

Toward a Generalized Model of Biomedical Query Mediation  
to Improve Electronic Health Record Data Retrieval

Gregory William Hruby

Submitted in partial fulfillment of the  
requirements for the degree  
of Doctor of Philosophy  
under the Executive Committee  
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2016

© 2016

Gregory William Hruby

All rights reserved

## **ABSTRACT**

Toward a Generalized Model of Biomedical Query Mediation

to Improve Electronic Health Record Data Retrieval

**Gregory William Hruby**

The electronic health record (EHR) is an invaluable resource for medical knowledge discovery. EHR data interrogation requires significant medical and technical knowledge. To access EHR data, medical researchers often rely on query analysts to translate their EHR information needs into EHR database queries. The conversation between the medical researcher and the query analyst is an information needs negotiation; I have named this process biomedical query mediation (BQM). There exists no BQM standard to guide medical researchers and query analysts to effectively bridge the communication gap between these medical and technical experts. The current practice of BQM likely varies among query analysts. This variation may contribute to the delivery of EHR data sets with varying degrees of accuracy. For example, a query analyst may return an EHR dataset that misrepresents the medical researcher's information need or another query analyst may return a different EHR dataset to the medical researcher for the same information need. The process used to formulate the medical researcher's information need and translate that need into an executable EHR database query may have severe downstream consequences affecting the reliability and quality of EHR datasets for medical research. This dissertation contributes early understandings of the BQM process and thereby improves the transparency and highlights the complexity of BQM by completing five studies: 1) survey the literature from other information intensive scientific disciplines to identify knowledge and methods potentially useful for BQM, 2) perform a review of existing tools and forms for assisting researchers in BQM, 3) perform a content analysis of the BQM process, 4) conduct a

cognitive task analysis to detail a generalized workflow, and 5) develop an enriched concept schema to capture comprehensive EHR data needs. This dissertation employs extensive qualitative methods using grounded theory, expert interviews, and cognitive task analysis to produce a deep understanding of BQM. Additionally, I contribute a promising concept class schema to represent medical researchers' EHR data needs to help standardize the BQM process.

# Table of Contents

List of Figures.....	viii
List of Tables.....	ix
Chapter 1. Biomedical Query Mediation for Electronic Health Record Data.....	1
1.1 Biomedical Query Mediation of EHR data for Research Use.....	3
1.2 What is an Information Need Negotiation.....	8
1.3 Secondary Use of EHR Data for Research is Limited Due to Access Constraints ....	9
1.4 Investigating BQM between Medical Researchers and Query Analysts.....	11
1.4.1 Summary of Approach.....	11
1.4.2 Aim I: BQM Gap Analysis.....	13
1.4.3 Aim II: The Biomedical Query Mediation Process.....	15
1.4.4 AIM III: The specification of EHR data needs through a conceptual schema ..	19
1.5 Contributions .....	22
1.5.1 Research results .....	23
1.6 Guide for the reader .....	24
Chapter 2. Facilitating Medical Researchers’ Interrogation of the Electronic Health Record: Ideas from outside of Biomedical Informatics.....	26
2.1 The Tradition of Clinical Data Reuse for Medical Research .....	26
2.2 Methods .....	27
2.2.1 Development of a Conceptual Framework for Interactive Data Retrieval .....	27

2.2.2	Literature Search Methods.....	29
2.3	Results.....	33
2.3.1	How the Data Source Facilitates User Access.....	33
2.3.2	USER.....	34
2.3.3	CHANNEL.....	36
2.4	Discussion.....	41
2.5	Conclusion.....	44
Chapter 3.	Initiating Electronic Health Record Data Requests.....	45
3.1	Introduction.....	45
3.2	Methods.....	48
3.2.1	Collection of data request forms.....	48
3.2.2	Development of a codebook.....	49
3.2.3	Form annotation by two annotators.....	51
3.2.4	Content analysis of the ten forms.....	52
3.3	Results.....	53
3.4	Discussion.....	58
3.5	Conclusion.....	62
Chapter 4.	Characterization of the Biomedical Query Mediation Process.....	63
4.1	Introduction.....	63
4.2	Data and Methods.....	65

4.2.1	Data.....	65
4.2.2	Annotation Schema Development.....	66
4.2.3	Dialogue Act Annotation.....	66
4.2.4	Data Analysis.....	67
4.3	Results.....	67
4.3.1	A Dialogue Act Classification Schema for Mediate Query Conversations.....	67
4.3.2	Temporal Distribution of Dialogue Act Classes.....	69
4.3.3	A closer look on the Discussion of the Clinical Process.....	71
4.3.4	Temporal Flow of Study Design and Research Workflow Dialogue Acts.....	71
4.4	Discussion.....	72
4.5	Limitations.....	74
4.6	Conclusion.....	74
Chapter 5.	Understanding and Generalizing the Biomedical Query Mediation Process....	75
5.1	Introduction.....	75
5.2	Methods.....	76
5.2.1	Participants.....	77
5.2.2	Semi-Structured Interview.....	79
5.2.3	Transcript Annotation and Analysis.....	80
5.2.4	Evaluation.....	81
5.3	Results.....	82

5.3.1	BQM hierarchical task model and process workflow.....	82
5.3.2	Face and content validation .....	87
5.3.3	Comparison to the Reference Interview .....	90
5.4	Discussion.....	92
5.4.1	Implications .....	92
5.4.2	Expected findings .....	93
5.4.3	Unexpected Findings .....	94
5.4.4	Limitations.....	95
5.5	Conclusion .....	95
Chapter 6.	Data-driven Concept Schema for Defining Clinical Research Data Needs.....	96
6.1	Introduction.....	96
6.2	Methods .....	98
6.2.1	Data Sources and Characteristics.....	99
6.2.2	Data Sampling .....	100
6.2.3	Dataset Annotation and Analysis.....	101
6.2.4	Evaluation .....	103
6.3	Results.....	105
6.3.1	Data-Enriched Schema .....	105
6.3.2	Evaluation .....	106
6.3.3	Participant-Enriched Schema.....	115



6.4	Discussion.....	118
6.4.1	Implications of Results .....	118
6.4.2	Intended Use Case .....	119
6.4.3	Limitations.....	119
6.5	Conclusion .....	120
Chapter 7.	Summary and Conclusions.....	121
7.1	The gestalt view of biomedical query mediation.....	121
7.2	Contributions to Clinical Research Informatics .....	123
7.3	Limitations.....	125
7.3.1	Limitations of using single use case scenario to understand BQM.....	126
7.3.2	Limitations of using a single stakeholder analysis to understand BQM .....	126
7.4	Future Work.....	127
7.4.1	Supporting cognition for BQM.....	128
7.4.2	BQM as a data source for EHR Phenotyping .....	129
7.5	Conclusions.....	129
References		131
Appendix A		145
1.1	Semi-structured interview .....	145
a.	Introduction.....	145
b.	Phase One.....	145

c.	Phase Two .....	145
d.	Phase Three .....	148
e.	Figures and Tables for the biomedical query mediation task .....	149
2.1	Query Analyst Workflows used during Member checking .....	152
a.	Definitions.....	152
b.	Validation Questions.....	152
c.	Interview ID_1 .....	153
d.	Interview ID_2 .....	154
e.	Interview ID_3 .....	155
f.	Interview ID_4 .....	156
g.	Interview ID_5 .....	157
h.	Interview ID_6 .....	158
i.	Interview ID_7 .....	159
j.	Interview ID_8 .....	160
k.	Interview ID_9 .....	161
l.	Interview ID_10 .....	162
m.	Interview ID_11 .....	163
	Appendix B	164
1.1	Adaptations made to the original Carpenter Framework using each dataset..	164
2.1	Semi-Structured Interview Material.....	166

a.	Introduction.....	166
b.	Concept Generation and Mapping .....	166
c.	Modeling Structure .....	167
3.1	Schema Class Definitions and Examples.....	169

## List of Figures

Figure 1-1. The Ammenwerth model .....	4
Figure 1-2. The average impact score.....	6
Figure 1-3. The BQM process. ....	7
Figure 1-4. Dissertation flowchart.....	12
Figure 1-5. Conceptual framework for interactive data retrieval. ....	14
Figure 1-6. Theme River. Content of BQM .....	17
Figure 1-7 Generalizable biomedical query mediation process workflow.....	18
Figure 1-8. The data enriched schema.....	21
Figure 2-1. A conceptual framework for interactive data retrieval. ....	29
Figure 2-2. The search strings and article selection flowchart .....	31
Figure 3-1. Data request form content analysis workflow .....	48
Figure 4-1. Broad BQM content workflow overview .....	65
Figure 4-2 Theme river. Temporal Distribution of Dialogue Acts.....	70
Figure 4-3. Theme River. Discussion of the Clinical Process.....	71
Figure 4-4. Theme River. Discussion about Study Design and Workflow Issues .....	72
Figure 5-1. High-level overview of the research process .....	77
Figure 5-2. Generalizable BQM task model.....	87
Figure 5-3. Content Validity Ratio (CVR) results.....	89
Figure 6-1. Research Design.....	99
Figure 6-2. The Venn diagram.....	105
Figure 6-3. The data enriched schema.....	106
Figure 6-4. Participant enriched schema. ....	116

## List of Tables

Table 1-1. The knowledge gaps.....	15
Table 2-1. The distribution of relevant topics in two bodies of literature .....	32
Table 2-2. EHR data access barriers and solutions.....	33
Table 2-3. A Taxonomy of Clarification Questions .....	38
Table 2-4. The knowledge gaps.....	42
Table 3-1. Form elements .....	50
Table 3-2. Summary of the coding analysis .....	55
Table 3-3. Data elements .....	56
Table 3-4. Noteworthy atypical form elements .....	57
Table 4-1. Example Dialogue Acts.....	66
Table 4-2. The Classification schema for Dialogue Acts in Query Mediation .....	68
Table 5-1. Study Participant Characteristics .....	78
Table 5-2. BQM preparation phase tasks/activities.....	83
Table 5-3. Face-to-face Task/Activity/Step.....	84
Table 5-4. Alignment of Reference Interview (RI) and BQM Tasks .....	91
Table 6-1. Datasets used .....	100
Table 6-2. Evaluation metrics.....	103
Table 6-3. Evaluator characteristics .....	104
Table 6-4. Class Preservation.....	107
Table 6-5. Participant breakdown for generalizability and class coverage .....	108
Table 6-6. The subjective metrics of understandability and structure.....	111
Table 6-7. Notable quotes.....	114

## **Acknowledgments**

This dissertation represents a significant investment of not only my own time but many members of my professional and personal communities. I owe gratitude toward my primary advisor. Dr. Chunhua Weng provided a rigorous mentorship that tested my ideals, challenged my methodologies, and addressed my deficiencies as a researcher. Her guidance is a direct reflection of the quality of this work and I can confidently leave this program with a foundation that promotes the paragon of academic rigor.

In addition, Drs David Hanuer, Vimla Patel, James Cimnio, Konstantina Matsoukas, and Eneida Mendonça provided invaluable insight from their collective wealth of research expertise improving the scholarship of this dissertation.

The details of this work would not have been possible without the research aid from my peers. In particular, Luke Rasmussen dedicated a significant amount of personal time to support my work. His generosity is something I hope to repay in kind in the future. Finally, my fellow students and post docs from the Department of Biomedical Informatics, Julia Hoxha, Praveen Ravichandran, Mary Boland, Dan Fort, Fernanda Polubriaginof, Rimma Pivovarov, William Brown III, and Jenny Prey, have provided me with emotional and academic support throughout these past five years. I am proud to call them my friends and colleagues.

Finally, I would like to thank my family members for their undying support of my academic goals. My brother Tim has grounded me throughout this effort with his admirable love of family. My partner in crime, Valentina, has become my best friend and checks me with a healthy dose of skepticism for my obstinate ways (no small task). Most important my parents, Frank and Cindy,

have provided me with countless opportunities to explore whom I am and what I wish to accomplish in life. I am extremely fortunate to have their unwavering support and love.

## **Chapter 1. Biomedical Query Mediation for Electronic Health**

### **Record Data**

Biomedical query mediation (BQM) describes the information needs negotiation process that medical researchers go through when defining a medical information need and translating that need into an executable Electronic Health Record (EHR) database query. BQM involves both an information seeking step and an information retrieval step [1, 2]. These components contain barriers to the successful resolution of the user's information need. For example, in information seeking, the information need is framed by the situation of the user, users of varying domain and technical expertise employ different information seeking strategies, and users tend to seek information that is easily accessed [1, 3, 4]. In information retrieval, obtaining relevant information resources are dependent on the formulated query. If the query was ill-conceived or is a vague representation of the information need, then the obtained resources, regardless of their relevance to the query, are insufficient to resolve the user's information need [5-8]. In the context of EHR data retrieval, the breadth and depth of EHR data available amplifies these barriers. As such, EHR data query formulation or BQM is a critical process to provide EHR data to medical researchers. However, we know little about this process.

My primary research goal is to understand and model BQM with the hope of improving the transparency and highlighting complexities of EHR data query formulation through a generalized BQM task model. Understanding the cognition involved for BQM lays the ground work for future cognitive computing applications facilitating EHR query formulation and EHR data retrieval. In this dissertation, query formulation refers to the process of eliciting a clear, unambiguous definition of medical researchers' information need. Query formulation is an



iterative process that involves feedback from an information intermediary, who can be an automated agent or a human query analyst. Finally, a generalized BQM workflow may provide a standard process that is independent of the information intermediary and optimize medical researchers' information seeking.

My dissertation consists of five parts. First, I performed a literature review and identified knowledge from the information science domain applicable to BQM. The manner in which information seekers specify their information need has severe downstream consequences on information retrieval. Second, I performed a qualitative analysis on institutional data request forms to understand how these forms enable medical researchers to specify their information need; often minimal content within the forms is dedicated to aiding medical researchers specify their information need. Third, to understand the BQM conversation space I performed a content analysis on the exchanges between medical researchers and query analysts. I produced visualizations of the content exchanged between medical researchers and query analysts. Fourth, the literature review identified information intermediaries as key players in the query formation process. I executed a cognitive task analysis on BQM to understand the tasks query analyst use to elicit information from the medical researcher. I identified multiple BQM processes and created a generalized workflow model representing multiple institutions' BQM. Finally, I enriched a concept class schema identified in my literature review with the goal of improving the data request form to aid the specification of an information need. I produced a comprehensive representation of EHR data needs to aid in the specification and elicitation of a medical researcher's information need. To provide additional context for the importance of BQM, I will elaborate on the roles of information mediators within an EHR data intensive research setting;

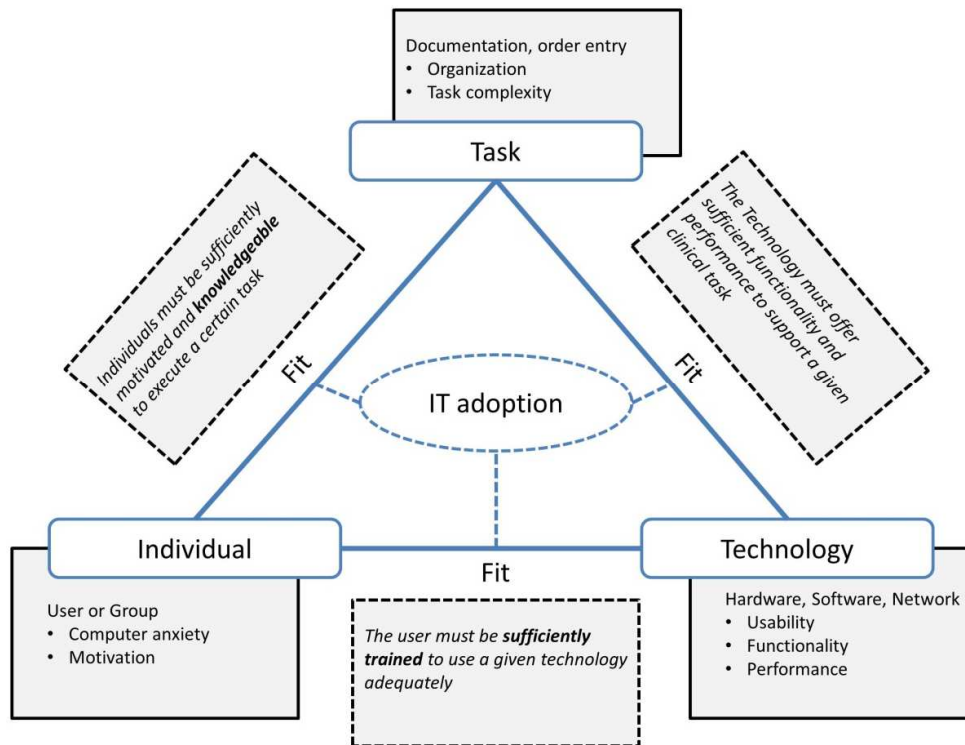
and present the significance of medical researchers leveraging EHR data for medical knowledge discovery in section 1.1.

### ***1.1 Biomedical Query Mediation of EHR data for Research Use***

BQM is an interactive data retrieval process that involves the exchange of the query analyst's system knowledge and medical researcher's domain knowledge. This exchange of information is a key component that facilitates query formulation. For example, the information exchange process takes abstract, vague medical concepts (e.g., Type 2 diabetes) and translates them into concrete data elements (e.g., ICD-9 codes and laboratory tests associated with Type 2 diabetes, as well as textual descriptions of Type 2 diabetes in assorted clinical notes) [9]. However, the literature provides little insight into this often opaque process [10]. Two mechanisms exist for medical researchers engaged in BQM, Self-service query tools and query analysts (intermediaries).

Self-service query tools have promised to provide comprehensive EHR data access to medical researchers[11]. Self-service query tools, such as Amalga [12], i2b2 [13], SHRINE [14, 15], VISAGE [16], and STRIDE [17], and Atlas [18], allow medical researchers to navigate the EHR data space autonomously [19]. Designers of Self-service query tools aimed to reduce the use of valuable human resources by allowing medical researchers to perform query formulation and translation. However, due to limited evaluation of these tools, their success is unclear. One evaluation of the i2b2 self-service query tool suggested the tool is useful for cohort selection related to common data requests, especially estimating cohort sizes, but not suitable for resolving complex queries with multiple constraints and complex temporal relationships between concepts [20]. This work complements other observations, the majority of users present with tasks related to cohort size estimation and can be facilitated by the i2b2 Self-service query tool application

[20] [21]. We can extend these results into the context of the Ammenwerth Fit framework for representing relationships among Individuals, Task, and Technology (**Figure 1-1**)[22].



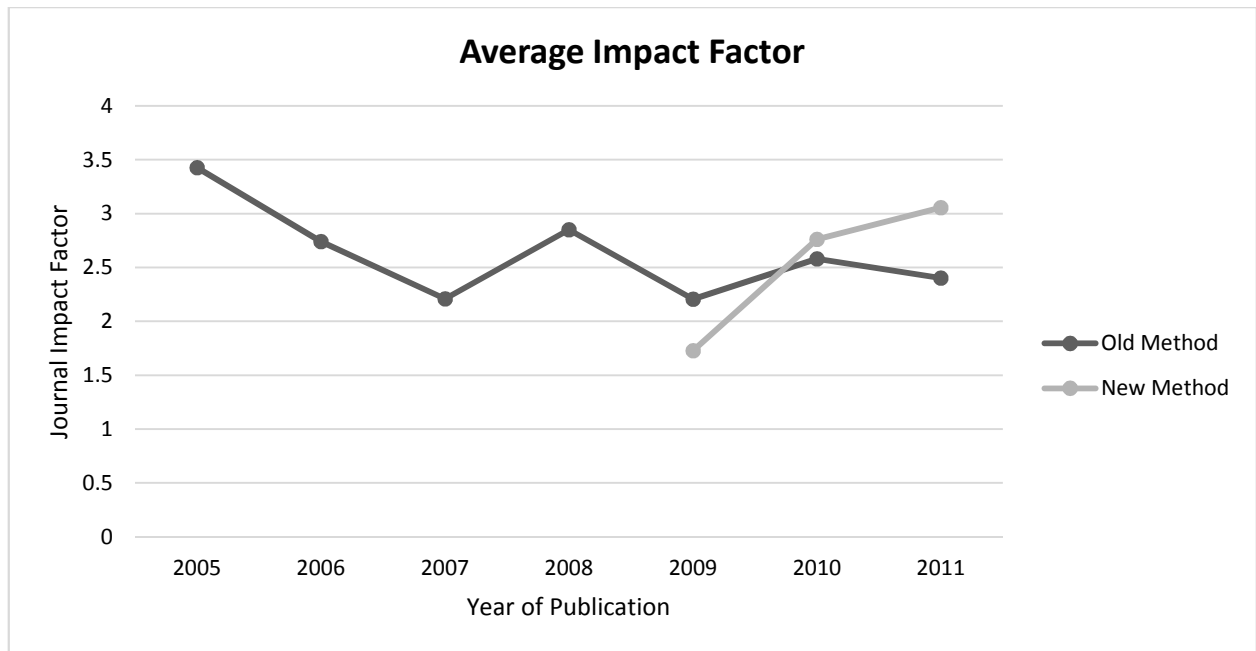
**Figure 1-1. The Ammenwerth model provides a standard to predict and measure the success of information technologies’ adoption.**

This model predicts the adoption of technology is dependent on a triad of relationships among users, tasks, and technology. Many institutions have implemented Self-service query tools to resolve the majority of tasks from their user base, simple cohort estimations. For these types of requests, Self-service query tools support the task of the user thus satisfying a key relationship of the Ammenwerth model. The other two relationships are an institutional dependent and a socio-technical component. For example, did the manager assign the task to the appropriate individual and does the institution provide adequate training for users of the technology. However, current self-service query tool designs do not provide the cognitive support features needed for medical researchers to formulate and translate complex EHR data

needs. To deal with complex requests, dedicated query analysts act as an information mediator aiding the researcher in formulating their information need and extracting an EHR dataset commensurate with that need.

