete an activity.

We define an activity as a subtask of a task, and steps as a sub-process of an activity. Additionally, for each level we identified the knowledge needed to perform and the expected outcome for that task. To contextualize the hierarchical task list, we created a harmonized process model from all the individual process models. We assessed the face and content validity of the generalizable task list and process workflow among the study participants.

Evaluation

We presented the consolidated model to the study participants. We asked them to complete a 29-item questionnaire developed using methods described by Lawshe et al. [33]. We assessed face validity by measuring the model's representativeness of the generic BQM workflow and usefulness for facilitating BQM on a ten-point Likert scale. The first item, representativeness, asked "to what extent does the task model simulate BQM" and the second item, usefulness, asked "how useful is this representation for a novice data analyst conducting BQM." We considered the model to have face validity if both dimensions obtained a median score ≥ 7 . We measured content validity for each BQM task by having our study participants rate each task as essential, useful, or non-useful. We established inter-rater agreement in the form of content validity ratio. Task content validity was achieved if the content validity ratio reached the minimum critical value or threshold of 0.620 [34]. We deemed tasks semi-valid if at least half of the data analysts rated the task as essential. Among all semi-valid tasks, those that received unanimous vote as being essential were categorized as valid. All other tasks were invalid. Furthermore, we asked each participant to assess if the task was automatable. Furthermore, positing that the reference interview is a well-documented technique for the extraction of detailed information from vague presentations of information needs and may be applicable for BQM, we compared our model to the reference interview process.

Results

BQM hierarchical task model and process workflow

The hierarchical task model defines all the tasks performed during typical BQM, divided into two phases, a preparation phase and a needs negotiation phase. Both phases contain the (1) tasks, (2) the activities for performing each task, (3) the steps for executing each activity, (4) the knowledge needed to perform the task, and (5) the expected outcome of that task. Table 2 and Table 3 display the corresponding tasks needed to complete the preparation and needs negotiation phases, respectively.

In the BQM preparation phase the data analyst prepares for the face-to-face session by identifying potential cohort case definitions, similar requests, and estimating the amount of time needed to perform the BQM. The needs negotiation phase contains five tasks to complete BQM. The most elaborative task is establishing the index and associated phenotype. The index phenotype is the EHR data representation of the primary medical condition(s) used to establish the patient cohort. Associated phenotypes are other medical conditions needed to identify the patient cohort. For example, if we have a cohort of treated, prostate cancer patients the index phenotype may look like this: one abnormal pre-treatment prostate specific antigen laboratory test, a prostate cancer ICD-9 code 185 assigned before

treatment, and the prostatectomy CPT code 55866. The outcome of prostate-specific antigen recurrence is an associated phenotype and can be defined as two consecutive rises of prostate specific antigen after treatment. Furthermore, the data element attribute ‘pre-treatment’ is not a concept regularly annotated in laboratory data and would need to be inferred. This process is highly iterative as issues surrounding EHR data limitations and phenotype validation will often initiate a new iteration for establishing the index/associated phenotype. The hierarchical task model is visualized in Figure 2.

Face and content validation

Eight of eleven participants completed our validation questionnaire. The model is face valid as the dimensions of representativeness and usefulness scored a median of 9 (7–9) and 8 (4–10), respectively. For content validity, three tasks, “Task 1.5 – Identify potential index phenotype,” “Task 4.3.5 – If needed, request EHR database access rights,” and “Task 4.4.3 – Perform query and present output to medical researcher” are valid. Nineteen out of 27 tasks are semi-valid, as at least half of the data analysts rated the tasks as essential. Eight tasks are invalid: 1.2 – Provide EHR data model educational documentation to MR, 1.6 – Identify similar requests (for query reuse), 1.7 – Consult with experienced data analyst, 1.9 – Profile MR, 1.11 – Verify request was assigned to correct data analyst/team, 4.1 – Request case definition, 4.4.2 – Clarify expected cohort size, and 4.5 – If exists, complete other IP/AP(s). Figure 3 displays the content validity evaluation results.

Comparison to the Reference Interview

Table 4 displays the five filters of the reference interview. For each filter, we match the BQM tasks that best represent the objectives of the filter. We present an example quote from our interviews associated with the task and then contextualize the task with a potential action that the data analyst may use to complete that task.

Discussion

In this paper, we present a method to capture the development process for a generalizable BQM task model used by data analysts to facilitate EHR data access for medical researchers. Below, we discuss the potential implications for this work, expected and unexpected findings, and limitations.

Implications

Our study established that a cognitive task analysis is effective at understanding the process details of a data needs negotiation. That is, we were able to extract procedural knowledge from subject matter experts through targeted interviews and mock needs negotiations in the context of a medical research information need [35]. By doing so, the greatest contribution of this work is the explication of the implicit. The model represents an amalgamated perspective of the tasks used to facilitate BQM and the knowledge needed to complete those tasks. We do not present this process as the ideal or common method used, but rather a comprehensive look at all the potential tasks used in BQM. We believe this representation could provide data analysts a knowledge resource to better document and describe the work they perform. It may serve the novice data analyst as a guide to navigate the BQM process or

even serve the expert data analyst as a reminder for potential tasks they may have forgotten. Furthermore, this knowledge provides the framework for which process re-design can be introduced to increase BQM efficiency, ultimately improving EHR data access for medical researchers.

Expected findings

Face and content validity—Our evaluators showed that the BQM task model generally had face validity. The ratings suggested that this task model was both representative of, and useful for, BQM. However, only a few tasks within the model met the criteria for content validity, while some others were considered invalid. On further investigation, the invalid tasks were found to be uncommon among our study sample of data analysts, and may be due to our small sample size. Nevertheless, our model included all tasks that data analysts may perform during BQM and, as such, some tasks may have been unfamiliar to some of our evaluators.

The low rate of valid and high rate of semi-valid tasks highlights that variation exists amongst the methods used by our participants to arrive at an understanding and translation of medical researchers EHR data needs. We suspect the tasks marked as valid and semi-valid correlate well with the utilization of those tasks by our participants. The high number of semi-valid tasks suggests a common workflow may exist across many institutions despite there being no standard process model and most of the institutions are amendable to an ideal workflow.

Related models of information seeking—Our analysis found that most data analysts completed the BQM process in two phases: (1) pre face-to-face and (2) face-to-face. During the face-to-face phase, most of our data analysts utilized task 3.0 – “Clarify project concept and methods” to understand the context of the project. This allowed the data analysts to minimize assumptions and frame the data in the context of the intended application. All data analysts performed task 4.0 and established the index/associated phenotype. The iterative nature of this task allowed the data analyst and researcher to focus on a clear definition of the EHR data need. This is analogous to other information seeking studies [36] and models, the ASK hypothesis [37] and Berrypicking [38].

Recently, Hoxha et al. investigated the email communications between researchers and query analysts in which they identified a set of dialog acts organized under the task of cohort identification [39]. They grouped dialog acts into the following topics: *Patient Characteristics, Medical Condition, Demographics, Data Source, Data Format, and Results Submission*. Further analysis identified a high occurrence of loops for the dialog act Patient Characteristics in the interaction between the data analyst and researcher. Their findings are consistent with our model. For example, the iterative loop associated with task 4.0 – “Establishing the index/associated phenotype(s)” mirrors the iterative discussion between the data analyst and researcher observed by Hoxha et al [39].

Finally, we posited that the reference interview serves as a gold standard for the elicitation of an information need from an information seeker. We aligned the goals of the tasks within the BQM model with those of the five components of the reference interview. Table 4 compares

and contrasts the five components of the reference interview with tasks from our model. The similarity between the procedural process of these need negotiations is promising and suggests many of the BQM tasks are powerful for the elicitation of non-vague details. It may benefit the data analyst to include all the corresponding tasks of the reference into their respective process.

Unexpected Findings—Our participants identified task 4.2 – “If known, explain EHR data limitations”, as semi-valid. This surprised us, as our previous understanding of data quality concerns was not with the data analyst but resided with the medical researcher. We believe BQM presents an opportunity to share knowledge about EHR data limitations between data analysts and researchers. Furthermore, Weiskopf et al. developed a guideline for data quality assessment that can be potentially used to facilitate data exploration and data quality awareness during information needs negotiation [40]. None of our participants thought any task can be easily automated. This response can be due to either the complexity of the BQM process or the protective attitude of participating data query analysts. Future studies are warranted to further investigate the feasibility of automating BQM tasks.