The goal of my dissertation is to study the complex interaction between the medical researcher and query analyst to resolve a complex EHR data need. In previous work, I detailed the effect of using EHR data for secondary research use[23]. Through this work, I will show supporting evidence on the potential of EHR data for research purposes, and an initial understanding of how BQM unfolds in an information intensive environment.

I completed a retrospective analysis of a centralized research data repository's impact on the Columbia University Medical Center, Department of Urology's research capacity during a pre-centralized research data repository period (2005-2008) and a post-centralized research data repository period (2009-2011)[23]. We implemented a new workflow model and the centralized research data repository. The new system allowed multiple research assistants simultaneous access to EHR data and permitted overlapping, multiple projects using the same system. As such, the average time for project completion dropped from 12 to 6 months. The department's average annual retrospective study publication increased from 11.5 to 25.6 publications. At the same time, the average journal impact score rose from 1.7 to 3.1. **Figure 1-2** illustrates the publication quality as demonstrated by average journal impact score over the study period from publications resulting from the old and new systems. In summation, there is evidence to suggest the centralized research data repository led to an increase in quality and quantity of the department's publications. My case study suggests the impact of improved access to EHR data on research is positive and the process by which medical researchers access these data should be optimized to further facilitate medical research.



**Figure 1-2. The average impact score for retrospective studies over the study period. We grouped publications resulting from the two periods; the solid and dashed lines indicate publications resulting from the old and new system, respectively. Notice the declining trend of publication quality from the pre-centralized research data repository period and the upward trend during the centralized research data repository period.**

This work also provides the earliest understanding for BQM between a medical research and a query analyst. As predicted, BQM is a highly complex and iterative communication process. BQM involves establishing a clear definition of the EHR data need and then translating that need into an executable database query. **Figure 1-3** depicts this process in detail.

## Key Players

Query Analyst (QA)  
Medical Researcher (MR)



## Resources

Electronic Health Record (EHR)  
Central Research Data Repository (CRDR)  
Transformed Content (TC)



## Cognitive Tasks

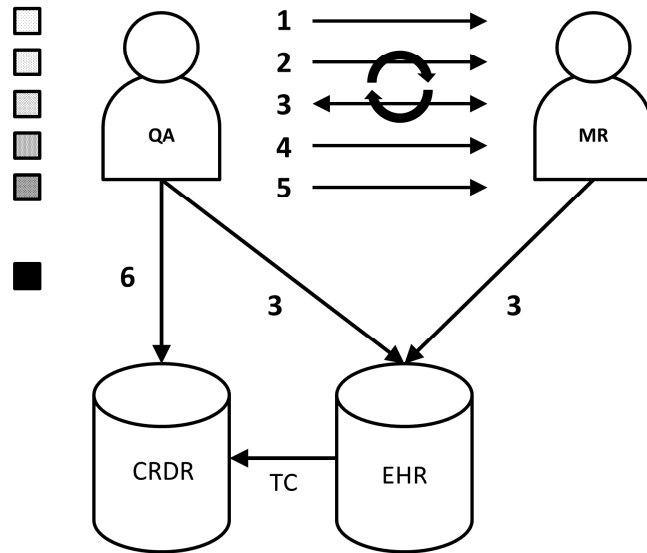
- 1 Define the research question
- 2 Define the clinical process
- 3 Locate data elements in the EHR
- 4 Establish patient characteristics not mentioned
- 5 Stop mediation
- 6 EHR database query formulation

## Actions

Unidirectional   
Bidirectional 

## Query Artifact

   
Empty Definition      Complete Definition



**Figure 1-3. The BQM process. This graph displays an abstract overview of the data needs negotiation process, including key players and resources that enable the process. In addition, the unidirectional arrows indicate which key player is directing a particular task. We posit the query analyst mediated the data needs negotiation.**

The core objective of BQM is to assist the medical researcher in providing a clear articulation of complex EHR data needs. In doing so, the query analyst minimizes the need to make assumptions, allowing for a more accurate executable database query. To improve the reliability of this process, an in-depth BQM model of the tasks used and the knowledge needed to complete those tasks is vital. An interactive BQM model may allow for improved support of medical researcher's data needs [1, 24, 25]. To bridge the knowledge gap between the query analyst and medical researcher, I propose an interdisciplinary approach, extending work performed in multiple fields such as information science, computer science, and linguistics to

define the BQM process. Next, I will examine the broad context of an information need negotiation process.

### ***1.2 What is an Information Need Negotiation***

An information need negotiation defines the process two or more individuals engage in to reach an agreement on the information need. A successful information need negotiation is dependent on the effectiveness of the iterative negotiation between the information seeker and the intermediary [1, 26, 27]. To complicate things further, information seekers initiate this process with a vague description of their information need. Several information seeking models, ASK [2] and Berrypicking [28], provide descriptive examples. The ASK hypothesis posits that information seekers are typically unable to construct a precise definition of their information need, and as such, are limited to vague descriptions or non-specific information requests. Similarly, the Berrypicking model describes the initial description of the information need as vague and details how the information seeker's ability to articulate an information need improves as relevant information becomes available to them. These models suggest an intermediary could play a critical role in aiding the information seekers articulation of what they need. The paragon example is between librarians and library patrons, the reference interview. The reference interview is a skilled needs negotiation between a librarian and an information seeker to convert a vague information need into an unambiguous query by iteratively eliciting tacit user needs, verifying implied assumptions and improving the specificity of information queries [26]. The information seeker and librarian leverage their expertise to aid in the needs negotiation. The librarian is a system expert, understanding how information is stored and accessed. The information seeker is the domain expert, identifying the terms to describe their need. The

information need negotiation process allows the information seeker to formulate a query using their terms to describe relevant concepts for that information need.

Query formulation places a significant strain on a user's cognitive resources. For example, users often experience this challenge in biomedical literature retrieval systems, where the technical expertise needed levies a large cognitive demand on the user [29]. EHR query formulation by medical researchers is a knowledge-intensive task. When medical researchers use Self-service query tools, the researcher's lack of EHR system knowledge and database architecture places a significant demand on their cognitive resources impacting the ability to extract EHR data elements corresponding to the information need [30, 31]. This aspect may explain why Self-service query tools are not adequate tools for the complex information needs of medical researchers [20]. Similarly, query analysts encounter difficulty during the query formulation process not due to their lack of technical knowledge, but their lack of medical domain knowledge. Query analysts spend cognitive resources comprehending the medical domain terminology expressed by the medical researcher, which contributes to the difficulties establishing a precise and accurate definition of the EHR data need.

### ***1.3 Secondary Use of EHR Data for Research is Limited Due to Access Constraints***

With nearly half of US hospitals now with one or more EHRs in place [32], a wealth of electronic data is available and carries with it implicit expectations that the secondary use of this data will facilitate comparative effectiveness research and public health initiatives to improve the quality of research and establish new knowledge to guide healthcare policy [33-40]. Clinical and translational research is a growing priority of the United States National Institutes of Health (NIH). To encourage greater advancements in this area the NIH has supported over 60 research



institutions through the Clinical and Translational Science Awards [41]. Additionally, several major consortiums exist to support the secondary use of EHR data; the Observational Health Data Sciences and Informatics (OHDSI) and the Patient Centered Outcomes Research Institute (PCORI) have been established to leverage a large network of electronic patient data for new knowledge generation for all aspects of healthcare[18, 42]. This concomitant investment in both the research enterprise and in health information systems that capture data electronically has presented unprecedented opportunities for advancing clinical and translational science, and is a necessary precursor for building the foundations of a broad-scale “learning health system” [43, 44]. Academic medical institutions have prioritized providing access to EHR data for medical researchers [45, 46].

However, providing EHR data access contains latent barriers. For example, early data access solutions provided minimal cognitive support for medical researchers interacting with EHR data and underestimated the resources necessary to resolve the medical researchers’ complex EHR data needs[47, 48]. Intermediaries aid medical researchers by properly formulating their information need and navigating the complex data structures and representations powering the EHR. The query formulation step establishes an information need framework with which intermediaries can aid in translating the query into an EHR data representation or more commonly known as an EHR phenotyping. Query formulation needs to be precise as inaccuracies, and/or inconsistencies may have severe downstream consequences rendering the resulting datasets unreliable. Understanding this process in detail and building a generalized model will in effect create an expectation for medical researchers that the information seeking process for EHR data from institution to institution is similar. A consistent process can improve the results obtained, and the medical researcher’s confidence the data is commensurate with their

need. To this end, my dissertation will study EHR data access facilitated by query analysts. Considering the potential for the use of BQM to mitigate known and latent barriers to EHR data access for research use, I will document an in depth understanding of BQM. I have conducted my research in the context of the medical researcher and the query analyst.

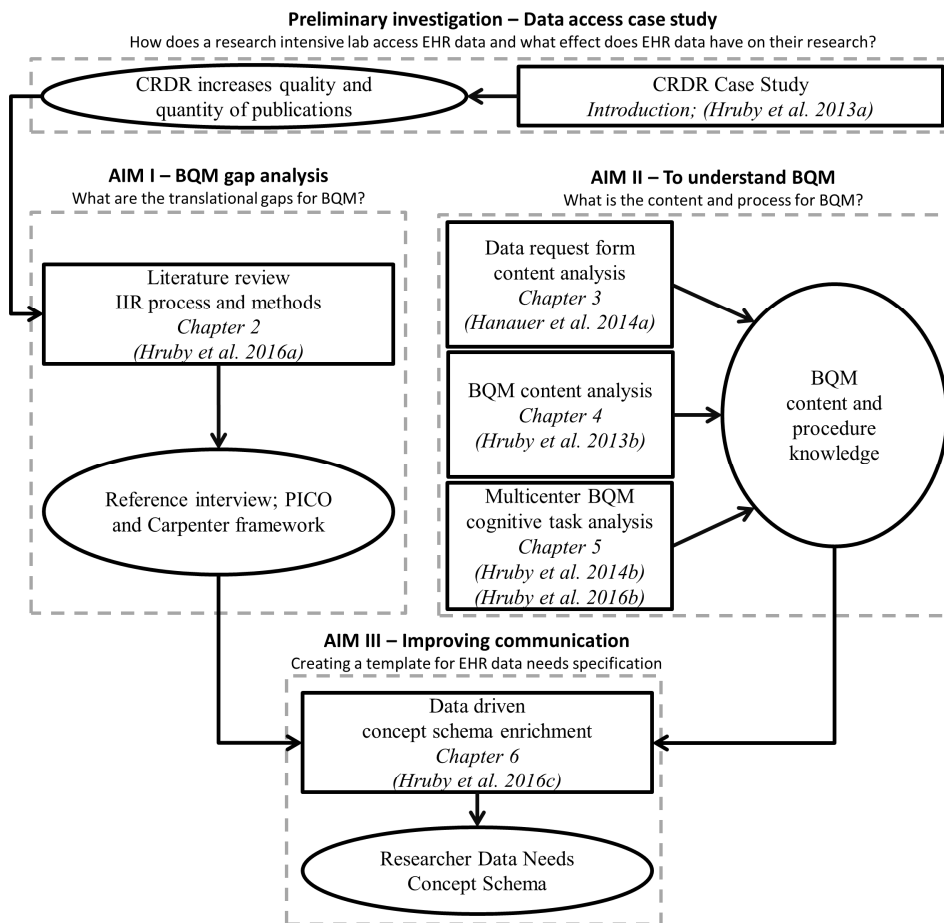
## ***1.4 Investigating BQM between Medical Researchers and Query Analysts***

### *1.4.1 Summary of Approach*

This dissertation contributes a full understanding of BQM. Furthermore, the dissertation implements novel methods to produce knowledge that may optimize BQM by providing a framework to increase specificity of the medical researcher's information need. To accomplish this, I proposed a mixed-methods approach, leveraging both a data-driven analysis and a cognitive task analysis to develop a deep understanding of the processes currently used to facilitate BQM.

**Figure 1-4** provides a graphical representation of the approach I used for each AIM of the dissertation. In AIM I, I conducted a literature review to identify methods and results that offer optimization of key BQM components. The major contribution from AIM I proposed the development of a dedicated query template to aid in the specification of EHR data needs. For AIM II, I produced two deliverables. First, I established a generalizable BQM task model. Second, I conducted an in-depth content analysis of EHR data request forms and discovered that the forms provide minimal support to the medical researcher for the specification of their EHR data needs. In AIM III, I used the major contributions from AIM I and II to inform AIM III's design and goal. I produced a conceptual schema of researcher data needs for eliciting and

specifying a medical researcher’s data need. In particular, AIM III bridges the conceptual knowledge identified in AIM I and the physical gap identified in AIM II.



**Figure 1-4. Dissertation flowchart. Within each AIM, the boxes denote a method and an oval represents the knowledge generated from the method. The corresponding chapters supporting each AIM and the research publications described in section 1.5.1 are noted in italics.**

#### *1.4.2 Aim I: BQM Gap Analysis*

Objective: Identify key knowledge and translational gaps in the facilitation of efficient data access in life sciences. Survey state-of-the-art approaches and key methodological considerations from the biomedical informatics and information science literature.

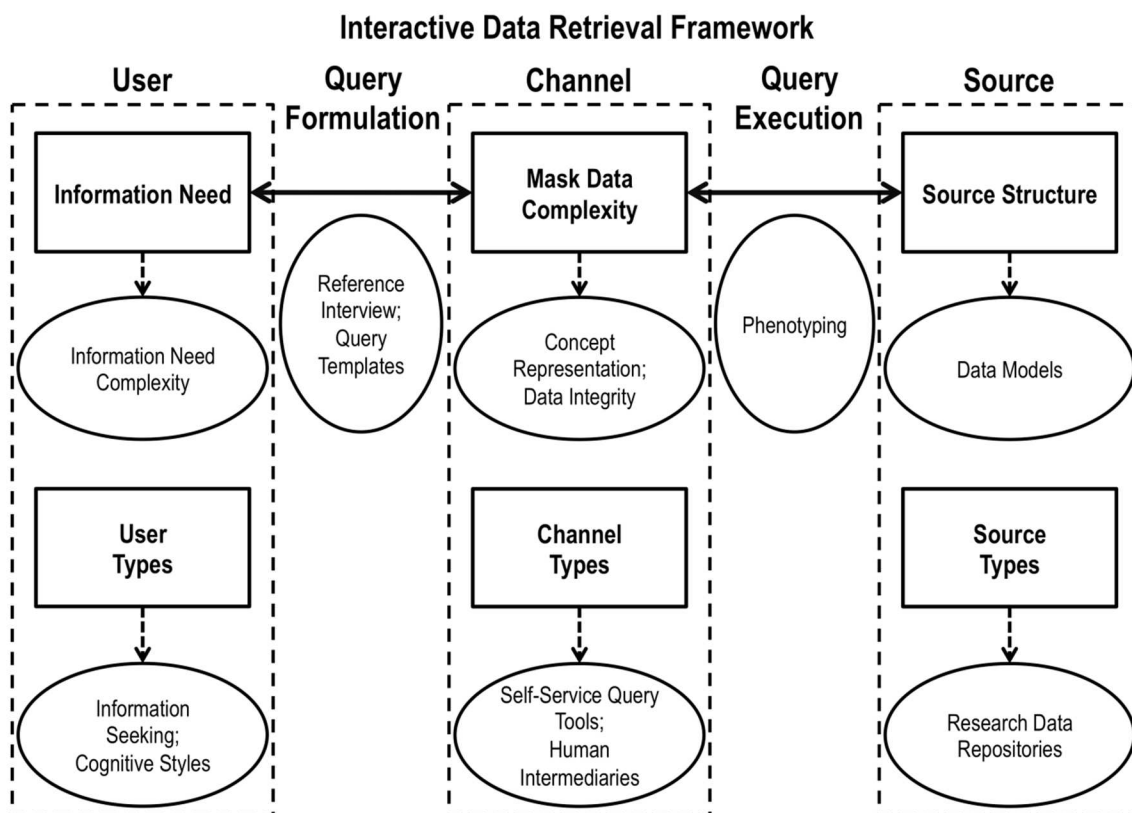
Hypothesis: Many of the barriers inhibiting EHR data access by medical researchers are not new, and the information science literature can identify translational gaps existing in the biomedical literature that can benefit interactive EHR data retrieval.

#### Research Questions:

- What are the characteristics of an information need and how do they affect the medical researcher's search process?
- What models exist to describe the users search process?
- What is known about the query formulation process in other informational intensive domains?

#### Primary Findings:

Information retrieval addresses information needs using a sequence of tasks [8]. Bystrom and Jarvelin modeled three entities for interactive information retrieval: user, channel, and source [49]. I extended this framework by elaborating on the channel entity, including the sub-entities of query formulation and query execution, respectively, as shown in **Figure 1-5**. Since this AIM targets end user augmentation and its process, user requirements, and the gaps in existing technologies for query formulation, I briefly summarize the literature on source and query execution but elaborate on user modeling and query formulation.



**Figure 1-5. Conceptual framework for interactive data retrieval. The entities, Query Formulation and Query Execution were added to better describe the interaction between the User, Channel, and Source entities.**

Relatively little is published on the topics of the biomedical researcher’s cognitive styles and information seeking strategies. Table 4 lists key knowledge gaps and potential recommendations to bridge those gaps. Understanding how biomedical researcher access EHR data and what barriers they encounter is needed. Query templates have the potential to standardize the expression of the medical researchers information need; a query template for expressing EHR data needs should be developed.

**Table 1-1. The knowledge gaps and recommendations for advancing EHR data interrogation**

Aspects	Knowledge Gaps	Recommendations
<b>User</b>	<ul style="list-style-type: none"> <li>• Lack of measure of information need complexity</li> <li>• Lack of knowledge of how the cognitive styles of medical researchers affect the information seeking process</li> </ul>	<ul style="list-style-type: none"> <li>• Develop metrics for measuring EHR information need complexities</li> <li>• Conduct qualitative studies of the information seeking processes of various medical researchers</li> </ul>
<b>Channel – Query Formulation</b>	<ul style="list-style-type: none"> <li>• No formalized structure for the medical researcher to express their information need</li> <li>• The need-negotiation process performed by data intermediaries is poorly understood</li> </ul>	<ul style="list-style-type: none"> <li>• Investigate other formal structures used for document retrieval</li> <li>• Leverage methods used to understand and improve the librarian reference interview</li> <li>• Support reference interview for EHR data interrogation</li> </ul>

I identified three promising concepts to inform the design and improvement of BQM. First, both information science and biomedical informatics have established the important role semantics play for optimal information retrieval; query templates, such as the Patient Intervention Control Outcome and Carpenter framework, offer some promising leads. Second, the complexity of an information need shapes search tactics used by medical researchers. Third, the established reference interview has effectively helped librarians to clarify user needs in their setting. EHR query analysts perform a role similar to librarians but lack a guideline on how to perform the reference interview.

*1.4.3 Aim II: The Biomedical Query Mediation Process*

Objective: Investigate the content exchanged between the query analyst and medical researcher during the BQM process as well as the tasks used to arrive at a successful transfer of the data need from the medical researcher to the query analyst. Determine the common tasks of the medical researcher and query analyst, the knowledge used for each task, and the barriers experienced by both during the BQM process.