Limitations

Our study has two major limitations. First, the semi-structured interviews may have contained biases. Nevertheless, this representation may be biased toward the primary author’s initial description of BQM [22]. To avoid this, each participant underwent several rounds of member checks, both during and after the interview to confirm the representation for each data analyst’s process. Second, we only used one coder to annotate the interview transcripts and generate new concepts. As such, our results may be skewed to the interpretation of the coder. To address this issue, we implemented two instances of member checks to ensure that the internal validity of our interpretations from the interviews aligned with the ideas held by our interviewees. Third, we only studied one stakeholder of the BQM process, the data analyst, without exploring the tasks and challenges involved on the biomedical researcher side.

Conclusions

To our knowledge, this paper reports the first effort for applying a cognitive task analysis to capture the process knowledge for BQM. We present BQM in the form of a hierarchical task model by harmonizing BQM processes from multiple institutions. This harmonized BQM model may enable us to optimize the BQM process to improve communication efficiency and accuracy.

Acknowledgments

This research was funded by NLM grant **R01LM009886 (PI: Weng)**, and **5T15LM007079 (PI: Hripcsak)**, and CTSA award **UL1 TR000040 (PI: Ginsberg)**. Its contents are solely the responsibility of the authors and do not necessarily represent the official view of NIH.

Appendix: Materials for the semi-structured interview

1.1 Introduction

1. What is your name?
2. What institution are you affiliated with?
3. What is today's date?
4. How long have you been a query analyst?

1.2 Phase One

Could you please describe the process you use to understand and interpret a data request?
Here are the aspects for helping you describe the process.

- a. What steps do you take?
- b. What are the goals of the steps you use?
- c. What information or knowledge do you use at each step? What is the source of this knowledge (e.g., expert, manager, or terminology dictionary)?

1.3 Phase Two

I would like to conduct a hypothetical situation; I will randomly select one of three situations representing a potential data request, and present it to you as if I were requesting the EHR data. I would ask you to enter into a needs negotiation with me in until you feel confident that you have the necessary information needed to complete a database query sufficient to resolve my data needs.

Scenario One

Study Title: Comparative Effectiveness of Olmesartan and other Angiotensin Receptor Blockers in Diabetes Mellitus[27]

Clinical Research Question: Does OLM increase CVD mortality compared to other ARBs in diabetic patients?

Research Hypothesis: OLM increases CVD mortality Risk in Diabetic patients

P – All patients receiving ARB who started their first does between 2004–2009. They must also be diabetic, have greater than 1 year of follow up, and can't change their ARB to or from OLM

I – Patients receiving OLM

C – Patients receiving other ARBs

O – Primary All cause hospital admission or Death, Secondary ICD-9 Codes for admissions 410, 411.1, 428, 430–438, 530–579, 555–558.

Scenario Two

Study Title: Treatment and 5-Year Survival in Patients With Nonmetastatic Prostate Cancer: The Norwegian Experience[28]

Clinical Research Question: what has been the Norwegian Experience with treating prostate cancer?

Research Hypothesis: Prostate cancer treatment in Norway is effective.

Variables needed:

P – Patients diagnosed with prostate cancer from 2010–2011. Under 75 years of age, clinical stage t1–t3, no t4 tumor, psa less than 100, prostate biopsy, clinical stage, and psa, performance status

I – What was their treatment for prostate cancer? Radiation, Surgical, Cryo, nothing?

C – NA

O – Overall and cancer specific survival,

Scenario Three

Study Title: A clinical nomogram and recursive partitioning analysis to determine the risk of regional failure after radiosurgery alone for brain metastases[29]

Clinical Research Question: What is a patient's regional failure risk as it pertains to brain metastases treated with stereotactic radiosurgery?

Research Hypothesis: A partition nomogram for regional failure of stereotactic radiosurgery for brain metastases can define high, intermediate and low risk patient groups.

Variables needed:

P – Patients with oligometastatic brain metastases

I – Patients treated with single modality stereotactic radiosurgery

C – NA

O – Primary: cumulative regional failure at 1 year (binary variable) defined as the presence of at least one regional failure occurring within one year of initiation of stereotactic radiosurgery. Secondary: overall survival, time to regional failure and cumulative regional failure.

1.4 Phase Three

1. For each new task the EHR data analyst presented that is not within the validated biomedical query mediation task list, the following question will be asked:

- a. Can you describe if this particular task may be related to or fit into the biomedical query mediation task list presented to you?
2. For each task represented in the biomedical query mediation task list the EHR data analyst did not mention in Phase one of the interview the following question will be asked:
 - a. Does this task represent a part of the process that you use? Why or why not?
 - b. Is there adequate information to describe this task?
 - i. Does this task require additional sub-tasks?
 - ii. Is the goal of this task accurate?
 - iii. Is additional knowledge needed to perform this task?
3. Has the presentation of the biomedical query mediation task list inspired additional tasks you use for biomedical query mediation? If so please answer the following:
 - a. What are the sub-tasks used to complete this task?
 - b. What is the goal of this task?
 - c. What knowledge is needed to perform this task?

1.5 Example process for biomedical query mediation

Table 1 illustrates an example list of biomedical query mediation tasks and lists the tasks, sub-tasks, goals, knowledge needed to perform the task, and examples of that task used by the EHR data analyst to conduct a biomedical query mediation. Figure 1 illustrates an example BQM task flowchart.

Table 1

Biomedical query mediation tasks and activities performed by the data analyst.

Task	Sub-task	Goal	Knowledge Required	Example
Define research statement	1.1 Elicit the clinical research scenario	To introduce core data elements of the information need	Study types	<i>What is the research question?</i>
	1.2 Understand the design of the proposed research	To establish the relationships among data elements	Study types	<i>Are you looking at pre-treatment factors that affect the outcome measure?</i>
Illustrate clinical processes	2.2.1 Elicit the clinical progression	To establish the temporal order of	Medical domain knowledge	<i>Patients with disease x that undergo treatment y, can</i>

Task	Sub-task	Goal	Knowledge Required	Example
	related to the information need	abstract data elements		<i>you describe the diagnosis, treatment and follow-up timeline?</i>
	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	<i>Do all doctors refer to treatment X as x? What billing codes/ image studies/lab tests are used for that type of visit?</i>
	2.3 Create list of unknown data elements	To provide inputs for task 3	Heuristics	<i>What is the data element X? Please describe.</i>
	2.4 Understand how to calculate derived variables from EHR data elements	To provide calculation parameters for derived variables	Heuristics	<i>The Duke University risk score takes into account variables x and y using this formula, $x/y + 5$.</i>
Identify related data elements	3.3.1 Elicit relevant abstract data elements not represented in the clinical process	To establish static variables required for the study	Medical domain knowledge	<i>What demographic information do you need? Any specific comorbidities?</i>
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>I'm unfamiliar with the data element X, where is it recorded in the EHR?</i>
	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	<i>Data element X is not collected in the EHR; Data element Y is available sporadically from patient to patient.</i>
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 biomedical query mediation	EHR data model; Data quality, accessibility, and reliability	<i>That data element is contained in a scanned image and can't be extracted from the EHR.</i>

Figure 1.
Example BQM Task.

References

1. Borycki E, Newsham D, Bates D. eHealth in North America. Yearbook of medical informatics. 2012; 8(1):3.
2. Hruby GW, McKiernan J, Bakken S, Weng C. A centralized research data repository enhances retrospective outcomes research capacity: a case report. J Am Med Inform Assoc. 2013; 20(3):563–567. [PubMed: 23322812]
3. Rasmussen LV. The electronic health record for translational research. Journal of cardiovascular translational research. 2014; 7(6):607–614. [PubMed: 25070682]
4. Ford N, Ford R. Towards a cognitive theory of information accessing: an empirical study. Information Processing & Management. 1993; 29(5):569–585.
5. Hanauer, DA.; Hruby, GW.; Fort, DG.; Rasmussen, LV.; Mendonça, EA.; Weng, C. AMIA Annual Symposium Proceedings. American Medical Informatics Association; 2014. What Is Asked in Clinical Data Request Forms? A Multi-site Thematic Analysis of Forms Towards Better Data Access Support; p. 616-625.
6. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc. 2013; 20(1):144–51. [PubMed: 22733976]