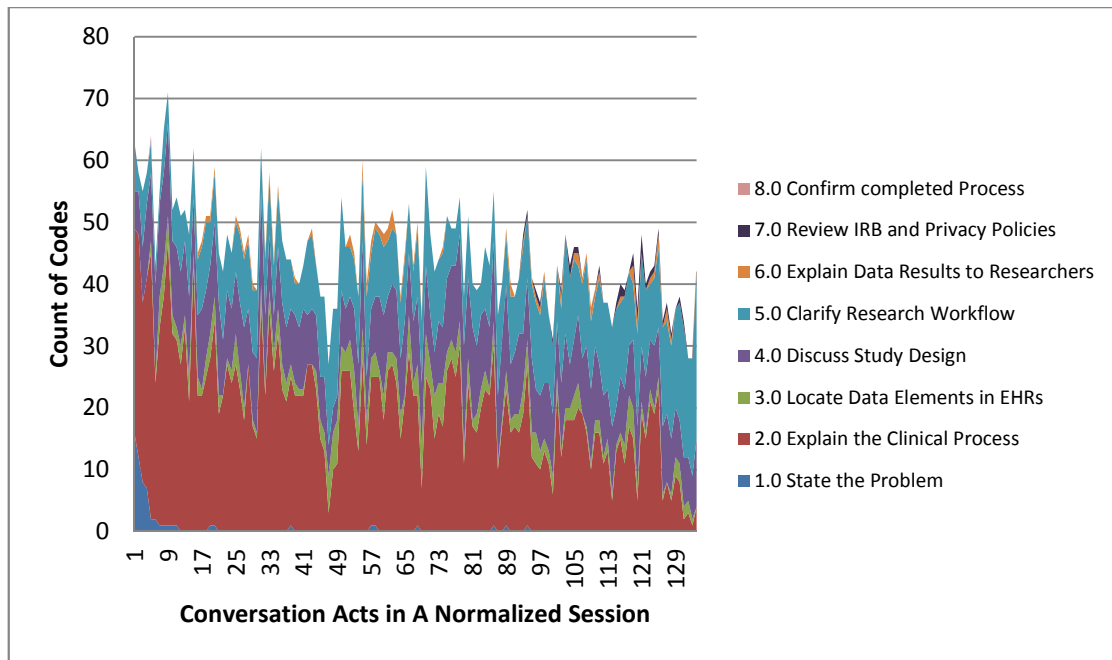
Hypothesis: BQM contains a set of common tasks used to elicit and translate information breadth (coverage) and depth (complexity) into executable EHR database queries.

Research Questions:

- What type of content is exchanged between the query analyst and medical researcher during BQM?
- What are the common tasks used to elicit and transfer information between the BQM participants?
- What BQM tasks can be documented using the cognitive task analysis?

Primary Findings:

My initial investigation of the BQM space describes the content exchanged between a medical researcher and a query analyst. **Figure 1-6** visualizes the content of several BQM conversations. The majority of the content surrounds a discussion of the clinical process related to the EHR data need, highlighting the breadth and depth of information exchanged in BQM. Furthermore, this provides additional evidence implying BQM is a complex information exchange process.

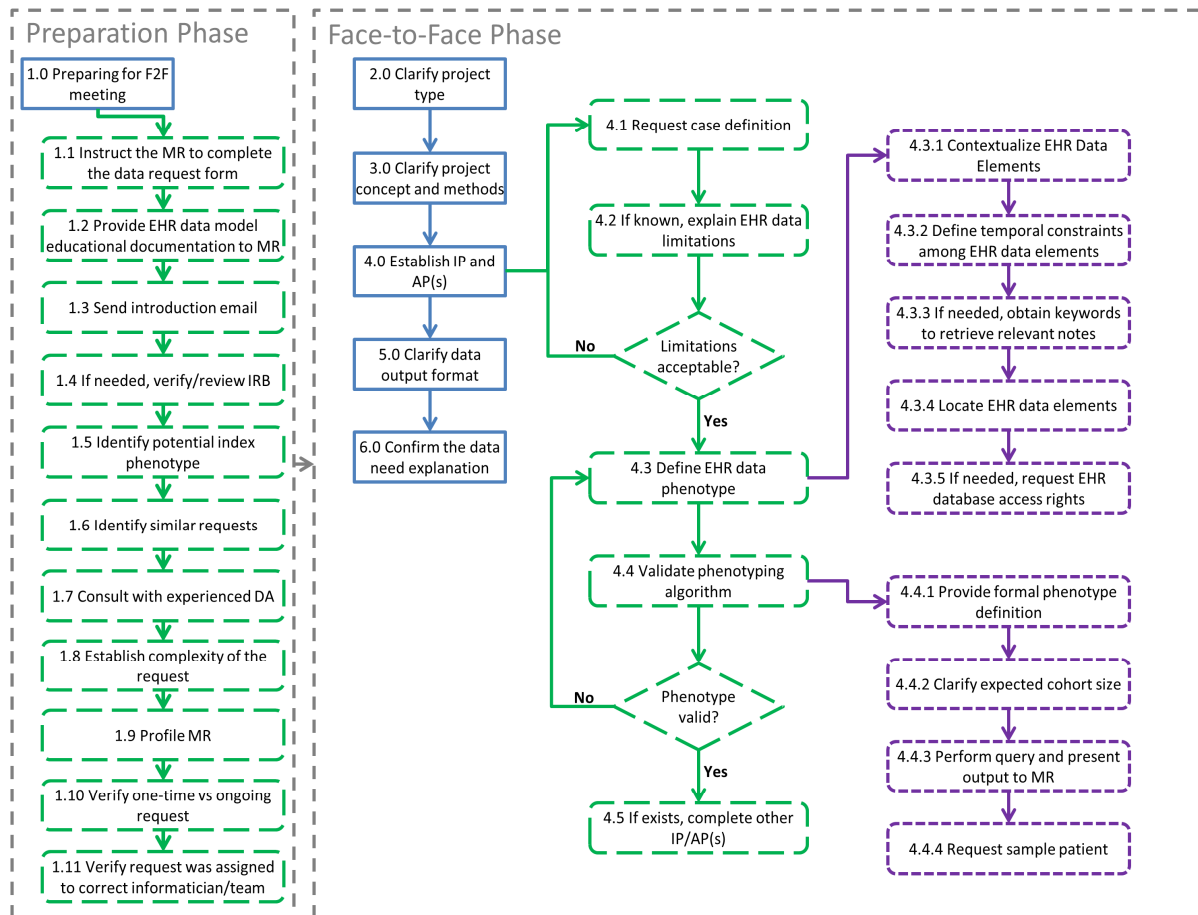


**Figure 1-6. Theme River. Content of BQM expressed over the course of the conversation. The majority of the content discusses the clinical process being studied. The y-axis represents the count of dialogue acts for a particular code, and the x-axis represent the linear progression of the conversation, starting with the first dialogue act and ending with the last.**

Next, I produced a generalizable, hierarchical task model for BQM. The model consists of two parts, a preparation phase, and a face-to-face consultation. **Figure 1-7** illustrates the components of these two parts. BQM is similar to the reference interview used by librarians. Taylor describes five filters that an information need passes through during a reference interview: determination of the subject, objective and motivation, personal characteristics of the inquirer, relationship of inquiry description to the file organization, and anticipated or acceptable answers[26]. Tasks within my model compliment these filters, for example, establishing the index phenotype, clarifying project concept and methods, profile the medical researcher, contextualizing the EHR data elements, and clarifying expected cohort size, respectively. That the two models are complementary is an encouraging and promising lead for the improvement of



this process by standardizing the BQM from institution to institution with a method known to be successful.



**Figure 1-7 Generalizable biomedical query mediation process workflow. This workflow represents the tasks, activities, and steps needed to elicit a clear information need from a researcher and then define that need with the corresponding EHR data elements. IRB – Institutional Review Board; IP – Index Phenotype; AP – Associated Phenotype; QA – Query Analyst; MR – Medical Researcher;**

I found the task, “If known, explain EHR data limitations”, of particular interest. It was unexpected that query analysts engage in an EHR data limitation discussion with the medical researcher. From the query analyst’s perspective, medical researchers have limited knowledge of the EHR and the medical researcher may confound the actual meaning of an EHR data element; the medical researchers may not know what the data element can tell them and how they can use

it appropriately for defining the medical researcher's EHR data need. Whether or not the EHR data limitation is known, the query analyst will dedicate time to either explaining the limitation or learning a potential new limitation of EHR data. This critical task of establishing the quality of the data is an active area of research [50].

The generalizable BQM task model provide a resource to both query analysts and managers. This resource can serve as a reminder to expert query analysts and a guideline to novice query analysts. Additionally, managers can use this knowledge to gauge the depth of work query analysts perform to resolve the EHR data needs of medical researchers, enabling them to allocate appropriate resources to support this work.

#### *1.4.4 AIM III: The specification of EHR data needs through a conceptual schema*

Objective: Construct and evaluate the counterpart to the Patient, Intervention, Control /Comparison, and Outcome (PICO) framework for the specification of EHR data needs.

Hypothesis: The data-driven conceptual schema represents researcher data needs and provides a reference that aids in the non-vague specification of researcher EHR data needs.

#### Research Questions:

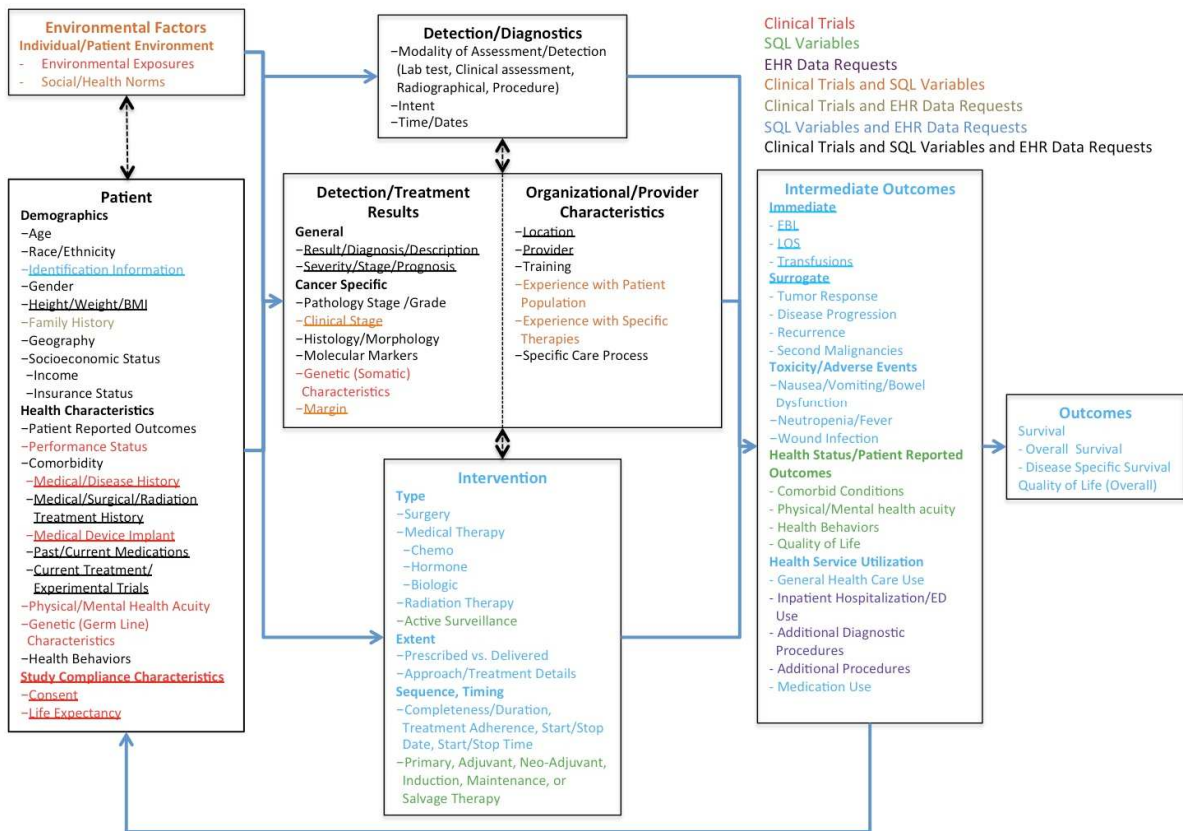
- Does the structure of the Carpenter model allow medical researchers to formulate an EHR data query commensurate with how the PICO allows medical researchers to compose an informational query?
- Is the Carpenter framework able to accommodate data needs outside the realm of cancer comparative effectiveness research?
- Across the three data sources used to enrich the carpenter model, EHR data requests, Clinical Trial inclusion/exclusion criteria, and EHR data request SQL files, what is the

distribution of mutually exclusive and inclusive concepts listed in the enriched conceptual model?

- Does the data enriched model capture medical researcher data needs?

Primary Findings:

I posit the way medical researchers organize medical concepts may aid the efficient elicitation of data needs, and may provide an easier interface for query analysts to map common EHR data elements to medical concepts described in medical researchers' data needs. I generated a data enriched schema presented in **Figure 1-8**.



**Figure 1-8. The data enriched schema. The blue directed edges represent the temporal process as the patient moves through the care continuum. The cyclical nature of this graph implies the patient can re-enter the care cycle. The bi-directional edges indicate an association between the sections. New additions to the schema are underlined, and color-coded classes correspond to the dataset that contains the class.**

My enrichment of the Carpenter framework utilizing three datasets provides some interesting findings. First, I confirmed that the Carpenter framework is a well-organized and comprehensive representation of medical concepts used in CER for cancer, as documented through the high preservation of many original classes from the Carpenter framework in the data-enriched schema. The data-enriched schema contained seventy-nine percent of the concept classes from the Carpenter framework. Additionally, the data-enriched schema contained 86% of the sections and 86% of the directed edges from the Carpenter framework. The high conservation from the

original framework suggested the conceptual organization was preserved. Additionally, my data-enriched schema extends the breadth of classes represented for other medical domains and research approaches.

Finally, the evaluation of my data-enriched schema provided significant insight regarding the understandability of the schema. Specifically, the reorganization of the core sections in line with the directed edge representing a temporal sequence was a major adjustment intended to convey a focus on the sections across a timeline. Additionally, I set out to produce a data needs template with the hope of eliciting more information from medical researchers. During the course of the evaluation, specifically the concept mapping component, the data-enriched schema reminded many participants to describe additional medical concepts they required to complete their research. Many saw the enriched schema as a mechanism to help aid the specification of their needs, and others saw it as a tool to be used during a data needs negotiation with a query analyst.

### ***1.5 Contributions***

This dissertation contributes a deep understanding of BQM to the field of Clinical Research Informatics. Specifically, this dissertation focuses on cognitive sciences in order to provide important insights into the nature of the processes involved in BQM with the goal of providing insight into the roles that knowledge, and strategies play in a variety of cognitive activities for BQM. This work lays the necessary foundation to build advanced cognitive computing systems to facilitate automated BQM applications. My work provides initial knowledge about the BQM process used to access EHR data. I developed a generalized hierarchical task model representing a combination of BQM processes used at multiple institutions. These tasks share common steps used in the reference interview implying the BQM model is able to extract a clear definition from vague EHR data requests. Finally, the dissertation developed and evaluated a data-enriched

conceptual schema for researcher data needs supporting the idea that query templates aid in the specification of medical researcher needs. The impact of the dissertation in a real world setting is unclear. However, this work provides the tools necessary to redesign a workflow facilitating EHR data access, which may then be available for study. Moreover, this dissertation contributes a theoretical framework to inform the design of a cognitive computer agent to serve as an information mediator between the medical researcher and EHR data. I will elaborate on these contributions in chapter 7.

### *1.5.1 Research results*

The following details the publications contributing to this dissertation.

(Hruby et al. 2013a) **GW Hruby**, J McKiernan, S Bakken, C Weng. A centralized research data repository enhances retrospective outcomes research capacity: a case report. *JAMIA* 2013. 20(3), 563-567. Presented in Chapter 1, section 1.1 of this dissertation.

(Hruby et al. 2013b) **GW Hruby**, MR Boland, JJ Cimino, J Gao, AB Wilcox, J Hirschberg, C Weng. Characterization of the biomedical query mediation process. *AMIA summits of Translational Science Proceedings*. 2013 89. Presented in Chapter 4 of this dissertation.

(Hanauer et al. 2014a) DA Hanauer, **GW Hruby**, DG Fort, LV Rasmussen, EA Mendonça, C Weng What is asked in clinical data request forms? A multi-site thematic analysis of forms towards better data access support. *AMIA Annual Symposium Proceedings*. 2014, 616. Presented in Chapter 3 of this dissertation.

(Hruby et al. 2014b) **GW Hruby**, JJ Cimino, V Patel, C Weng. Toward a Cognitive Task Analysis for Biomedical Query Mediation. *AMIA Summits on Translational Science Proceedings*. 2014, 218. Presented in Chapter 5 of this dissertation.

(Hruby et al. 2016a) **GW Hruby**, K Matsoukas, JJ Cimino, C Weng. Facilitating biomedical researchers' interrogation of electronic health record data: Ideas from outside of biomedical informatics. JBI 2016, 60:376-384. Presented in Chapter 2 of this dissertation.

(Hruby et al. 2016b) **GW Hruby**, LV Rasmussen, D Hanauer, V Patel, JJ Cimino, C Weng. A Multi-Site Cognitive Task Analysis for Biomedical Query Mediation. IJMI 2016 93, 74-84. Presented in Chapter 5 of this dissertation.

(Hruby et al. 2016c) **GW Hruby**, J Hoxha, PC Ravichandran, EA Mendonça, DA Hanauer, C Weng. A Data-driven Concept Schema for Defining Clinical Research Data Needs. IJMI 2016 91, 1-9. Presented in Chapter 6 of this dissertation.

### ***1.6 Guide for the reader***

Chapter 2 is a systematic literature review on interactive data retrieval across two literature domains, biomedical informatics, and information science. I identified three promising concepts for optimizing the biomedical query mediation process: (1) Query templates optimize information retrieval, (2) information need complexity influences the information seeking process, and (3) the reference interview is a model for query analyst to elicit non-vague details from the medical researcher.

Chapter 3 analyzes the current forms used to initiate a data request. I found data request forms contain considerable dissimilar in form content, both in the breadth and depth of the topics covered, most offered limited aid for medical researchers to articulate their information need.

Chapter 4 lays out the content exchanged during BQM between medical researchers and query analysts. I produced several visualizations of the content exchange. Additionally, I show

this process contains a cyclical component suggesting multiple iterations between the medical researcher and query analyst until a consensus is met.

Chapter 5 performs task analysis of the biomedical query mediation process between medical researchers and query analysts. I identified a set of tasks used by multiple query analysts to extract a clear understanding of the medical researcher's EHR data need.

Chapter 6 is a follow-up on one of the promising concepts identified from chapter 2, the important role semantics play for optimal information retrieval. I present a data-driven concept schema for defining researcher data needs.

Chapter 7 presents a summary of the contributions from this dissertation. Additionally, I will present my prioritized list of future endeavors needed to enable an automated data access engine.



## **Chapter 2. Facilitating Medical Researchers' Interrogation of the Electronic Health Record: Ideas from outside of Biomedical Informatics**

### ***2.1 The Tradition of Clinical Data Reuse for Medical Research***

Biomedical research has long benefited from a valuable and cost-effective data source: patient health records [23]. For example, the Apgar Scale [51] and the Goldman multifactorial index of cardiac risks [52] were both derived from analyses of patient health records. With the increasingly pervasive adoption of EHR systems worldwide [53], many have recognized the rich clinical data increasingly made available by EHRs as a promising data resource for accelerating medical knowledge discovery [54] and for enabling comparative effectiveness research [34-37, 55]. Subsequently, the demand for reusing EHR data for research among biomedical researchers has been rising rapidly [19, 45, 47, 56, 57]. Assisting biomedical researchers to interrogate EHR data has been a vital mission for the biomedical informatics research community. However, this task faces significant human and technological barriers [55, 58, 59]. Current data captured by EHRs are not optimized for secondary uses beyond clinical care or administration-centered documentation practices so that many institutions employ intermediating query analysts to retrieve EHR data for biomedical researchers, with varying degrees of assistance from self-service query tools. The use of intermediaries may not scale to large data networks such as the clinical data research networks (CDRNs) as part of the PCORnet [60] established by the Patient Centered Outcomes Research Institute [42]. For example, the heterogeneity of data representations across institutions and the complex, idiosyncratic local data collection processes

that often remain “black boxes” to intermediaries are serious barriers facing users of the data contained in PCORnet. To contain cost for involved expensive operations, many institutions have to charge clinician scientists for reusing such data collected during patient care for research. Meanwhile, self-service query support is still at its early stage of development and may not support sophisticated data queries [20, 61].