7. Hruby, GW.; B, M.; Cimino, JJ.; Gao, J.; Wilcox, AB.; Hirschberg, J.; Weng, C. AMIA Summits on Translational Science Proceedings. San Francisco: 2013. Characterization of the Biomedical Query Mediation Process; p. 89-93.
8. Deshmukh V, Meystre S, Mitchell J. Evaluating the informatics for integrating biology and the bedside system for clinical research. BMC medical research methodology. 2009; 9(1):70. [PubMed: 19863809]
9. Lowe HJ, Ferris TA, Hernandez PM, Weber SC. STRIDE—An integrated standards-based translational research informatics platform. AMIA Annu Symp Proc. 2009; 2009:391–5. [PubMed: 20351886]
10. Zhang GQ, Siegler T, Saxman P, Sandberg N, Mueller R, Johnson N, Hunscher D, Arabandi S. VISAGE: A Query Interface for Clinical Research. AMIA Summits Transl Sci Proc. 2010; 2010:76–80. [PubMed: 21347154]
11. Horvath MM, Winfield S, Evans S, Slopek S, Shang H, Ferranti J. The DEDUCE Guided Query tool: providing simplified access to clinical data for research and quality improvement. Journal of biomedical informatics. 2011; 44(2):266–276. [PubMed: 21130181]
12. Conway, M.; Berg, RL.; Carrell, D.; Denny, JC.; Kho, AN.; Kullo, IJ.; Linneman, JG.; Pacheco, JA.; Peissig, P.; Rasmussen, L. AMIA Annual Symposium Proceedings. American Medical Informatics Association; 2011. Analyzing the heterogeneity and complexity of Electronic Health Record oriented phenotyping algorithms; p. 274-283.
13. Meystre, SM.; Deshmukh, VG.; Mitchell, J. AMIA Annual Symposium Proceedings. American Medical Informatics Association; 2009. A clinical use case to evaluate the i2b2 Hive: predicting asthma exacerbations; p. 442
14. Bernstein CN, Blanchard JF, Rawsthorne P, Wajda A. Epidemiology of Crohn's disease and ulcerative colitis in a central Canadian province: a population-based study. American journal of epidemiology. 1999; 149(10):916–924. [PubMed: 10342800]
15. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, Basford M, Chute CG, Kullo IJ, Li R. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. Journal of the American Medical Informatics Association. 2013; 20(e1):e147–e154. [PubMed: 23531748]
16. Kho AN, Pacheco JA, Peissig PL, Rasmussen L, Newton KM, Weston N, Crane PK, Pathak J, Chute CG, Bielinski SJ, Kullo IJ, Li R, Manolio TA, Chisholm RL, Denny JC. Electronic medical records for genetic research: results of the eMERGE consortium. Sci Transl Med. 2011; 3(79): 79re1.
17. Hansen P, Järvelin K. Collaborative information retrieval in an information-intensive domain. Information Processing & Management. 2005; 41(5):1101–1119.
18. Janes J. Question Negotiation in an Electronic Age. The Digital Reference Research Agenda. 2003:48–60.
19. Lin J, Wu P, Abels E. Toward automatic facet analysis and need negotiation: Lessons from mediated search. ACM Transactions on Information Systems (TOIS). 2008; 27(1):6.
20. Lankes RD. The Digital Reference Research Agenda. Journal of the American Society for Information Science & Technology. 2004; 55(4):301–311.
21. Taylor RS. Question-negotiation and information seeking in libraries. College & research libraries. 1967; 29(3):178–194.
22. Hruby GW, C J, Patel VL, Weng C. Toward a Cognitive Task Analysis for Biomedical Query Mediation. 2014 Summit on Clinical Research Informatics. 2014:218–222.
23. Clark RE, Estes F. Cognitive task analysis for training. International Journal of Educational Research. 1996; 25(5):403–417.
24. Chipman SF, Schraagen JM, Shalin VL. Introduction to cognitive task analysis. Cognitive task analysis. 2000:3–23.
25. Padwal R, Lin M, Etmnan M, Eurich DT. Comparative Effectiveness of Olmesartan and Other Angiotensin Receptor Blockers in Diabetes Mellitus Retrospective Cohort Study. Hypertension. 2014 p. HYPERTENSIONAHA. 113.02855.

26. Fosså SD, Nilssen Y, Kvåle R, Hernes E, Axcrona K, Møller B. Treatment and 5-year survival in patients with nonmetastatic prostate cancer: the norwegian experience. *Urology*. 2014; 83(1):146–153. [PubMed: 24238563]
27. Rodrigues G, Warner A, Zindler J, Slotman B, Lagerwaard F. A clinical nomogram and recursive partitioning analysis to determine the risk of regional failure after radiosurgery alone for brain metastases. *Radiotherapy and Oncology*. 2014
28. Padwal R, Lin M, Etmnan M, Eurich DT. Comparative Effectiveness of Olmesartan and Other Angiotensin Receptor Blockers in Diabetes Mellitus Retrospective Cohort Study. *Hypertension*. 2014; 63(5):977–983. [PubMed: 24535009]
29. Rodrigues G, Warner A, Zindler J, Slotman B, Lagerwaard F. A clinical nomogram and recursive partitioning analysis to determine the risk of regional failure after radiosurgery alone for brain metastases. *Radiotherapy and Oncology*. 2014; 111(1):52–58. [PubMed: 24444529]
30. Schardt C, Adams MB, Owens T, Keitz S, Fontelo P. Utilization of the PICO framework to improve searching PubMed for clinical questions. *BMC medical informatics and decision making*. 2007; 7(1):16. [PubMed: 17573961]
31. Hoffart N. A member check procedure to enhance rigor in naturalistic research. *Western Journal of Nursing Research*. 1991; 13(4):522–534. [PubMed: 1897212]
32. Weston C, Gandell T, Beauchamp J, McAlpine L, Wiseman C, Beauchamp C. Analyzing interview data: The development and evolution of a coding system. *Qualitative sociology*. 2001; 24(3):381–400.
33. Lawshe CH. A quantitative approach to content validity¹. *Personnel psychology*. 1975; 28(4):563–575.
34. 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.
35. Shortliffe EH, Patel VL. Human-Intensive Techniques. *Clinical Decision Support: the Road Ahead*. 2007:207–26.
36. Kannampallil TG, Franklin A, Mishra R, Almoosa KF, Cohen T, Patel VL. Understanding the nature of information seeking behavior in critical care: implications for the design of health information technology. *Artificial intelligence in medicine*. 2013; 57(1):21–29. [PubMed: 23194923]
37. Belkin NJ, Oddy RN, Brooks HM. ASK for information retrieval: Part I. Background and theory. *Journal of documentation*. 1982; 38(2):61–71.
38. Bates MJ. The design of browsing and berrypicking techniques for the online search interface. *Online Information Review*. 1989; 13(5):407–424.
39. Hoxha J, Chandar P, He Z, Cimino J, Hanauer D, Weng C. DREAM: Classification scheme for dialog acts in clinical research query mediation. *Journal of biomedical informatics*. 2016; 59:89–101. [PubMed: 26657707]
40. Weiskopf, NG. Enabling the Reuse of Electronic Health Record Data through Data Quality Assessment and Transparency. Columbia University: Graduate School of Arts and Sciences; 2015. (Unpublished)

Research Highlights

- We present a validated task model to characterize biomedical query mediation
- We found the biomedical query mediation process to contain similar elements to the Librarian reference interview
- Our study participants believed that BQM tasks cannot be automated.

Summary Points			
What is known about the biomedical query mediation process?		What this study adds to the knowledge?	
•	Query analyst engage in a needs negotiation process with medical researchers seeking EHR data	•	We validated a generic representation of the tasks involved for the biomedical query mediation process
	• To some success, this process is also facilitated by self-service query tools	•	This process shares similar characteristics to that of the librarian reference interview, suggesting tasks within this process are powerful for the elicitation of non-vague details
	• The needs negotiation is a complex, iterative process that maps specific medical concepts requested by the medical researcher to EHR data representations	•	This formal representation may provide a more reliable method for the facilitation of biomedical query mediation

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 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.

Figure 2.

Generalizable BQM task model displayed as a process workflow. The major phases are separated by dashed boxes. Tasks, activities, and steps are represented by solid, dashed, and dotted boxes, respectively. IP – Index Phenotype; AP – Associated Phenotype; DA – Data Analyst; MR – Medical Researcher

Figure 3. Content Validity Evaluation Results. Only three tasks: Tasks 1.5, 4.3.5, and 4.4.3 met the minimal threshold for content validity, Content Validity Ratio (CVR) = 0.62. Nineteen tasks were considered semi-valid as at least half of the participants labeled the task as Essential (CVR=0).