By identifying and reviewing existing theories and best practices for general data interrogation, I aim to inform the design of next-generation EHR data interrogation aids that directly facilitate biomedical researchers to autonomously retrieve and reuse this data for clinical and translational research. Towards this goal, this paper contributes a literature review on this topic. I summarized existing approaches, identified research gaps, and recommended research priorities. Although this review focuses on EHR data, the knowledge gained may generalize to interactive end-user data interrogation for other reusable health data resources.

## **2.2 Methods**

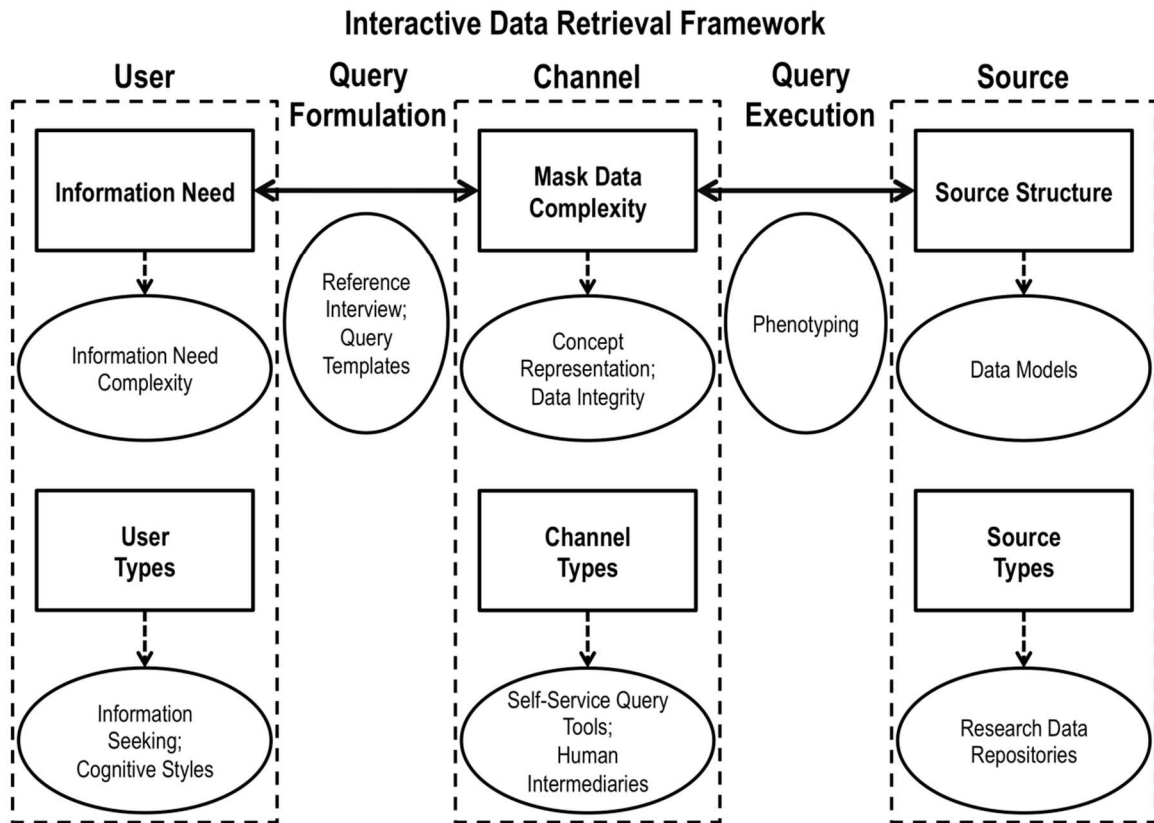
### *2.2.1 Development of a Conceptual Framework for Interactive Data Retrieval*

An information retrieval process addresses information needs using a sequence of tasks [8]. The complexity of the task sequence is dependent on the information retriever’s *a priori* knowledge of the information need, the information retrieval process stipulated by data owners, and the complexities of each of the tasks used to complete the information retrieval process [2, 29, 49, 62, 63]. Many models have been developed for characterizing the information retrieval process or for investigating how information systems enable users during this process [28, 64-

72]. For example, the berry-picking model [28] and the sense-making model [65] focus on how the user iteratively refines search terms and information needs based on their conceptualizations of the information space. Among all existing models, only one developed by Bystrom and Jarvelin explicitly defined three entities that influence the complexities of an information retrieval process: user, channel, and source, to characterize the information retrieval process [49]. The user entity focuses on the user's profiles, communication styles, and knowledge of data. The channel masks the complexities of the source and translates user information needs to data representations. The source concerns data representations towards optimal data retrieval efficiency. Therefore, source is the container of information and channel guides efficient navigation of the source. I adopted this conceptual framework to organize the literature to discuss interactive EHR data retrieval.

In this paper, I survey related methods and theories in the context of EHR data retrieval for secondary use by end users who are unfamiliar with the data, such as biomedical researchers and clinician scientists. Since this paper aims to augment end users with improved query formulation, I focus primarily on effort supporting the user and the channel, while briefly describe existing efforts on the source. I adopted the constructs of user, channel, and source combining them with the concepts of query formulation and query execution, as shown in **Figure 2-1**. For example, a researcher may want to identify an institution's mortality rate among its patients undergoing coronary artery bypass. Query formulation transforms vague data requests (e.g., "adult patients younger than 75 years old with coronary artery bypass surgery last year") into contextualized data requests consisting of specific EHR data elements (e.g., "patient DOB, Current Fiscal Year, billing code for coronary artery bypass billed in this current fiscal year"). The step of query execution further translates this query from contextualized data elements into executable

database queries consisting of disparate data types and represented by local terminologies using The Structure Query Language (SQL).



**Figure 2-1. A conceptual framework for interactive data retrieval.**

### 2.2.2 Literature Search Methods

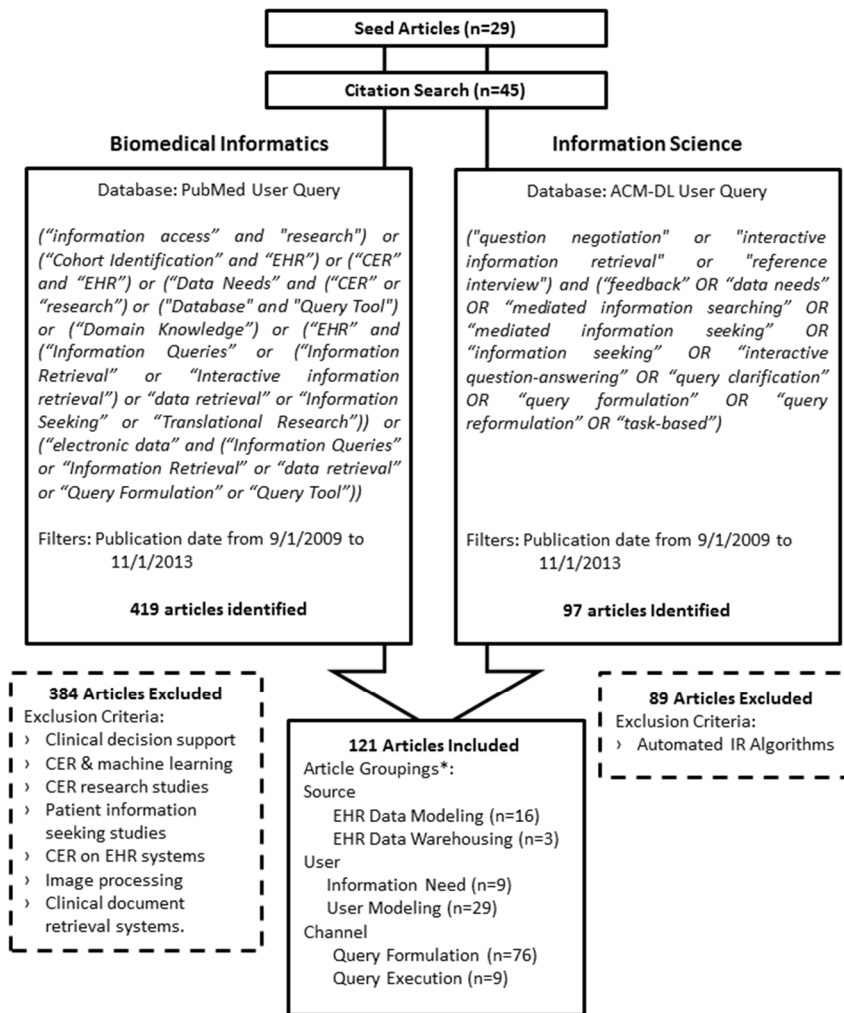
I adopted the post-positivist model for research [73]. I iteratively searched for related work published between 2009 and 2013. Following this model, I searched beyond the field of biomedical informatics or clinical research informatics that were obviously relevant and included the literature in informatics and computer and information science. Also, I categorized all included citations by their focus on user, source, and channel so that significant amounts of

qualitative information were categorized to produce quantitative information to help draw the big picture and deduce evidence gaps.

I seeded my search with 29 articles proposed by my research advisor, Chunhua Weng [2, 3, 10, 26, 28, 30, 42, 48, 49, 64, 74-92]. Additionally, citation searches within these articles provided an additional 45 references [1, 4, 6-9, 11-13, 29, 58, 62, 63, 65-69, 72, 93-124]. These 74 articles served as the basis for the development of the search query. With my initial search query, I iteratively searched and reviewed the identified articles, incorporated new search keywords as they emerged, revised my search string and article inclusion/exclusion criteria iteratively according to their relevance as determined by manual review. I surveyed both the information science literature (i.e., <http://dl.acm.org>) and biomedical informatics literature (i.e., MEDLINE). I limited my search to the main journal citation databases, the ACM Digital Library, and Medline, for the respective fields of information science and biomedical sciences, concluding that these databases provides a representative sample for my topic.

**Figure 2-2** is a flow chart that highlights the final search strings for PubMed and ACM databases and the inclusion and exclusion criteria for selecting articles for this review. I generated the final search string and reviewed the title and abstracts of the returned articles; articles containing any of the exclusion criteria were removed from the pool. Next, I iteratively reviewed and annotated the 125 included articles using the conceptual framework developed in Section 2.1. For each annotation, the first author wrote a summary and justification paragraph. After annotation, the first author reviewed the summary and justification paragraphs for each set of articles against the conceptual framework components and derived themes within these sections. For the source, the major themes identified were EHR data modeling (how data is structured and the standards used to store data elements) and warehousing (dedicated silos of

data for secondary use). For the User, the major themes identified were Information Need (defining the complexity of the need) and user modeling (understanding the user attributes and the information seeking strategies used). For the channel, the major themes identified were query formulation (the process of defining an information need) and execution (the process of translating an information need into an executable database query).



**Figure 2-2. The search strings and article selection flowchart (\*Articles could be classified into multiple categories as some content spans multiple categories).**

**Table 2-1** organizes the articles according to the Source, User, and Chanel conceptual framework. More work focused on user modeling, human intermediaries, and reference interview in the field of information science than in the field of biomedical informatics. In the following sections, I will synthesize the major themes from each discipline and compare and contrast their ideas from difference sources.

**Table 2-1. The distribution of relevant topics in two bodies of literature**

	<b>Biomedical Informatics</b>	<b>Information Science</b>
<b><u>SOURCE</u></b>		
EHR Data Modeling	[36, 56, 75, 80, 82, 85, 99, 100, 116, 125-131]	
EHR Data Warehousing	[13, 17, 74]	
<b><u>USER</u></b>		
<i>Information Need</i>		
Information Need Complexity	[83, 86, 101, 132-134]	[3, 49, 117]
<i>User Modeling</i>		
Information Seeking Processes	[87, 90, 105, 112]	[1, 2, 8, 10, 49, 63, 71]
User Cognitive Styles	[48, 59]	[2, 4, 9, 28, 62, 64-72, 93, 113]
<b><u>CHANNEL</u></b>		
<b>Task 1: Query Formulation</b>		
Concept Representation	[29, 39, 59, 80, 84, 92, 99, 100, 121, 127, 135-139]	[123]
Characterization of Data integrity	[50, 59, 135, 140]	[63]
Query Templates	[81, 89, 101, 104, 109, 110]	[141]
Self-Service Query Tools	[11-16, 20, 75, 84, 85, 102, 111, 134, 142-144]	[76, 118, 145-149]
Human Intermediaries	[23, 79, 150, 151]	[6, 7, 26, 78, 91, 115, 119]
Reference Interview		[26, 77, 88, 94-98, 103, 108, 114, 119, 120, 124]
<b>Task 2: Query Execution</b>		
Phenotyping	[30, 92, 106, 107, 122, 137, 152-154]	

## 2.3 Results

### 2.3.1 How the Data Source Facilitates User Access

The barriers to task-based data access in life sciences fall into two categories: (1) human factors (e.g., a user lacking a correct conceptualization of task complexities); (2) system factors (e.g., technology limitations in the existing systems such as data heterogeneity and fragmentation) [59]. Table 2-2 presents examples of known barriers and corresponding recommended solutions. Human factors relate to the user, while system factors relate to the metadata of the source, or in this instance, the lack of metadata concerning known and underlying EHR data quality issues.

**Table 2-2. EHR data access barriers and solutions**

<b>Level</b>	<b>Barriers</b>	<b>Solutions</b>
<b>Human factors</b>	Known and latent EHR data quality issues [50, 59, 140]	Transparent reporting of data limitations for intended uses [63, 135]
<b>System factors</b>	System interoperability[130]; Real time analytics [36, 56]; Data fragmentation[130]; Data heterogeneity [50]	Streamlining data access workflow[100] [85]; Data warehousing [30, 107]; Data modeling and integration strategies (i.e. The Observational Medical Outcomes Partnership Common Data Model) [80] [82, 131]

In the context of this study, I discuss data warehousing at the institutional level rather than at the state or national level, although the same principles may apply to both. Data warehousing has been a focus in the clinical research informatics community for overcoming technical barriers to data access by providing efficient access to integrated EHR data, which can be centralized or federated. The centralized infrastructure [74, 75, 85] avoids data heterogeneity but can introduce challenges for incorporating new or latent data elements over time because each update affects the entire database, hence not being able to scale easily. In contrast, the federated architecture is



flexible [116, 125-127], allowing autonomous data control and growth over time. It permits data representation heterogeneity [99, 128, 129]. It also enables the leveraging of distributed local expertise for data modeling and data quality control and supports geographic distribution by multiple stakeholders. The National Center for Education statistics has summarized the tradeoffs between centralized and federated data repositories [155]. Briefly, centralized systems ease data governance, increase data retrieval performance, provide uniform data for efficient data mining, entail a high-cost burden for ensuring data currency and completeness, and are harder to scale for evolving data needs and different data access workflows [156]. Many institutions have centralized data repositories such as STRIDE [17]. The most widely used platform, i2b2 and its SHRINE architecture, supports distributed and federated data repository [13]. The federated architecture represents an established model for most large data networks such as PCORnet [60, 157].

### 2.3.2 *USER*

#### 2.3.2.1 Information Need Complexity

One component of user modeling is understanding the complexity of information need, which depends on the diversities in the use context [101], variations in information seeking behaviors [49], and heterogeneity in languages used to express the information need [117]. Others have proposed a categorical scale of data need complexity by measuring the amount of work required to accomplish the task for satisfying the information need [3, 113]. Structures have been defined to characterize a complex information need, including a problem statement, an event of interest, a comparison event (if necessary), and potential effects of the event of interest [83]. Unfortunately, little is known about the data needs of biomedical researchers [132-134].

The very few studies available have largely focused on identifying sets of the major data elements needed to facilitate research in particular medical domains of interest. Cimino et al. leveraged these key data elements needed by researchers to inform the development of a user-centric query tool [134]. What has been lacking includes a thorough understanding of user preferences and search behaviors, as well as communication patterns between biomedical researchers and query analysts for clarifying data needs iteratively.

### 2.3.2.2 User Cognitive Styles

Five user characteristics influence a user's information seeking tactics:

- (1) Phases of mental model of information – confusion, doubt, threat, hypothesis testing, assessing, and reconstructing [93]
- (2) Levels of need – visceral, conscious, formal, and compromised [26]
- (3) Levels of specificity – new problem, new situation, experiential needs, and well understood situation [2]
- (4) Expression – questions connections, and commands gap [2, 26]
- (5) Mood – invitational, and indicative [93]

Kuhlthau provides a great amalgamation of these user characteristics in her theoretical foundation of the information seeking process [1], which was well supported by Vakkari's review [10]. Additionally, the information seeking process has been modeled within the biomedical literature. Mendonca et al. [87] and Hung et al. [90] have provided models for the biomedical literature information seeking process, which propose to aid user's search strategies through well-structured clinical queries and by leveraging the knowledge of human search experts, respectively.

User cognitive styles shape information seeking processes [71]. Many describe these styles along two orthogonal axes: analytic and descriptive. The analytic cognitive style captures an active approach to information seeking in which conceptual level questioning is used to resolve information need, whereas the descriptive cognitive style represents a passive approach, where concentration on the most detailed level of the subject matter is used to resolve an information need. User cognitive styles are either passive (high descriptive and low analytic with attention to detailed questions) or active (high analytic and low descriptive questioning). The active styles represent more effective and efficient search strategies than passive styles [72].

A user's domain knowledge and technical knowledge are both associated with their cognitive styles and effective search strategies [9, 59, 83]. The users' cognitive styles can be differentiated [4, 113] by search tactics. Users of varying cognitive styles have different sense making strategies or processes. Studies of cognitive styles offer a more generalizable mechanism to stratify users and to predict individual information seeking styles. Cognitive science allows classification of the demand characteristics for a particular task and to focus on the appropriate problem dimensions [48]. Although user cognition during EHR data interrogation is rarely studied in the biomedical informatics literature, especially for facilitating EHR data interrogation, such studies are much needed. Cognitive studies of users can enable user centered EHR data interrogation designs aiming to improve the user experience and effectiveness.

### 2.3.3 *CHANNEL*

The complexity of a data source is multidimensional, including heterogeneous semantic representations [80, 99, 100], opaque data integrity, complex time expressions [121, 123], and fragmented knowledge of logical data constructs [29, 59, 135]. A channel enables users to

navigate data despite complexities by providing users with an abstract mechanism to interact with data sources during query formulation or query execution.

### 2.3.3.1 Query Formulation

The query formulation component facilitates the iterative interaction between the user and the source to formulate a query in a user's language. Hripcsak et al. investigated two EHR data retrieval channels, AccessMed and Query by Review, and found neither achieved adequate performance, indicating the difficulty of query formulation for EHR data [84]. I also reviewed common aids for query formulation and related execution challenges, human intermediaries, query templates, as well as self-serving query tools.

#### 2.3.3.1.1 Human intermediaries

Human intermediaries are often employed to formulate user queries to ensure feasibility and precision [7, 78, 79, 115, 150]. Intermediaries usually have received formal training and possess a deep understanding of work culture and technical skills for data querying [6, 23, 91, 151]. The biomedical literature provides scant knowledge regarding how human intermediaries operate in the biomedical information rich domain. Information science has extensively studied human intermediaries, most notably, librarians and their development of the reference interview technique [26, 119]. Reference interviews elicit tacit user needs, specify vague queries, narrow overly broad questions, and suggest further dimensions of the information need that the user may not have expressed but are logically related to the user-stated objective. It enables a skillful interrogation process widely adopted by librarians for converting vague and general data request provided by users into specific data queries expressed using user language [26, 88, 95-98, 114, 119, 120].

Elicitation strategies used in reference interviews have been explored to improve negotiation of information needs [77, 94, 103, 108]. Specifically, interrogation strategies are primarily developed to obtain the user’s objective surrounding the information need. When users are aware of the reference interview’s purpose, they are willing to provide additional information on objective and intent. In a related study, Lin et al. analyzed need negotiations and extracted a taxonomy of clarification questions that fit within a set of six classes [124].

**Table 2-3** illustrates the taxonomy applied in the context of EHR data interrogation with example clarification questions. These results imply that in the context of interactive EHR data retrieval, the reference interview may provide human intermediaries with a more efficient workflow to best extract a non-vague description of the data need from the user.