Table 1

Participant	Site	Data Infrastructure	Training	Title	Years of BQM Experience
1	Columbia University	CDW;	BS	User Services Consultant	5
2	Columbia University	CDW;	MS	Data Analyst	2
3	University of Colorado	CWD; i2b2	BS	Data Analyst	2
4	Kansas University	CWD; i2b2	PhD	Query Analyst	2
5	Northwestern University	CDW; i2b2	BS	Informatician	9
6	Northwestern University	CDW; i2b2	BS	Data Architect	2
7	Northwestern University	CDW; i2b2	MS	Statistical Analyst Programmer	3
8	Northwestern University	CDW; i2b2	BS	Data Architect	4
9	Northwestern University	CDW; i2b2	BS	Data Analyst	5
10	NYC Department of health and Mental Hygiene	CDW	MS	Hub Manager	2
11	University of Wisconsin	CDW	MD, PhD	Associate Professor of informatics	1

Table 2

The hierarchical task model for the BQM preparation phase. Each task is described with the knowledge required to complete the task and the expected outcome for that task.

Task	Activity	Knowledge Required*	Expected Outcomes
1.0 Preparing for F2F meeting	1.1 Instruct the MR to complete the data request form	Institutional Policy	Make sure MR is compliant with internal protocols
	1.2 Provide EHR data model educational documentation to MR	Internal Documentation	Educate the MR as to what is potentially available in the EHR
	1.3 Send introduction email	Heuristics	Introduce the data analyst to the MR, Acquire additional clarification, and to set up the F2F meeting
	1.4 If needed, verify/review IRB	Heuristics; Institutional Policy	Comply with policies; Identify study concept and methods
	1.5 Identify potential index phenotype	Heuristics	Provides an initial attempt to define the patient cohort of interest
	1.6 Identify similar requests	Request tracking system	Provide knowledge to potential phenotypes used to map the index phenotype
	1.7 Consult with experienced Data analyst	Heuristics	Identify phenotypes not documented previously
	1.8 Establish complexity of the request	Heuristics	Provide an expectation as to the time needed to perform the needs negotiation
	1.9 Profile MR	Heuristics	Provide an expectation as to the time needed to perform the needs negotiation
	1.10 Verify one-time vs ongoing request	Heuristics	Establish the scope of the project
	1.11 Verify request was assigned to correct data analyst/team	Heuristics	Ensure the request is matched with the appropriate resources

* Institutional Policy – Institutional protocol for submitting an EHR data request; Internal Documentation – Institutional EHR data model; Heuristics – empirical knowledge gained from BQM practices; Request Tracking System – Internal work management application

Table 3

The hierarchical task model for the BQM face-to-face communication phase for the data analyst to negotiate with the medical researcher (MR) to arrive at an understanding of the MR's data need. Each task is associated with the knowledge required to complete the task and its expected outcome. Index Phenotype (IP); Associated Phenotype (AP)

Task	Activity	Step	Knowledge Required*	Expected Outcomes
2.0 Clarify project type			Heuristics	Establishes if the project is research or a business process project
3.0 Clarify project concept and methods			Study design and methodology	Introduces key medical condition and concepts and how they will be used in the research plan
4.0 Establish IP and AP(s)	4.1 Request case definition		Medical domain knowledge; Search engine	Provides a non-technical definition of the index/ associated phenotype
	4.2 If known, explain EHR data limitations		EHR data model; EHR data collection	Provides a nuanced discussion surrounding the limitations of EHR data used to map the index medical condition
	4.3 Define EHR data phenotype	4.3.1 Contextualize EHR Data Elements	Medical domain knowledge; EHR collection	Map the index/associated medical condition to EHR data elements
		4.3.2 Define temporal constraints among EHR data elements	Medical domain knowledge; EHR data collection	Define temporal relationship of EHR data elements used to represent the index/associated phenotype
		4.3.3 If needed, obtain keywords to retrieve relevant notes	Medical domain Knowledge; Heuristics	Sets of keywords that identify relevant clinical documents with key medical concepts
		4.3.4 Locate EHR data elements	EHR data model; GUI	Identify the database location of data elements needed
		4.3.5 If needed, request EHR database access rights	Heuristics	Gain access to database systems/tables that contain needed data elements
	4.4 Validate phenotyping algorithm	4.4.1 Provide formal phenotype definition	Expected cohort size; Gold standard	Assess the validity of the phenotype in identifying the index/associated medical condition(s)
		4.4.2 Clarify expected cohort size	Medical domain knowledge; Institutional practice patterns	Establishes a rudimentary surrogate marker for phenotype accuracy
		4.4.3 Perform query and present output to MR	Heuristics	Allows MR to inspect the output and verify accuracy of query results
		4.4.4 Request sample patient	Medical domain knowledge	Provides both a key list of data elements representing the index and associated phenotypes as well as an accuracy marker for the phenotype
	4.5 If exists, complete other IP/AP(s)		Heuristics	Moves the conversation toward other associated medical

Task	Activity	Step	Knowledge Required*	Expected Outcomes
				concepts within the cohort of patients
5.0	Clarify data output format		Data structures	Identifies the expected query output the MR would like
6.0	Confirm the data need explanation		Heuristics	Establishes an agreement between the MR and data analyst as to what the MR is requesting to minimize future disagreements regarding the output of data

* Heuristics – Empirical knowledge gained from BQM practices; Study design and methodology – Rationale differentiating various research approaches; Medical domain knowledge – Disease, treatments, and potential outcomes; Search engine – accessing new information; EHR data model – Information structure; EHR data collection – Clinical care documentation in the EHR; GUI – User facing application data entry; Expected cohort size – Count of patients with particular condition; Gold standard – Formal clinical case definition

Table 4

Alignment of Reference Interview (RI) and BQM Tasks

Concept	BQM Task Code	Quote	Contextualized
RI			
Determination of Subject	2.0	<i>Just because there is not a great standardization across researchers or you know not only on our campus but I think just nationally, I have to go to the investigator and say, hey this is how I am going to define [the patient cohort] and sometimes they agree and sometimes they disagree. Does that answer your questions?</i>	<i>To complete this project, I would like to establish a case definition, for the medical condition you are studying. In language that you are comfortable with, please describe the (index/associated) medical condition.</i>
Object and Motivation	3.0	<i>Typically, I will try and get a high level overview of what they are trying to accomplish. Sometimes people just jump into specific exclusion or inclusion criteria. Unless I can see the big picture, I might misunderstand what they are trying to accomplish and that won't become clear until a dataset is delivered.</i>	<i>I see that you are looking for diabetic and hypertensive patients taking hypertensive medications, do you mind describing your research questions and how you plan to answer it?</i>
Personal Characteristics of Inquirer	1.9	<i>A lot of the [data requesters] are really beginners or haven't [worked with EHR data before]. On the other side, you see senior personnel that just need help with figuring out what they can actually extract from the EHR in an automated way.</i>	<i>Based on the MR experience with both research and using EHR data for research, the data analyst assigns the MR as novice or expert. This helps establish an expectation for the time and effort needed for the needs negotiation.</i>
Relationship of inquiry with file organization	4.3.1	<i>We help [medical researchers] figure out what ICD-9, procedure, or medication codes used represent the medical condition they need so they can actually ask the hospital to retrieve the data.</i>	<i>Ok, so we are going to use two or more ICD-9 codes, 250. *, or if that does not exist elevated blood tests measuring abnormal glucose, fasting glucose, and A1C. I am clear on where to find the ICD-9 codes. Also, we can identify all the terminology used to label these particular blood tests, but what is your threshold for each of these tests to establish a patient as diabetic? Additionally, our medical entities dictionary does not have medication information, would you be willing to provide a list of all ARBs, both generic and name brand?</i>
Anticipated or acceptable answers	4.4.2	<i>One is to figure out how many patients we might have with the criteria that they've already given me to see if our numbers are in line with the medical researchers expectation.</i>	<i>Before I run the phenotype on the database, what is your expected cohort size, how many patients do you expect to meet this criterion at this institution?</i>
	5.0	<i>Look, there are multiple things happening at the same time. In addition to the data format, I also try to explain to them the data itself. If it is a snapshot, it is very easy, but when it's longitudinal data, it's more difficult.</i>	<i>When the query is complete, how should the output be formatted, for example, excel, SAS, STATA, text file, etc..</i>