**Table 2-3. A Taxonomy of Clarification Questions Utilized During Need Negotiation**

Question Type	Definition	EHR Data Clarification Examples
Relevance threshold	These clarification questions (CQ) are used to better understand the user’s relevance threshold, a mapping of a continuous scale into a binary decision.	<i>What is the A1c threshold for the diabetes patients that you are looking for?</i>
Ambiguity in conjoined facets	Information needs are composed of multiple conceptual facets. These CQ establish the relationships between these facets.	<i>Do you only want patients that have both a diagnosis of diabetes and hypertension or patients with a diabetes diagnosis with or without a hypertension diagnosis?</i>
Example concept	These CQ address whether a particular concept within a set of documents is an example of a concept referenced initially?	<i>Would patients without Diabetes diagnoses, but taking a medication for diabetes be of interest to you?</i>
Closely related or subset concept	These CQ highlight the user’s interest in a particular concept X extend to a closely related concept X’.	<i>Do you want patients with both types of Diabetes? Type 1 and 2?</i>
Related topical aspects	These CQ highlight if the user is interested in facets that are conceptually related, but not directly requested.	<i>Does it matter what treatment protocol the diabetic patients are on?</i>

Acceptability of summaries	These CQ highlight if the user would be interested in a general summary or overview.	<i>Do you just want to know how many patients fit your criteria?</i>
----------------------------	--	--

### 2.3.3.1.2 Query Templates

Templates are another effective technique for expressing standards-based structured data needs free of ambiguity and vagueness [105, 112]. A query template provides an organizational structure for the user to describe their information need in a non-vague structure [109]. Templates have been developed to access clinical data [81] and medical literature [89, 101, 104, 109]. The Patient, Intervention, Control/Comparison, and Outcome (PICO) framework is extensively used to explore the medical literature for relevant resources [110, 141]. Currently, there is no well-accepted standard template based on community consensus. Instead, many medical institutions require users to complete data requests using free text.

### 2.3.3.1.3 Self-service query tools

Since human intermediaries are expensive and time-consuming, self-service query aids have been pursued in many institutions in recent years [12-14, 16, 75, 84, 102, 134, 142, 144]. Some are form-based, while others support queries in natural language [76, 84]. Visual query formulation is a recent trend and is expected to reduce user cognitive load by presenting information intuitively to the user [145]. The Informatics for Integrating Biology and the Bedside (i2b2) project represents the most widely adopted self-service EHR data retrieval system. The system's terminology explorer and query builder allow the user to search the terminology for applicable terms and build cohorts using a frame system with Boolean constraints [11, 13-16, 85, 111, 134, 142-144]. Deshmukh et al. studied the various types of data requests these applications

were able to resolve. The study suggested that i2b2 facilitated relatively simple, cohort identification queries. They also acknowledged that the majority of requests they studied were “simple” queries [20]. These reports indicate that the majority of self-service tools for EHR data support a limited scope of data specification. Many complex data request require more than simple constraints, e.g. “all patients diagnosed with diabetes between May and July of 2012”, but complex relations between data elements, e.g. “all patients with their first recorded diagnosis of diabetes between May and July of 2012, and all lab glucose tests after their diagnosis and before the start of treatment.” For these complex temporal queries, temporal query tools can be used to visualize raw data or concepts over absolute and relative temporal timelines [12, 118, 146-149]. Though experimental, these tools offer a solution to a complex problem of temporal specification and visualization of EHR data. Meanwhile, significant EHR data processing and transforming are needed for these systems to work appropriately. Finally, these tools place the burden to identify the correct terms associated with a particular medical concept on the user.

For the properly trained user, these self-service query tools represent an acceptable solution for EHR data retrieval of simple patient cohorts. It is clear the ultimate goal for these efforts is a fully functional self-service model, however, self-service query tools provide minimal support for linking EHR data representations with medical concepts. Additionally, complex temporal relationships amongst the medical concepts cannot be expressed using these tools.

### 2.3.3.2 Query Execution

The query execution component focuses on the conversion of a user query into an executable database query by mapping medical concepts specified by the user to the EHR data elements that define that concept. The Electronic Medical Records and Genomics (eMERGE)

consortium [137] has studied the problem of phenotyping disease concepts through enumerable data elements within the EHR. The EMERGE consortium has shown that each disease phenotype contains significant heterogeneity, underlying elements representing nested Boolean logic, complex temporality and ubiquitous ICD-9 codes [137, 152]. As of 2013, the group has validated 13 phenotypes [92]. Although the temporal nature of EHR data was considered only in some of the eMERGE phenotypes, temporal abstraction is an important technique for EHR phenotyping. Post et al. have established the PROTEMP method, which allows for the abstraction of temporal data events [122]. Additionally, Shahar's framework on temporal abstraction has described promising methods for formally representing temporal patterns [106, 153, 154].

## ***2.4 Discussion***

Interactive EHR data retrieval involves complex interactions among users, sources, and channels. The healthcare industry has heavily invested in infrastructures for data integration. To maximize the return on investment and to use these resources to advance medicine, my goal is to make such data accessible to biomedical researchers for various computing needs.

Self-service query tools have not fully addressed user needs and hence make human intermediaries indispensable in many institutions. These intermediaries utilize a needs-negotiation process with the user. Barriers facing this process include the lack of medical and technical knowledge by the intermediary and the user, respectively. Bridging these knowledge gaps for the intermediary and user may engender efficient communication. Furthermore, a user often presents a vague understanding and description of their information need. Intermediaries may benefit from a standardized structure through which requests can be organized, which may reduce the ambiguity of the request and allow the intermediary to focus on query execution.



Relatively little is known regarding biomedical researcher’s cognitive styles and information seeking strategies. **Table 2-4** lists key knowledge gaps and potential recommendations for bridging those gaps. Additional exploratory studies are needed to bridge the knowledge gaps concerning how biomedical researchers interrogate EHR data and what their barriers are. Additional investment is needed in interactive information retrieval that augments not only the source but also the user.

**Table 2-4. The knowledge gaps and recommendations for advancing EHR interrogation**

<b>Aspects</b>	<b>Knowledge Gaps</b>	<b>Recommendations</b>
<b>User</b>	Lack of measure of information need complexity Lack of knowledge of how the cognitive styles of medical researchers affect the information seeking process	Develop metrics for measuring EHR information need complexities Conduct qualitative studies of the information seeking processes of multidisciplinary medical researchers and their barriers to clinical data access
<b>Query Formulation Channel</b>	Lack of formalized structure for the medical researcher to express their information need Poor understanding of the data need-negotiation process performed by data intermediaries	Investigate other formal structures used for document retrieval, i.e. PICO framework Support reference interview for EHR data interrogation

Information-seeking models explain the sub-optimal outcomes resulting from current methods used for EHR data interrogation. The granularity of data that biomedical researchers are seeking adds more complexity to existing information seeking models. Additional understanding of process-oriented EHR data access by biomedical researchers is needed. To this end, I extracted from the literature three promising concepts to aid in the construction of an ideal process-oriented EHR data interrogation model.

First, both information science and biomedical informatics have established the important role semantics play in optimal information retrieval. For example, the PICO framework is an excellent user aid, which helps organize and express the information need of clinicians. In the

context of EHR data interrogation, the PICO framework can be potentially a good starting point for supporting the expression of biomedical researchers' EHR data need.

Second, the complexity of information need shapes information search tactics. A metric for complexity assessment of the information need can optimize resource allocation while resolving the user's information need. Complex data requests could be directed to a query analyst, whereas simple request would be facilitated through improved self-service query tools. A standardized method for EHR data need complexity assessment can further enable global resource optimization.

Third, the established reference interview has effectively helped librarians to clarify user needs. Query analysts provide a similar role as librarians but lack a guideline on how to conduct a reference interview for users seeking EHR data. Experience is the only way for query analysts to gain insights and expertise for this task. To increase the efficiency and effectiveness of the EHR data needs negotiation, an EHR-based reference interview, conducted by a query analyst, may aid the query formulation process in the translation of vague EHR data requests into specific data queries. More studies are needed in this area to enable reference interview for EHR data.

My study has two major limitations. First, I developed the search criteria based on pre-selected topics. This method may be biased towards self-selected topics and leave out topics relevant to this review but not searchable by the query derived from the pre-selected topics. Nevertheless, I used an established method to identify a focused topic set and believe this review is representative of what is available in the literature. Second, my focus on the most recent four years of literature may have excluded seminal articles in the field from the past; however, I believe my exhaustive citation search should have largely mitigated this problem.

## **2.5 Conclusion**

This review surveys the methodological considerations for interactive EHR data interrogation. I have identified knowledge gaps and research opportunities for advancing EHR data interrogation. My results show that application of the reference interview technique for EHR data is a promising direction for improving communication with biomedical researchers during EHR data interrogation. More user understanding is needed to enable such support cost-effectively. I suggest that cross-disciplinary translational research between biomedical informatics and information science is needed to apply theories and techniques from information science to facilitate efficient end user data interrogation in life sciences.

## **Chapter 3. Initiating Electronic Health Record Data Requests**

### ***3.1 Introduction***

Research tasks that were previously impractical, if not impossible, to perform with paper-based health records have now become achievable due to the large volumes of data stored in “readily accessible” electronic format. However, in addition to privacy and security constraints, numerous difficulties remain with respect to access and use of the data [158]. Compared to paper records, EHR data should be much easier to aggregate across large numbers of patients, but the complexity of the underlying systems, including the heterogeneity in metadata, data structures, and even the data itself, often hinders their computational reuse by a broad range of stakeholders [159-161]. Further, prior work has shown that it is not uncommon for a hospital to have hundreds of different IT systems [162]. Data from multiple health information systems are thus often aggregated into databases commonly referred to as data or information warehouses, data repositories, data marts, or data networks [23, 93, 163-167]. A major theme of the NIH roadmap has been the idea of “Re-Engineering the Clinical Research Enterprise” [168], and one of the recognized challenges has been providing the means to “facilitate access to research...resources by scientists, clinicians” and others [169]. However, two major barriers exist with respect to data access.

First, to help meet the needs of clinical and translational research, “self-service” tools have been developed to provide a means for data access as well as analysis and visualization [14, 16, 17, 134, 170-173]. Many of these tools have been widely implemented and have achieved a good level of adoption. While self-service tools have been demonstrated to work well for various scenarios [174], by nature of their intended simplicity for a broad user base, these systems often

cannot handle all of the complex data needs that are required by biomedical research teams [20, 134]. It is not common for researchers to have the database knowledge or to have the understanding of what is involved in data retrieval. In contrast, data managers or query analysts do not know how to ask questions to elicit data needs using non-technical language understandable by researchers [175]. Data need negotiation involves several “trial-and-error” iterations. As a result, many institutions have recognized the need to invest in informatics or IT experts (often called query analysts or report writers) to serve as an intermediary between the complex data sources and the biomedical researchers, the latter of whom have significant domain expertise but often lack training in data access approaches such as the use of structured query languages (SQL) [151, 176, 177].

Second, for liability considerations and Health Insurance Portability and Accountability Act or regulatory compliance, data owners need to carefully check the credentials and qualifications of data requesters. These requirements involve lengthy review processes with multiple institutional review offices. An important artifact, the *data request form*, is the nexus linking all the stakeholders in the process of providing data access for researchers. Such forms are generally meant to serve documentation and communication needs for multiple stakeholders, including researchers, query analysts, data owners, and regulatory officers [178]. They can provide a means for researchers to list their credentials and specify their needs through a formal request process. They also help data stewards verify if the appropriate regulatory approvals are in place and to help with other administrative bookkeeping. Importantly, data request forms are also meant to provide a means for research teams to communicate complex data needs in a manner that can be understood by a query analyst and converted into executable database queries [121]. It follows, then, that the manner in which these forms define the data need can have major

downstream consequences for the subsequent research on which the request is based. Yet there are no published standards for designing EHR data request forms, or even best practices about which an institution can turn to in constructing a form. It is up to each institution to develop their own form with the hope that the right questions are being asked of data requestors in order to ensure that data needs are being met accurately and efficiently.

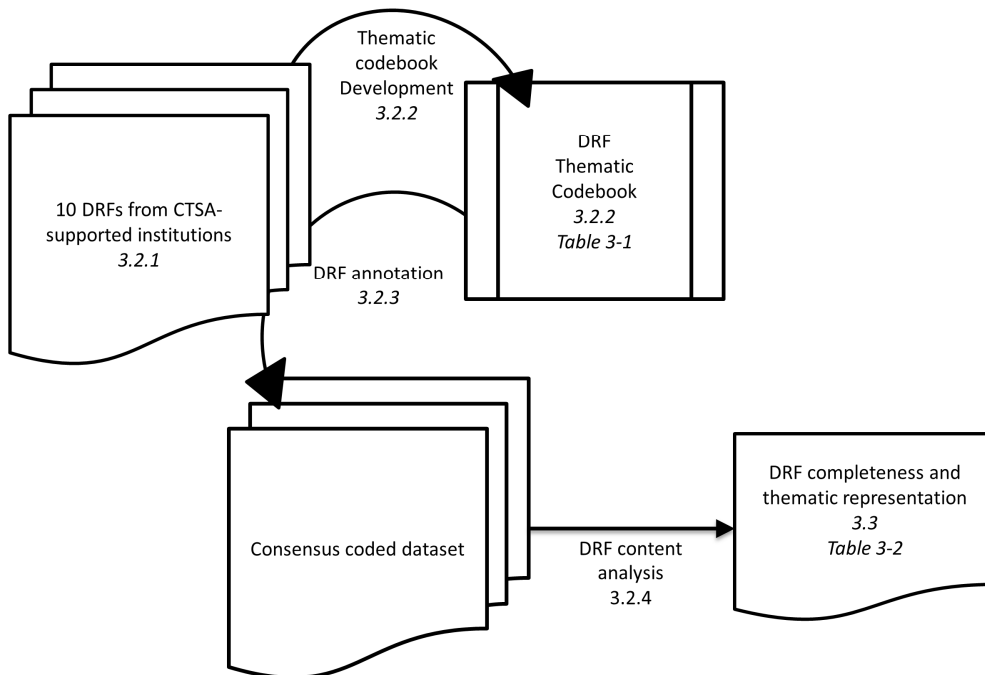
The data request form plays an indispensable role in facilitating data access for researchers in many institutions. I conclude that to provide better data access to the broad clinical and translational research community, we need to understand (1) if the current forms efficiently collect information needed by data owners and effectively communicate data needs of researchers, and (2) if they collect necessary and relevant information that cannot be extracted for reuse from existing institutional information systems. I conducted a formal content analysis of data request forms from multiple academic institutions affiliated with a Clinical and Translational Science Awards. My goals were to develop a deeper understanding of what questions are typically asked on the forms, to ascertain whether current data request forms provide adequate coverage of salient details, and whether they capture the medical researcher's data needs effectively.

This paragraph summarizes the steps taken to achieve these goals. I obtained ten, blank data request forms from Clinical and Translational Science Awards supported academic medical centers in the United States. Then I developed a form annotation schema based on the consensus of two annotators and used this coding book to annotate the forms. I then conducted a detailed content comparison and analysis of the forms and identified information deficiencies as well as unnecessary workload imposed on researchers that exist across many forms in use today. Finally, I suggest insights and recommendations from my analysis that could be used to improve the

content of data request forms and, ultimately, improve the process for obtaining complex data from institutional repositories in support of clinical and translational research. These steps and recommendations are detailed in the balance of this chapter.

### 3.2 Methods

Figure 3-1 presents the workflow overview for this research project.



**Figure 3-1. Data request form content analysis workflow with italic numbers representing corresponding subsections within this chapter with further detailed information.**

#### 3.2.1 Collection of data request forms

Ten data request forms were obtained for this study. All forms were in use at Clinical and Translational Science Awards supported academic medical centers around the US as of February 2014. Four of the forms were obtained through my personal contacts, whereas the remaining six

were identified through an online search with the Google search engine using the strings “EHR data request” and “medical research data request.” The ten Clinical and Translational Science Awards supported institutions from which these forms were actively in use are: Boston University, Columbia University, Northwestern University, University of California - San Diego, University of California - San Francisco, University of Colorado Denver, University of Kansas, University of Michigan, University of Wisconsin, and Vanderbilt University. Note that these institutions are listed here in alphabetical order, which does not match the order in which they are presented in the results section, wherein only a letter is used to identify each form.

### *3.2.2 Development of a codebook*

I selected two expert reviewers to develop the codebook. Both reviewers are senior PhD students with extensive experience navigating EHR data requests and working with EHR data. Additionally, I selected my advisor as a third reviewer to advise and consent on the final codebook. My advisor has significant experience working with EHR data for secondary use.

Five of the ten data request forms were randomly selected for developing the coding schema for the content analysis. The two reviewers independently evaluated the five forms and developed a list of themes derived from the forms. These themes were based only on the actual questions asked on each of the forms. Several research questions helped guide the analysis. These included: (1) what high-level organizational categories can data request form elements be assigned to? (2) what percentage of metadata could potentially be obtained from source systems without asking research teams to copy it to a form? (3) how are request form items distributed between administrative data (i.e., ‘bookkeeping’) and actual data requests? and (4) how much detail does each element on a form seek to obtain from a user completing the form?



The theme lists from both reviewers were then compared, discussed, and consolidated into a single list. My advisor evaluated the merged list and refined it further. Finally, the two reviewers compared two randomly selected forms to finalize the codebook and address additional gaps in code coverage. Similar themes were then grouped into logical categories (e.g., “Compliance”, “Data Use”) and numbered. This final list served as the codebook, which is shown in **Table 3-1**.

**Table 3-1. Form elements comprising the codebook for the content analysis of the data request forms, including examples of each type of element. When an element could be coded as Simple [S] or Extensive [E], an example of each is provided. Basic elements were only coded as Simple if present; thus no Extensive example is provided.**

Code	Name	Description	Example(s)
1.0	<i>Requester Metadata</i>	<i>Any form elements that describe the user requesting data</i>	<i>not a coding element</i>
1.1	Name	This element may include the name, and/or contact data of the requester	[S] Requester Name [E] Requester Name, Department, Email
1.2	PI/Supervisor/ Department Head	This element may include the name, and/or contact data of the requester's PI, supervisor and/or department head	[S] Supervisor Name [E] Supervisor Name, Department, Email
1.3	Billing/ Administrative	This element may include the name, and/or contact data of the requester's administrator or other billing information	[S] Administrative Name [E] Administrative Name, Department, Email
1.4	Other	Any other attributes associated with the requester, and not associated with the content of the request	[S] Are you a part of the Clinical and Translational Science Awards?
2.0	<i>Request Metadata</i>	<i>Any form elements that describe the actual request</i>	<i>not a coding element</i>
2.1	Study Title/Request	This is a brief summation of the request.	[S] Project Title; Research Question
2.2	Existing/New Request	This element specifies if the request is new or a modification to an existing request	[S] Is this a new request or a modification to an existing report
2.3	Funding Source	This element is asking who is financially supporting the use of this data.	[S] What are your funding sources? [E] Will funds be used to pay subcontractors; do funding sources have restrictions on the use of the data collected for this project?
2.4	Request Purpose	Concerns the use of the data being request. For example will it facilitate an internal administrative report, research or preparatory for research, cohort/Clinical trial recruitment?	[S] Will the requested data be applied to any of the following areas? Non-research, Patient Care, Operations, Research, etc.
2.5	Request Type	This element is allows the user to specify the degree of data access	[S] Multiple Choice: Self-service, Super user [E] Study Design Consultation, Research Navigator
2.6	Data Sources	Any element that asks the user to specify the source of data, for example this maybe a particular database, or a particular clinical site where the user thinks the data may originate.	[S] Sources of data? (Text Box) [E] Sources of data? (Multiple Choice)
2.7	Data Element Specification	This element refers to any description of the medical data elements the requester is after.	[S] Describe the data you need. [E] What is your selection criteria, From what time period... What data fields do you need?
2.8	Recurring Requests	This element is specific to the frequency of data delivery. A clinical trial that submits a request to aid recruitment may wish to receive a weekly dump of potential matches.	[S] Is this a one-time request or recurring?
3.0	<i>Compliance</i>	<i>Form elements related to a compliance attribute, such as IRB, PHI, internal regulations, or documentation requirements</i>	<i>not a coding element</i>
3.1	Institutional review board (IRB)	If the request is research, this element request details on the IRB number or if the protocol is IRB exempt.	[S] IRB number
3.2	IRB Proof	Elements that require IRB proof	[S] Please upload your approved IRB protocol
3.3	Protected Health Information (PHI)	Regardless of request purpose, this element specifies HIPAA compliance and asks to what level of identified data (if any at	[S] Will the data be identified or de-identified [E] Please select the type of data you will need:

		all) are needed.	identified, de-identified, limited decedent, aggregate counts...
3.4	Compliance Other	This element concerns any type of compliance attribute, whether it be IRB, PHI, internal regulations, or documentation requirements that could not be classified elsewhere	[S] Provide your consent (or waiver of consent)
4.0	<i>Data Use</i>	<i>Refers to how the requester is going to use or share the data</i>	<i>not a coding element</i>
4.1	Internal Sharing	Data This element represents how the user is sharing the data within their team, where the data is going to be stored, how the data is to be delivered, or the format of the data.	[S] Please describe data storage and use plan [E] Who will have access to the data, where the data is to be stored, data delivery & format
4.2	External collaborators data use agreement (DUA)	If the requester is sharing the information with an external collaborator, is there a formal data use agreement	[S] Is there a DUA? [E] Name non-affiliated project team members that will have access to the data; upload DUA.
4.3	Public Sharing of Original Dataset	This elements refers to the intent of the requester to publish the original dataset	[S] Will data be made publically available? [S] Do you plan on making this data publically available, how so?
4.4	Terms and conditions of use	This element refers to any mention of terms and conditions the requester must agree to for the release of the data to them.	[S] Please read/agree to these terms and conditions for the use of this data.
4.5	Data Use Other	This element includes items that were not specifically covered in the other data use categories	[S] Who is your intended audience for data reporting?
5.0	<i>Miscellaneous</i>	<i>Form element that cannot be categorized elsewhere</i>	<i>not a coding element</i>
5.1	Elements not classified elsewhere	Items that did not fit into other categories.	[S] Is this an emergency request due to a grant deadline [S] Will you be contacting patients?

I also utilized a ‘comprehensiveness’ measure to indicate the breadth of each element: Simple or Extensive. Simple elements were related to a very focused, narrow question on a form (e.g., “Your Name”), whereas Extensive elements had a much broader scope. For example, an Extensive element asked the requestor to “indicate all identifiers (PHI) that may be included in the study research record”, followed by a list of all 18 HIPAA identifiers with a checkbox next to each. Examples of Simple and Extensive elements with respect to the codebook are also shown in Table 3-1. Note that some elements (e.g., codes 1.4, 2.1, 2.2 in Table 3-1) were judged by the team to only be coded using a Simple “comprehensiveness” measure; others could be either Simple or Extensive.

### 3.2.3 Form annotation by two annotators

Each data request form was divided into individual, granular form elements based on the questions asked on each form. For example, one of the forms had a single numbered question comprised of two sub-questions, (1) “describe the data security procedures” and (2) “who will

have access to the data”. These were split into two distinct elements for coding. Each data request form element was then entered into the Coding Analysis Toolkit (Texifter, Amherst, MA). The Coding Analysis Toolkit provided the capability for each element to be shown to an annotator on a computer screen along with the codebook so that all elements could be reviewed and coded efficiently. Using the Coding Analysis Toolkit, two annotators independently reviewed and coded all of the data elements from each of the ten data request forms, including whether each was Simple or Extensive in terms of comprehensiveness. Inter-rater agreement for each form was assessed with the kappa statistic. Coding disagreements were then discussed between the two coders and code assignment consensus was reached.

#### *3.2.4 Content analysis of the ten forms*

From the coded elements on each form I estimated the completeness of information about data needs captured in each form. This was done by assigning a numerical score to each element in the codebook based on the comprehensiveness measure (Simple=1, Extensive=3) that represented the maximum score each item could be assigned. Forms that had  $\geq 3$  Simple elements assigned to the same code were considered to have an Extensive comprehensiveness measure of that code by nature of having multiple elements covering the same concept. I then computed the percent coverage of all possible elements by summing the scores per form and dividing by the total number of possible points a theoretical, all-inclusive form would have had. Finally, I assessed the form elements coded with either code element 2.1 and 2.7 for their ability to capture the salient details for the context and content of data requests in a reliable manner. This may serve as a communication channel between biomedical research teams and query analysts.

### **3.3 Results**

The primary results from the analysis are shown in **Table 3-2**. There was substantial variation in how much detail each form covered and in the elements that were covered. Based on my metric of coverage, the top three forms (A, C, and J) had coverage of 52%, 48%, and 48%, respectively. Form B was much more sparse with only 11% total coverage. In general, forms that had more overall elements (or individual questions) also had better coverage, but the relationship was not completely linear. For example, Form A with the highest percentage of coverage (52%) only had 15 total elements whereas form F had 19 total elements but only 35% overall coverage. Such discrepancy was most often due to either the number of Simple versus Extensive elements used on a form (e.g., fewer elements, but more extensive coverage by each element) or due to many elements disproportionately being related to only a handful of related questions (e.g., one form had four elements dedicated to the funding source).

**Table 3-2** is a summary of the coding analysis performed on the ten data request forms. If a cell is shaded it means that the specific code (row) was found to exist in the specific form (columns A-J). Additionally, the comprehensiveness measure of each element is shown with either an S (Simple, light shading) or E (Extensive, dark shading); those with  $\geq 3$  Simple elements on a form related to a single code were assigned an ‘E’ label even if it was not originally coded as being Extensive. The “Max Score” column represents the total number of points a form element could be assigned as a representation of its comprehensiveness. The total coverage of all elements for each form is shown at the bottom of the table as both a sum and percentage. Note that cells with an Extensive comprehensiveness label were given a score of 3 and those with a Simple comprehensiveness label were given a score of 1. The “# Forms with element” column is a sum of the number of distinct forms that had at least one element on the form that had the respective code in it. For example, nine forms contained code 2.1 (“Study Title/Request”).

**Table 3-2. Summary of the coding analysis performed on the ten data request forms.**

Code	Description	Max Score	Form										# Forms with element
			A	B	C	D	E	F	G	H	I	J	
<i>1.0 Requester Metadata</i>													
1.1	Name	3	E				E		E	E		E	5
1.2	PI, supervisor, department head	3	E	S	E	E		E			E	E	7
1.3	Billing/Administrative content	3			E		S	S				E	4
1.4	Other	1		S		S					S		3
<i>2.0 Request Metadata</i>													
2.1	Study Title/Request	1	S	S	S		S	S	S	S	S	S	9
2.2	Existing/New request	1	S		S					S			3
2.3	Funding source	3			E			S			S	S	4
2.4	Request purpose	1	S	S	S		S	E*	S		S	S	8
2.5	Request type	3		S		E					S		3
2.6	Data sources	3	E					S		S			3
2.7	Data element specification	3	E					E	S	S	S	S	6
2.8	Recurring requests	1	S		S								2
<i>3.0 Compliance</i>													
3.1	IRB	1	S					S			S	S	4
3.2	IRB proof	1	S									S	2
3.3	PHI	3	E				E						2
3.4	Compliance other	3	S				E		S			E	4
<i>4.0 Data Use</i>													
4.1	Internal data sharing	3			E			S	E	S			4
4.2	External collaborators DUA	3			E							S	2
4.3	Public sharing of original dataset	1							E				1
4.4	Terms and conditions of use	1	S								S		2
4.5	Data use other	1								S			1
<i>5.0 Miscellaneous</i>													
5.1	Elements not classified elsewhere	3	S		E	S	S	E		S	E	E	8
<b>Total Score</b>		46	24	5	22	8	13	16	13	10	14	22	
<b>Percent coverage of all possible elements</b>		100%	52%	11%	48%	17%	28%	35%	28%	22%	30%	48%	
<b>Total number of distinct form elements identified for coding</b>			15	5	25	10	9	19	11	11	21	36	

\* This was labeled Extensive because there were 4 distinct Simple elements related to category 2.4; however, this category was considered to be a Simple category. Thus, in this row it still only counts as 1 towards the total score.

Nine out of the ten forms asked about the title of the study/request, and this was the most common question asked across the forms. Other questions were less commonly asked. Only two forms (A and J) explicitly requested proof of study approval from an institution review board, and only one form (G) asked if there was a plan to share the original data set publically. At a category level, four forms did not have a single element related to “Compliance” and three did not have a single element related to “Data Use”. All forms incorporated at least one element related to the categories of “Requester Metadata” and “Request Metadata”, the latter of which is most important for understanding the actual data needs for a request. Within the “Request Metadata”, codes 2.1 (“Study Title/Request”) and 2.7 (“Data element specification”) were determined to be the most relevant for a query analyst to understand the specific needs of the research team. Therefore, I list the specific elements for codes 2.1 and 2.7 derived from all 10 forms within

**Table 3-3.** Some forms asked detailed questions (e.g., five distinct elements coded 2.7 on form F) whereas others asked very basic questions (one element coded 2.1 on form C).

**Table 3-3. Data elements related to codes 2.1 (“Study Title/Request”) and 2.7 (“Data element specification”). These two codes were judged to be the most relevant for a query analyst to understand the information needs of the research team. Note that form D did not contain any elements for which these codes could be applied.**

Form	Code	Simple/ Extensive	Element Header Excerpt	Element Question	Element Options
A	2.1	S	General Reason for Request	Brief description of intent for use of data and/or associated project	Text Box
A	2.7	E	Research Request Reason	Please included as applicable: Request Information (Please include Request Description and if known) - Data Elements, Date Range/Parameters, Sort Sequence, Included Population (e.g. nursing units, DRG codes), Excluded Population (exceptions to the included population), Associated Form (Eclipsys Use Only)...	Document Upload
B	2.1	S	Please provide the following information	I need the new report because...	Text Box
C	2.1	S	Data Type	Full Study Title	Text Box
E	2.1	S	If the purpose of your request is for Patient Care, Education, Administrative, Billing/Payment...complete the	Give a brief description of your project in the space below:	Text Box

following				
F	2.1	S	Data request form	Study Title/Study Idea Text Box
F	2.7	E	Data and/or Records Needed for Research Protocol: Include the following...	Selection Criteria (e.g., all patients with a visit with an ICD-9 780.3x and/or 345.x, English speakers whose age > 50 and age <= 75, etc.) Text Box
F	2.7	E		Counts (if applicable): (e.g., number of patients seen by Firm A, B, C grouped by under 65 and 65 or older) Text Box
F	2.7	E		Dates of Records: (e.g., January 1, 2004 March 31, 2005) Text Box
F	2.7	E		Number of Records: (e.g., 2000 patients with specified diagnosis, 10% sample of patients with diagnosis, all patients admitted thru ED) Text Box
F	2.7	E		List of Data Fields: (e.g., age, race, diagnosis, service area, PCP, etc.) Text Box
G	2.1	S		Complete the following questions
G	2.1	S	What is the purpose of the project or study? Text Box	
G	2.7	S	Describe the data elements needed, such as cancer type (site and histology), geographic location and dates... Text Box	
H	2.1	S	What are the objectives of this project?	What question(s) are you trying to answer? Text Box
H	2.1	S		What problem(s) are you trying to solve? Text Box
H	2.7	S	What are the data requirements?	How much historical data are needed to meet the targeted reporting scope? Text Box
H	2.7	S		How current do the data need to be to support the targeted reporting? Text Box
I	2.1	S	Project Details	Project Title Text Box
I	2.7	S		Please explain below and describe, in detail, the nature of your request to BMI/ICTR. Please do not include any protected health information (PHI) Text Box
J	2.1	S	General Question	Protocol Title Text Box
J	2.7	S		Anticipated Enrollment Text Box
J	2.7	S		Is your anticipated enrollment period greater than a year Y/N/NA

During the coding process I also came across form elements that stood out from the rest, based on the unusual or interesting nature of the questions. These are detailed in **Table 3-4**. This table also contains descriptions based on consensus opinion on why those specific elements were noteworthy. Overall, coding the forms was challenging due to the highly variable manner in which questions were worded. For the ten forms in the analysis, the initial Kappa scores measuring the inter-rater agreement were quite variable, ranging from 0.14 to 0.86 (full list for the forms in the order presented in Table 2: 0.83, 0.86, 0.57, 0.14, 0.64, 0.65, 0.52, 0.55, 0.43, 0.76). Thus, some forms required considerable effort to reach consensus on the final coding of each element.

**Table 3-4. Noteworthy atypical form elements grouped from different forms.**

Element	Why noteworthy
---------	----------------



“I want to write my own SQL queries”	Allows for the possibility of self-service of the complex databases for advanced users. It is unclear what type of guidance or oversight is provided for such requests.
“Please specify what type of Biomedical Informatics Services you are requesting: REDCap, Velos...”	This form combined questions related to data requests and those related to data storage.
““Will you be contacting patients? ___ No ___ Yes. If yes, please justify the need.”	This form seemed to conflate the role of data request fulfillment with that of an institutional review board (IRB). A judgment about the appropriateness of contacting patients is generally handled within the framework of an IRB.
“Principal Investigator: Degree(s):”	It is unclear what the need is for the academic degrees of the principle investigator. It is possible that some institutions limit data access to investigators with a terminal degree.
“What question(s) are you trying to answer” “What problem(s) are you trying to solve”	These questions appear to be aimed at developing a broader perspective about the specific needs and goals of the research term. This information could be useful to help the analyst better understand the context for the data request.

### 3.4 Discussion

Our analysis of research data request forms revealed several interesting findings. Foremost was the substantial variability in the content and comprehensiveness of the forms. The variability suggests that there is no universal or community-based consensus on the optimal way in which a data request form should be designed. The ‘right’ questions to ask and how they should be asked (i.e., expecting simple or extensive answers) are unknown. This could cause downstream consequences including an inability to meet regulatory requirements (e.g., no record of IRB approval verification) or an inability to track research data use in trustworthy ways, as well as problems developing the right queries to meet the fine-grained needs of research teams.

My analysis raised the important question about how well the forms were designed. Being able to answer this question adequately depends, in part, on how well the forms could capture complex data needs accurately and in a reproducible manner. Some forms were very vague or brief about asking researchers what was needed, whereas others asked about specific elements (Table 3-3). Yet I did identify one form that contained questions that seemed to be aimed at

helping the analyst develop a deeper understanding of what data were being sought (**Table 3-4**, row 5) and this may be a useful approach to improve communication.

Because data request forms might serve as the first point of contact between a data management team and a research team, improvement of these forms could provide a positive effect. It has been shown that work focused on redesigning pathology test request forms has been beneficial [179-181], so it may be reasonable to extrapolate that similar benefits could be achieved with redesigned data request forms. The process of developing appropriate data queries from complex user needs can take multiple rounds of refinement [151], but current forms do not appear to be designed to support this process well. It has been noted in the literature that adequately meeting the data needs of investigators for a single request can take a long time [182] so any efficiencies that can be gained would be welcomed.

Data requests forms have been mentioned in the literature [178, 183] (often as a side note) but little attention has been paid to their role in helping investigators obtain data accurately and efficiently. Relative to other form elements, the analysis indicates elements used to elicit the context and content of the requester's data need are lacking. The utilization of frameworks such as PICO (problem/population, intervention, comparison, and outcome) might prove to be advantageous in this setting [89, 184]. With PICO, requesters are encouraged to structure the information need along each of the four dimensions, which could help convey a more realistic description of the request.

Additionally, the effectiveness of forms could likely be improved by providing additional education to investigators about the nature of the data in the systems while at the same time helping to guide researchers through the request form in a more logical manner to ensure that all

the important aspects are covered. It has been observed that familiarity with the database fields by research teams is essential even when working with query analysts [185] but the forms I analyzed did not provide such details. It is possible that some of the forms I reviewed were meant to be accompanied by additional descriptive documents, but I did not come across them in my search. I also did not identify any forms that discussed the issues about data in coded format versus free text narratives, or what types of data are generally found in either of those types of sources.

The forms that comprised my analysis appeared to be constructed to meet the needs of multiple stakeholders (researcher, compliance, IT, etc.). What was surprising, however, is that many forms were unbalanced and placed a greater emphasis on capturing administrative (i.e., bookkeeping) data rather than on the details necessary to execute an effective data query. At the large academic centers generally funded by Clinical and Translational Science Awards, it is likely that many of these data elements already exist in electronic format in administrative databases and might not even need to be transcribed onto a form. Additionally, asking about the degrees of the principal investigator, for example **Table 3-4**, row 4, may be a reflection of a data governance concerns; that is, trainees or temporary employees without terminal degrees may not be granted access to the data at some institutions.

Future work should include a careful analysis of actual data requests in order to be able to map the type of data needs to appropriate elements on existing forms, or to create new form elements when needed. Understanding these needs is a first step towards developing solutions to meet those needs [21]. Cimino *et al.* recently described their work related to understanding complex queries to better develop data retrieval capabilities in the self-service tool BTRIS (Biomedical Translational Research Information System) in use at the NIH [134]. Their goal was

to better empower users to obtain the needed data rather than having to rely on query analysts to retrieve the data for them. Several of their observations could likely improve the design of data request forms, specifically the recognition that the requirements from users “included types of data, constraints on data, and data sets formed from inclusion from multiple data sources” [134].

In addition, future work should seek to quantify the time it takes to complete the elements on a data request form, and if there may be a reasonable tradeoff between form length and the subsequent quality and efficiency of the data extraction. Additionally, observing investigators as they fill out the forms could provide insights about what form elements may be confusing or ambiguous.

From my analysis I am able to make several recommendations about future data request form development: (1) more effort should be made to standardize the types of questions being asked across institutions; (2) whenever possible, forms should de-emphasize the collection of administrative metadata and expand the scope of elements related to the request itself; (3) despite decrease administrative metadata, forms should capture enough information to ensure that regulatory requirements about data use, privacy, and human subjects protection are being met; (4) form design should match the data requirements of investigators--since this is not well described, further research will be needed to elucidate these requirements; (5) because data requirements may vary based on the intended use (e.g., research versus administrative), a ‘one-size-fits-all’ form may not always be ideal, and forms customized to various use cases may be more effective; and (6) forms should provide at least a minimal level of detail to ensure that users understand their selections and options, including details about data sources and data types.

### **3.5 Conclusion**

To serve people I must first understand them. A data request form is meant to be a tool to facilitate an understanding between data owners and data requesters, rather than a burden on researchers serving bureaucratic purposes. This analysis of research data requests forms revealed considerable heterogeneity in form content, both in the breadth and depth of the topics covered. Additionally, most forms over-emphasize the collection of administrative metadata and under-emphasize the collection of important details necessary to communicate a complex data request to a query analyst team. Future work should focus on better understanding the content and nature of data requests from the perspective of multiple stakeholders to help inform the design of new data requests forms that can better capture the complexity of clinical and translational research teams.

## **Chapter 4. Characterization of the Biomedical Query Mediation Process**

### ***4.1 Introduction***

Expanding data access for clinical and translational researchers has long been an important priority for accelerating clinical and translational research. Many institutions employ query analysts to translate data requests from medical researchers into executable database queries. As discussed in the previous chapter, EHR data request forms now in use, provide minimal structure and guidance for the medical researcher to define explicitly their EHR data need. The existing forms lead to a time consuming process where query analysts, who usually have limited medical domain knowledge, must consult with the medical researcher to clarify vague or nonspecific concepts over the course of many emails, phone calls or meetings.

To relieve the burden on query analysts, a variety of data query tools were developed [14, 16, 81, 137, 143]. Notable ones include:

Informatics for Integrating Biology and Bedside (i2b2) [14, 172, 186]

Visual Aggregator and Explorer (VISAGE) [16].

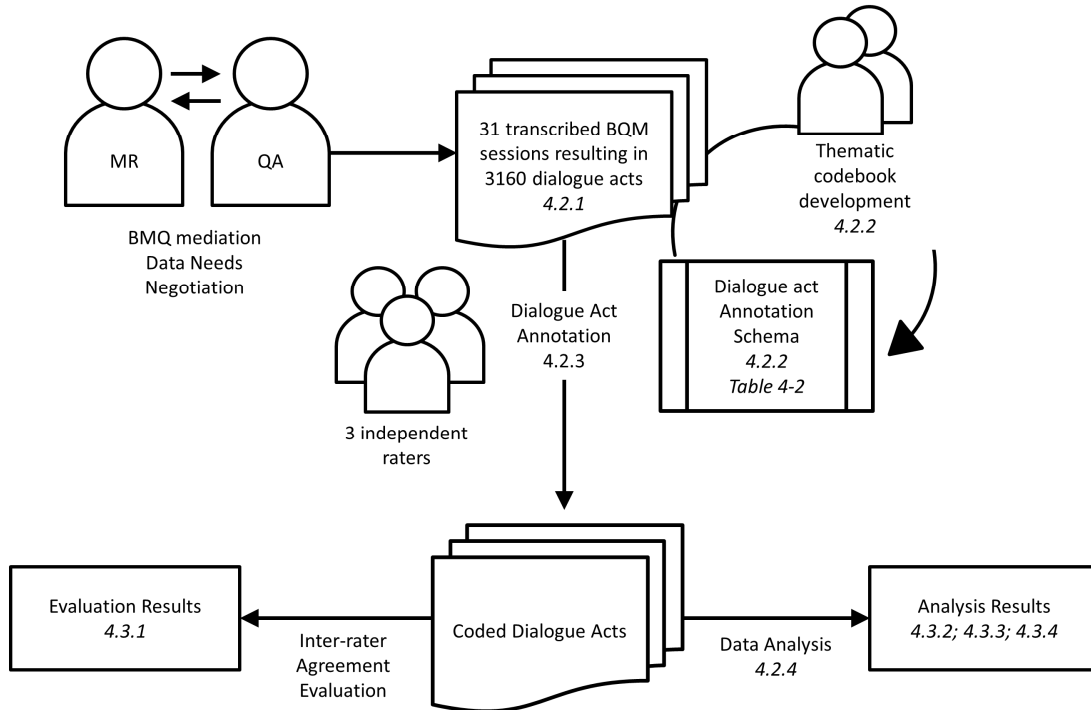
I2b2 enables users to drag and drop concepts to construct queries. The modified Web version, SHRINE, also enables federated queries across multiple databases[14]. Similarly, VISAGE is an ontology-driven visual query interface that recommends concepts for query formulation. Such tools generally require users to specify or select concepts for query formulation, which can be a significant challenge for researchers who usually have limited knowledge of the organization and

coding of the data or the sensitivity and specificity of terms in the database. This problem becomes worse as databases increase in size and complexity.

Ideally, researchers should interrogate databases on their own. As a practical matter it is unrealistic to equip all medical researchers with the systems knowledge which would allow them to efficiently execute such tasks. One approach that could maintain significant researcher involvement is to support the biomedical researchers with computer-based reference interviews. “A reference interview is a conversation between a librarian and a library user, usually at a reference desk, in which the librarian responds to the user's initial explanation of his or her information need by first attempting to clarify that need and then by directing the user to appropriate information resources” [119]. Query analysts, like librarians, often use a negotiation process to comprehend the needs of the researcher [79, 84]. However, at this point, little is known about common steps and their temporal relationships during the biomedical query mediation process. Query analysts often do not have a reference interview template to guide them through the query mediation process. Therefore, this study reports the analysis of the query mediation dialogues between a query analyst and medical researchers and my findings of the characteristics of the biomedical query mediation process. This study extends the work of a poster presented at the 2012 AMIA Fall Symposium entitled, “Analysis of Query Negotiation between a Researcher and a Query Expert”[187].

## 4.2 Data and Methods

Figure 4-1 presents a broad overview of the methods used in this chapter.



**Figure 4-1. Broad BQM content workflow overview in this chapter. The italic numbers indicate the corresponding sub-sections describing work specifics.**

### 4.2.1 Data

Between July 2011 and January 2012, I recorded and transcribed 31 discussions for 22 medical research projects between one query analyst and eight medical researchers at the Columbia University Department of Urology. The Columbia University Medical Center Institutional Review Board approved this study (**IRB-AAAJ8850**). Table 4-1 shows five example dialogue acts. In the context of this paper, a dialogue act is one exchange of speech. I arrived at 3160 dialogue acts for the 31 query mediation sessions.



**Table 4-1. Example Dialogue Acts**

<b>Speaker</b>	<b>Dialogue Act Exchange</b>
Query Analyst	Alright. So we're going to be talking about your study so I guess briefly describe to me what you want to do.
Medical Researcher	So, I haven't really put much thought into it, I just talked with a guy and he suggested that he had talked with umm a pathologist and with other urologists and it would be like very, very interesting to see like after cystectomies see if the urethra was involved.
Query Analyst	Uh huh
Medical Researcher	Umm because that could umm like possibly umm affect you know the outcomes of like long term outcomes of the of the like complications and overall prognosis, that's what he told me. But I haven't like
Query Analyst	So we're looking at the effect of urethral involvement, urethral or ureteral?

#### *4.2.2 Annotation Schema Development*

I used the dialogue acts from 10 randomly selected projects to develop a dialogue act classification schema. I first derived the common tasks of dialogue acts, such as understanding the clinical process, identifying available data, and explaining data characteristics. Then, I grouped the tasks by their corresponding aspect of the query mediation process, such as stages of mediation, data request complexity, and interpretation of requester response. I decided to classify dialogue acts along the “Stages of Mediation” aspect in order to see if temporal patterns of dialogue acts emerged. I iteratively designed and tested a classification schema on sample transcripts and finalized the schema with group consensus among three independent raters. I select two PhD students and one post-doc student from the informatics department.

#### *4.2.3 Dialogue Act Annotation*

I selected two doctoral candidates and one post-doc in biomedical informatics as expert raters. These raters independently annotated all 3160 dialogue acts. Each dialogue act was annotated with at least one classification code. I assessed inter-rater agreement with the kappa

statistic. For dialogue acts with inter-rater disagreement, I reached consensus by accepting the pair-wise consensus between the raters.

#### *4.2.4 Data Analysis*

I used the consensus annotation results for further dialogue flow analysis. I normalized the query negotiation space for the 22 projects to the median number of dialogue acts by either condensing or expanding the conversation sets for the 22 projects. I aggregated the annotated content of the 22 projects into one representation of the negotiation space. I used descriptive statistics and graphs to visualize this space.

### **4.3 Results**

#### *4.3.1 A Dialogue Act Classification Schema for Mediate Query Conversations*

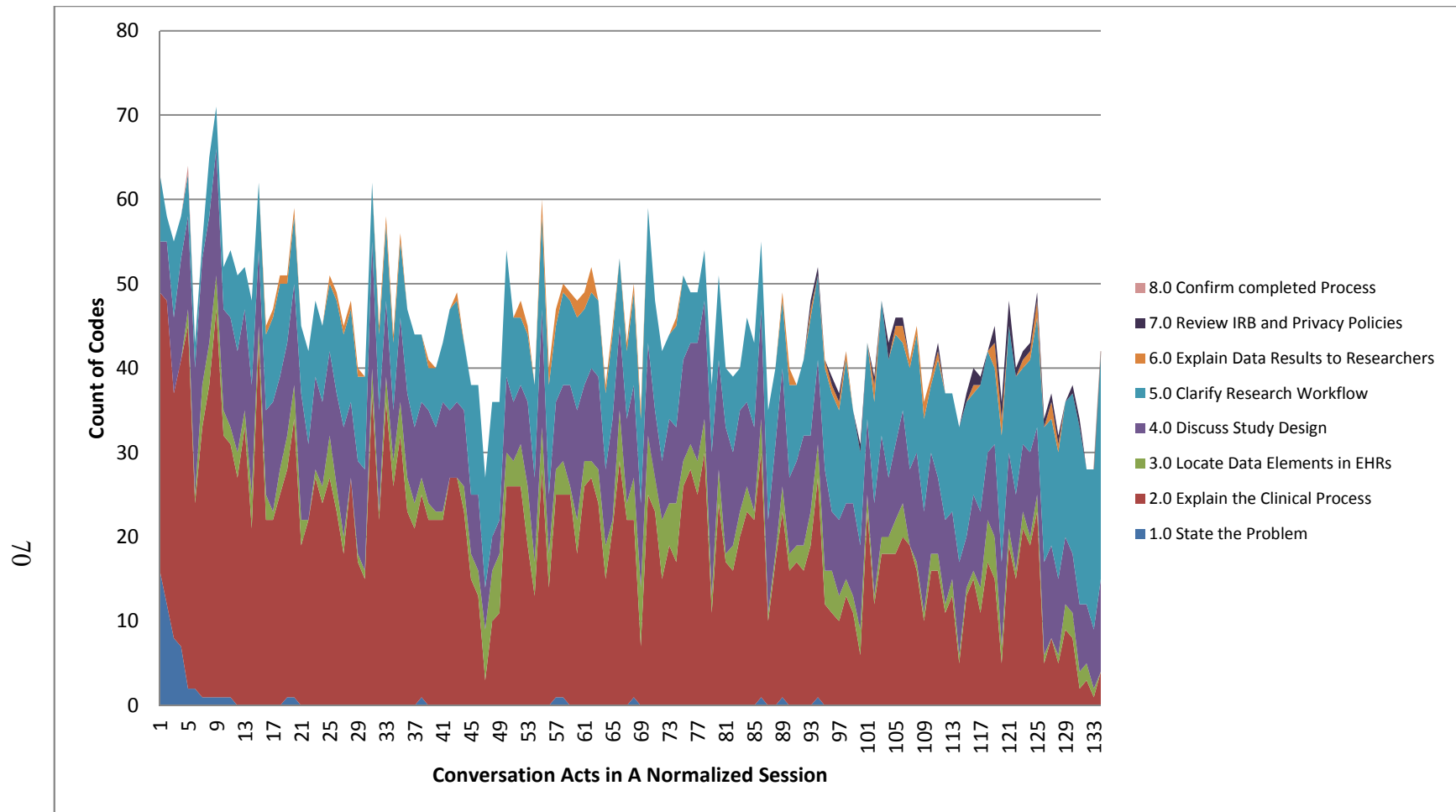
The minimum, median, and maximum numbers of dialogue acts in a project were 27, 134, 323, respectively. The tasks I identified corresponding to the aspect of “Query Mediation Steps” are (1) State the Problem, (2) Locate Data Elements in EHRs, (3) Project Re-Iteration, (4) Discuss Study Design, and (5) Confirm Completed Process. These served as the basis for the coding book displayed in Table 4-2. Tasks were iteratively organized into a hierarchical structure to be used to describe the dialogue acts of the mediation process between the query analyst and medical researcher. The inter-rater kappa score over all the dialogue acts was 0.61.

**Table 4-2. The Classification schema for Dialogue Acts in Query Mediation**

Dialogue Act	Example Dialogue Acts
1.0 State the problem	<i>Alright, So we're going to be talking about your study so I guess briefly describe to me what you want to do.</i>
2.1 Patient demographics	2.21 Initial Diagnosis of disease
2.2 Temporal aspect of the clinical process	2.2.2 Primary treatment of disease
	2.2.3 Follow-up/Surveillance of disease
	2.2.4 Salvage treatment of disease
2.3 Laboratory tests	
2.4 Radiographical studies	2.5.1 Disease confounders and comorbidities
2.0 Explain the clinical process	<i>And then they are diagnosed with cancer after the image?</i>
2.5 Clinical findings	2.5.2 Social history
	2.5.3 Family history
	2.5.4 Clinical stage/Risk assessment/Disease status
	2.5.5 Disease specific/Overall survival
2.6 Surgical procedure	
2.7 Pathology	
2.8 Medical therapy	
2.9 Radiation therapy	
2.10 Other treatments	
2.11 Treatment toxicities, complications and adverse events	
3.0 Locate data elements in EHR	<i>You will have to look in the operative note.</i>
4.0 Discuss study design	<i>Because I want to exclude any disease that could potentially have an effect on the GFR.”)</i>
5.0 Clarify research workflow	<i>It's gonna be rare. So you're probably gonna have to update it as well.</i>
6.0 Explain data results to researcher	<i>So follow-up is last time known alive. So this is corresponding to overall survival information.</i>
7.0 Review IRB and privacy policies	<i>It is expedited because it is de-identified.</i>
8.0 Confirm completed process	<i>Alright. I think I have enough information.</i>

### 4.3.2 *Temporal Distribution of Dialogue Act Classes*

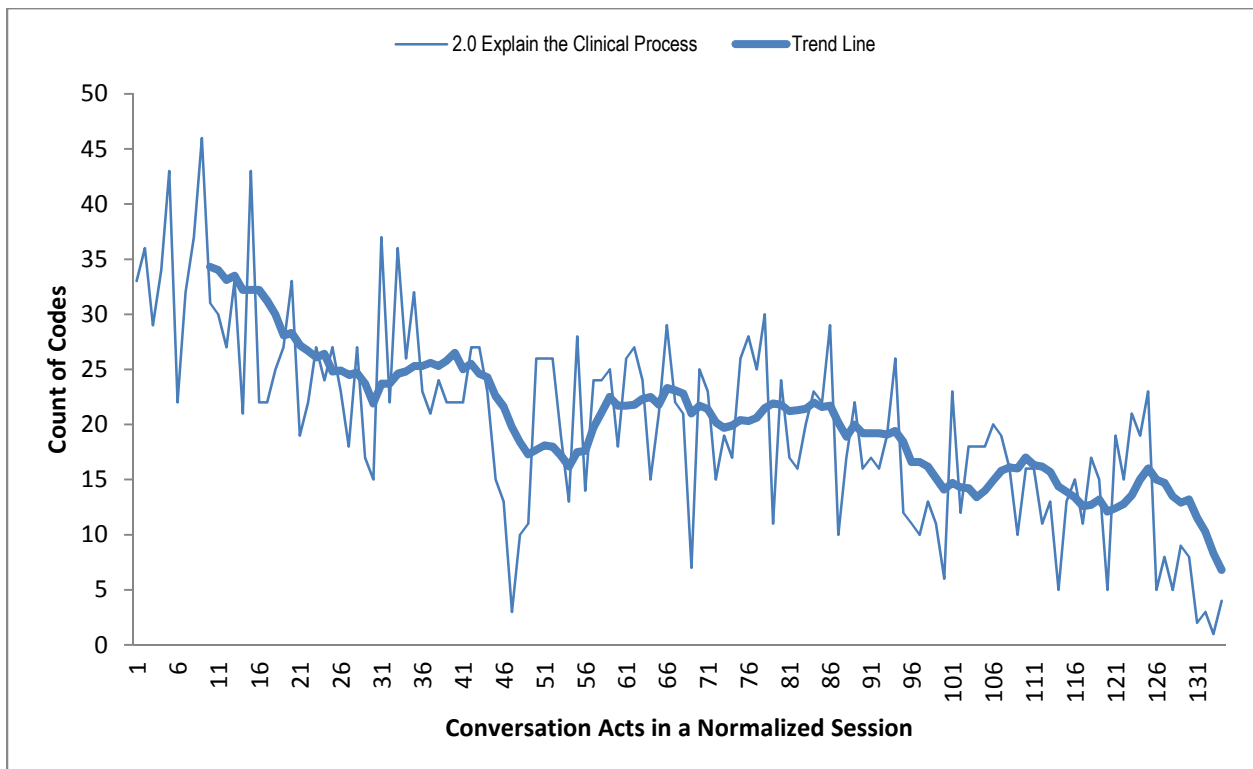
**Figure 4-2** illustrates the broad variety of issues discussed between a query analyst and medical researcher. This figure also represents the aggregate of all 22 projects into one normalized space. The y-axis represents the total number of codes used to annotate a particular conversation act defined by the x-axis. For example, 62 codes were used to annotate the first conversation act of all 22 projects. Throughout the conversation, the majority of the discussion surrounds the clinical process. However, as the conversation concludes, greater attention is drawn toward the research workflow clarification. Additionally, as the conversation concludes, the query analyst and the medical researcher discuss IRB and privacy policy.



**Figure 4-2 Theme river. Temporal Distribution of Dialogue Acts across a Normalized Mediated Query Conversation Session. The y-axis represents the sum of all codes used. The x-axis represents the sequence of the conversation starting with the first dialogue act and ending with the last dialogue act. I normalized all conversations to the average number of dialogue acts for the group of conversations.**

### 4.3.3 A closer look on the Discussion of the Clinical Process

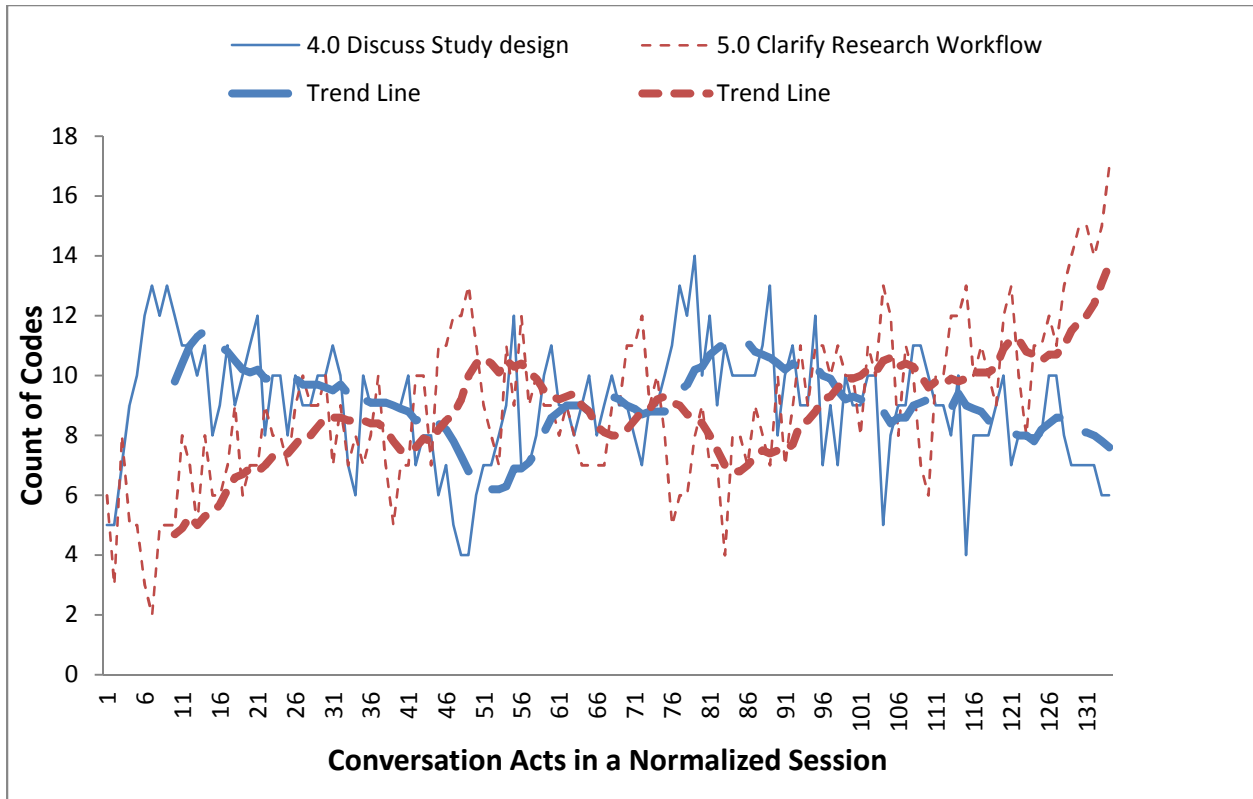
**Figure 4-3** shows how the clinical content of the space is left-skewed towards the beginning of the conversation and trails off at the end. The blue thin line represents the aggregated clinical variable codes, 2.0 (“Explain the Clinical Process”). The blue thick line represents the trend of this variable.



**Figure 4-3. Theme River. Discussion of the Clinical Process over the Course of a Normalized Conversation Session**

### 4.3.4 Temporal Flow of Study Design and Research Workflow Dialogue Acts

**Figure 4-4** shows two classes from the coding schema, 4.0 Discuss Study Design (blue line) and 5.0 Clarify Research Workflow (red line). The start of the conversation supports the development of the study design. The middle of the conversation exchanges these two classes cyclically until the end of the conversation, where research workflow emerges as the dominant class.



**Figure 4-4. Theme River. Discussion about Study Design and Workflow Issues throughout a Normalized Conversation. The oscillation between the two classes of dialogue acts suggest an interactive discussion switching between the theoretical and practical aspects of a project. As the conversation ends, the discussion focuses on the practical aspects to complete the project.**

#### **4.4 Discussion**

As the health record transforms and migrates to the electronic form, data requests for research purposes are likely to increase. The volume of these requests will quickly overwhelm human agents who might remain responsible for querying these data. Non-mediated means for data queries exist but fail to fully satisfy researcher’s data needs. Instead, a mediated data extraction process is needed. However, little is known about the negotiation space between the query analyst and the medical researcher. Zhang et al. briefly describe this process in their “data access paradigm model” [16].

I identified several classes that fall under “stages of the negotiation process.” After several iterations and reductions to the class list, the granularity of the classes was expanded to create the annotation schema for dialogue acts for mediated queries. Although, we had an inter-rater kappa score of 0.61, I do not expect for this coding book to generalize to all other research query mediation processes, but rather to describe the content of this specific negotiation space. This coding schema will allow us to study the progression of conversation and inform the design of a structured interview between query analyst and medical researcher.

The initial illustration of the negotiation space (**Figure 4-2**) is a clear representation of the complexity that exists. I interpret this result as a clear refutation of the idea that data needs assessment is a simple and easy process. A significant amount of query analyst and medical researcher investment is needed to reach an understanding of what the data needs are for any given project. This represents a critical part of the process that occurs in order for a consensus to be reached regarding the researcher’s data needs. The clinical content illustration (**Figure 4-3**) presents a clear view of potential clinical variables that may be presented by the researcher. Difficult clinical concepts, discussed over the course of the conversation, are explored until an understanding is reached and the clinical content drops off toward the end of the conversation space. **Figure 4-4** provides insight regarding how a conversation reaches consensus. It shows how a conversation moves from a theoretical description of data elements to a practical project management discussion. Of particular interest is the middle of the conversation space, where an iterative exchange is occurring between these two classes (Study Design and Research Workflow) of dialogue acts.



#### ***4.5 Limitations***

This study contains two major limitations. First, the study only analyzes the conversation space of one query analyst with medical researchers from one academic department. Furthermore, this query analyst was intensively involved with the department's research program. The query analyst facilitated not just data access but also study design and project management. As such, the conversation space may cover more issues than traditional query negotiations that exist between other query analyst and medical researcher.

#### ***4.6 Conclusion***

To the best of my knowledge, this study represents the first attempt to understand the mediated query dialogues between a query analyst and a medical researcher. The results confirmed that the query negotiation space is not a straightforward translation of a researcher's needs, but rather an iterative process necessary to reach an understanding of the research needs. Query mediation represents a process-based needs assessment and clarification. The results of this study prepare us for the next steps, which are to extract common dialogue elements in mediated query processes and to model the conversation flow in order to inform the design of structured query negotiation, towards the development of an intelligent virtual medical data librarian.

## **Chapter 5. Understanding and Generalizing the Biomedical Query Mediation Process**

### ***5.1 Introduction***

Rich clinical data made available by the Electronic Health Records (EHRs) are invaluable for medical knowledge discovery [23, 54, 188]. However, as such data increases in volume, velocity, and variety, biomedical researchers face significant data access barriers [72], including convoluted regulatory processes [189], inconsistent and limited data quality reporting [50], and opaque data representations [151]. To facilitate data access, data analysts have developed self-service query tools to enable biomedical researchers to navigate and query EHR data autonomously [16, 17, 20, 144]. 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 [61, 152]. They have reduced the barrier for some medical researchers but do always resolve complex queries.

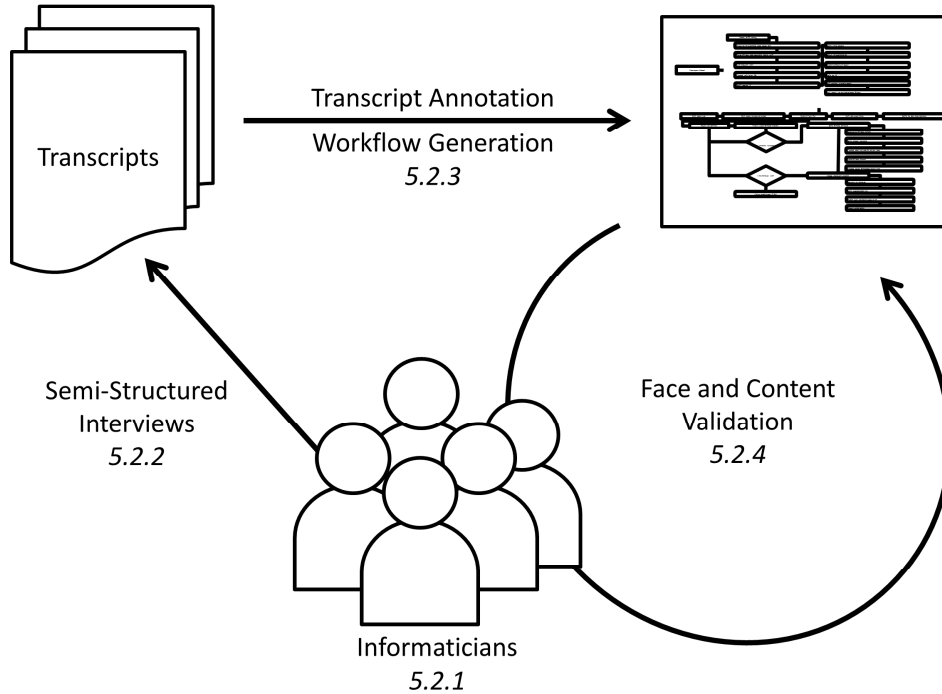
Each medical condition may have multiple data representations in EHRs, which can be structured or unstructured and are collected for billing or clinical care purposes with varying data quality [50]. If structured, the coding schema can be from a broad range of clinical terminologies, such as ICD-9, ICD-10, 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 [190]. Computable representations for a disease may vary across institutions due to phenotype differences in population subgroups and variances in EHR documentation or data representation. Selection of cost-effective EHR data representation for

identification of a cohort with the condition is non-trivial [92, 137] so that query analysts are often indispensable for assisting with the data extraction process [6].

In this chapter, I will identify key tasks in the BQM process and align it with the reference interview approach. Previously, I established a preliminary understanding of BQM processes [23, 151, 187, 191] for one institution [191]. To gain a deeper and more generalizable understanding of the task complexity of the BQM process, I conducted a multi-site cognitive task analysis of the BQM processes to construct a harmonized representation for the BQM process and its common tasks. I utilized the cognitive task analysis protocol described by Clark et al. [192] to yield information about the knowledge, thought processes, and steps for each task [193]. This analysis and its results are reported below.

## ***5.2 Methods***

**Figure 5-1** provides a high-level review of the research process used for this BQM study. Through an iterative process using 11 query analysts, I conducted semi-structured interviews to extract knowledge of the process they use to extract from medical researchers the concepts needed by the medical researcher and translate those concepts into query terms that can effectively identify relevant EHR data. I evaluated the final representation of this process with the same participant pool using a face and content validity questionnaire.



**Figure 5-1. High-level overview of the research process. I initiated the with semi-structured interviews. I annotated the transcripts from these interviews to generate individual and general task flow representations. Finally, I produced and evaluated a harmonized task model. Numbers indicate section headers in this chapter.**

### *5.2.1 Participants*

Between May 2013 and May 2014, I recruited a convenience sample of 11 query 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 5-1 provides additional detail about the query analysts interviewed for this project. All the participants consented to be recorded. I used the interview transcripts for the analysis. This study has received the approval from Columbia University Institutional Review Board (#AAAJ8850).

**Table 5-1. Study Participant Characteristics. (Note: CDW standards for Central Data Warehouse)**

<b>Participant</b>	<b>Site</b>	<b>Data Infrastructure</b>	<b>Training</b>	<b>Title</b>	<b>Years of BQM Experience</b>
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

### 5.2.2 *Semi-Structured Interview*

I conducted a semi-structured one-on-one interview with each data analyst to elicit the details of the BQM process used by each data analyst. The interview questions were organized into three parts. In part one, to establish a general understating of the query analyst's process for BQM, I asked each participant to elaborate on their actions, the goals of those actions, and the knowledge required to perform those actions and the source for that knowledge. This part also prepares the participants for performing a hypothetical BQM in the second part, in which I presented three information need scenarios from published comparative effectiveness research studies [194-196]. I asked each participant to randomly select a scenario, which I decomposed into its information components using the PICO (Patient, Intervention, Control/Comparison, and Outcomes) framework [89]. Next, I 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 query analyst's process with earlier findings [191]. First, I compared the tasks mentioned in part one of the interview to tasks observed in part two of the interview. For new tasks identified in part two, I then asked the interviewee to elaborate on those tasks. Then, I addressed tasks mentioned in the material presented to the query analyst but not identified as an action by the query analyst. I asked the participant if these tasks represented a part of the process they use, 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, I investigated whether or not the presented material served as a reminder of additional tasks the query analyst used for BQM. If so, I asked them to elaborate on those tasks by describing the steps to complete the task, the goal of the task, and the knowledge needed to perform the task. **Appendix A, section 1.1** provides the interview instrument.

To provide a validity check for my interpretations of the interviews, I implemented two member checks [197]. First, during the interviews I reiterated concepts presented by the query analysts to ensure the clarity and completeness of information presented during the interview. Second, after completing transcript annotation, I constructed a process workflow representation of the query analyst's BQM process and contacted the participants to verify if the organization and concepts within each respective representation reflected their view of the BQM process. **Appendix A, section 2.1** displays the 11 participant's individual workflows.

### *5.2.3 Transcript Annotation and Analysis*

To identify a comprehensive list of tasks used to conduct a BQM I performed a thematic analysis by iteratively annotating the interview transcripts [198]. I annotated the eleven transcripts using previously developed task representation [191]. After I annotated all transcripts; I assigned new tasks identified a general description and then grouped them based on that description into a new code. I used the new codes to perform the next round of annotations on the transcripts. This iterative process continued until I could not generate new codes. After the final annotation round, I constructed the workflow process for each query analyst. I used these representations to perform a second round of member checking with each participant. Through e-mail communication, I presented each participant with his/her process 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?" I augmented the individual process flow models according to the query analyst's input.

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

Additionally, for each level I identified the knowledge needed to perform and the expected outcome for that task. To contextualize the hierarchical task list, I created a harmonized process model from all the individual process models. I assessed the face and content validity of the generalizable task list and process flow among the study participants.

#### *5.2.4 Evaluation*

I presented the consolidated model to the study participants. I asked them to complete a 29-item questionnaire developed using methods described by Lawshe et al. [199]. I assessed face validity by measuring the models representativeness of BQM and usefulness for 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 query analyst conducting BQM.” I considered the model to have face validity if both dimensions obtained a median score of seven or greater.

I measured content validity for each BQM task by having study participants rate each task as essential, useful, or non-useful. I 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 [200]. I deemed tasks semi-valid if at least half of the query analysts rated the task as essential. Tasks that did not meet either of these criteria were non-valid. Furthermore, I asked each participant to assess if the task was automatable.

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, I compared my model to the reference interview process.



### **5.3 Results**

#### *5.3.1 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 tasks, the activities for performing each task, the steps for executing each activity, the knowledge needed to perform the task, and the expected outcome of that task. **Table 5-2** and **Table 5-3** display the corresponding tasks needed to complete the preparation and needs negotiation phases, respectively.

**Table 5-2. BQM preparation phase tasks/activities. These are the tasks the query analyst uses to prepare for the face-to-face (F2F) meeting with the medical researcher (MR). In addition, 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 query 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 Query 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 query analyst/team	Heuristics	<b>Ensure the request is matched with the appropriate resources</b>

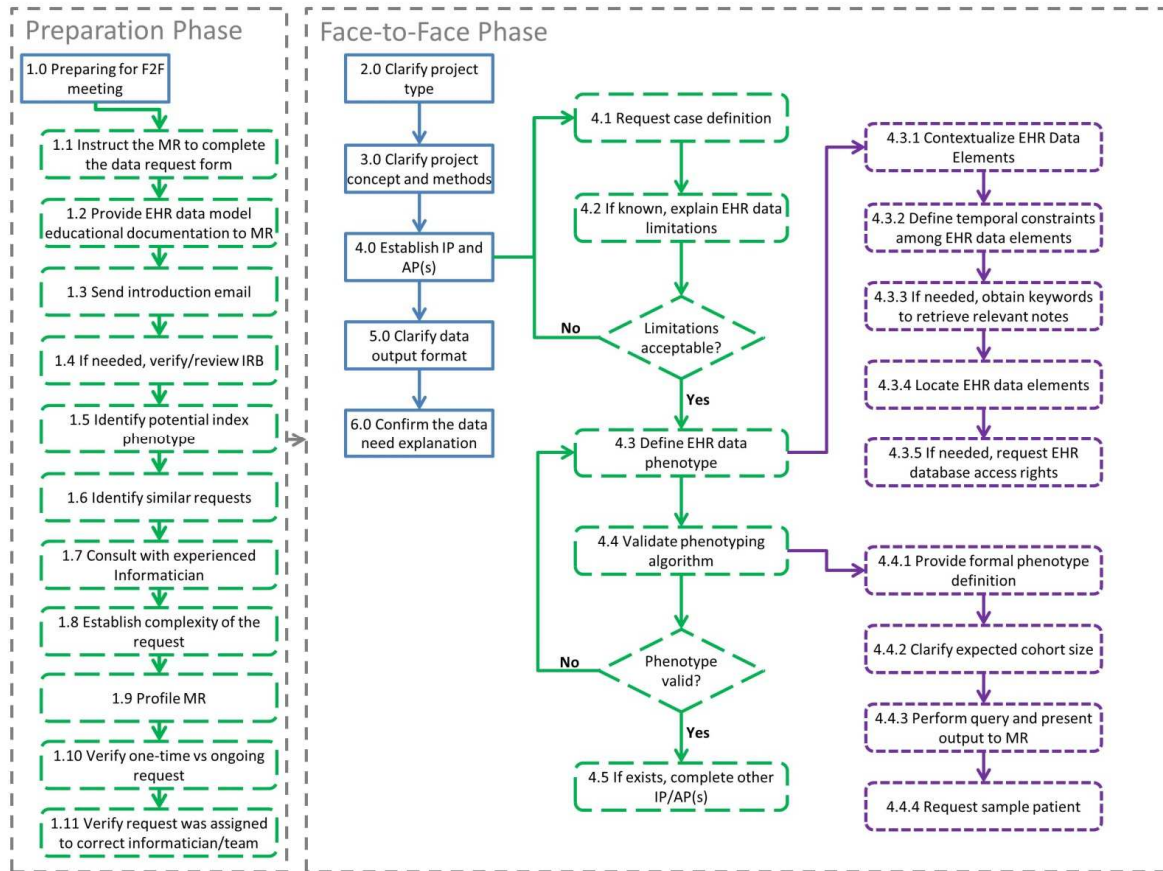
\* Institutional Policy – Institutional protocol for submitting an EHR data request;  
 Internal Documentation – Institutional EHR data model;  
 Heuristics – empirical knowledge gained from BQM practices;  
 Request Tracking System - Internal work management application

**Table 5-3. Face-to-face Task/Activity/Step. These are the tasks used by the query analyst with the medical researcher (MR) to arrive at an understanding of the MR’s data need. In addition, each task is described with the knowledge required to complete the task and the expected outcome for that task. 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
		4.3.1	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
		4.4.1	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	Validate phenotyping algorithm	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 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 query 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

In the BQM preparation phase the query 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 medical condition(s) used to establish the patient cohort. Similarly, associated phenotypes are EHR data representations for any medical condition or concept needed for the proposed study, but do not represent the patient cohort. For example, if I 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. In the example of prostate cancer, I consider the outcome of prostate-specific antigen recurrence an associated phenotype and define it 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 from other element attributes. 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. To visualize the hierarchical task model, **Figure 5-2** displays the BQM workflow process.



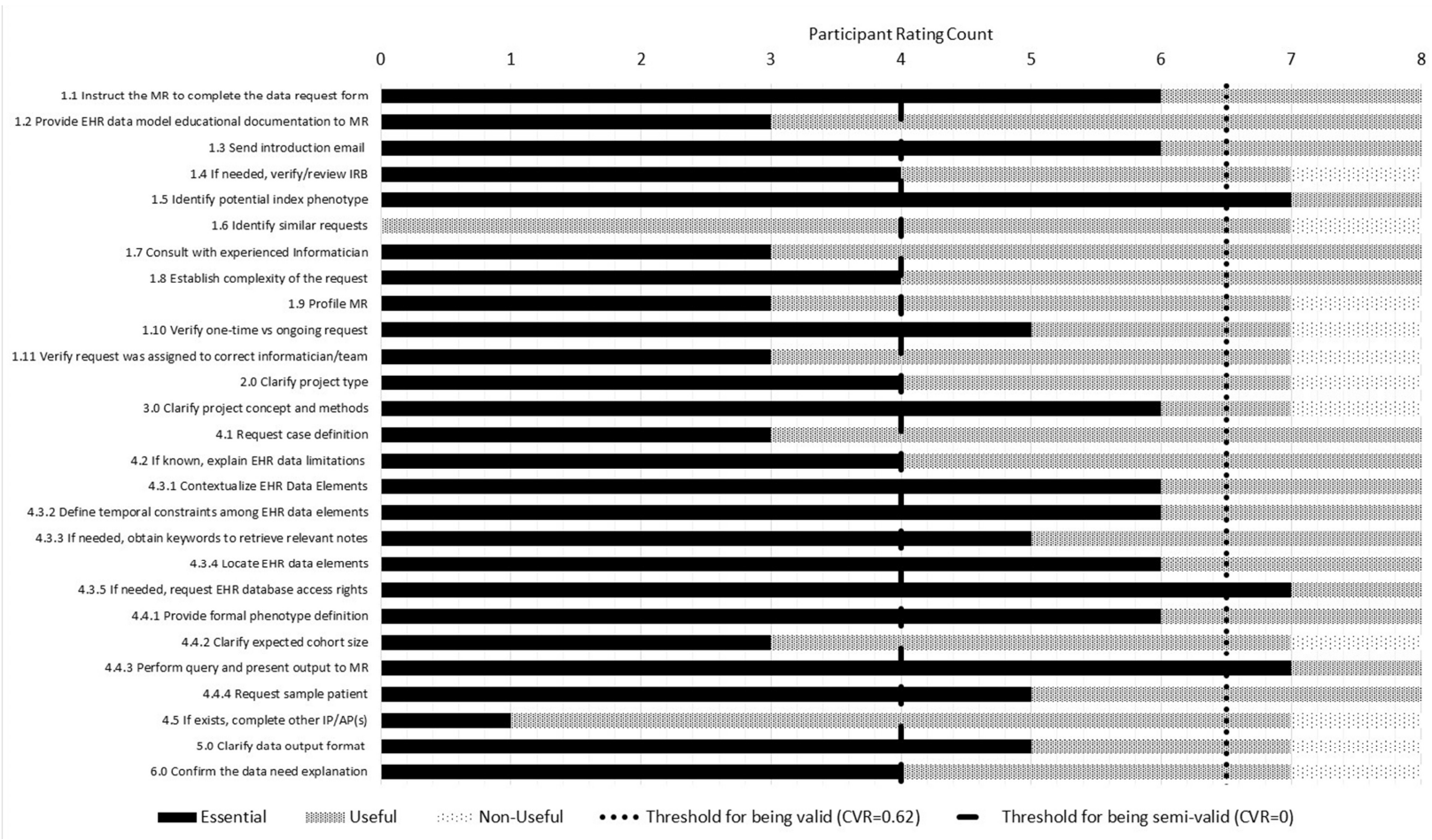
**Figure 5-2. Generalizable BQM task model displayed as a process workflow. The major phases are separated by dashed boxes. Tasks, activities, and steps are represented by solid, dashed, and dotted boxes, respectively. IP – Index Phenotype; AP – Associated Phenotype; QA – Query Analyst; MR – Medical Researcher**

### 5.3.2 Face and content validation

Eight of eleven participants completed the 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 query analysts rated the tasks as essential. Eight tasks are non-valid: 1.2 - Provide EHR data model educational documentation to medical researcher, 1.6 - Identify similar

requests (for query reuse), 1.7 - Consult with experienced Query analyst, 1.9 - Profile medical researcher, 1.11 - Verify request was assigned to correct query 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 5-3** displays the results of the content validity evaluation.



**Figure 5-3. Content Validity Ratio (CVR) results for all tasks in the Preparation and Face-to-face phases. Only three tasks: Tasks 1.5, 4.3.5, and 4.4.3 met the minimal threshold to be have content validity (CVR=0.62). Nineteen tasks were considered semi-valid as at least half of the participants labeled the task as Essential (CVR=0).**



### 5.3.3 *Comparison to the Reference Interview*

Table 5-4 displays the five filters of the reference interview. For each filter, I compare tasks that best represent the objectives of the filter. I present an example quote from my interviews associated with the task and then contextualize the task with a potential action that the query analyst may use to complete that task.

**Table 5-4. Alignment of Reference Interview (RI) and BQM Tasks**

Concept	BQM Task Code	Quote	Contextualized
<b>RI</b>			
1. Determination of Subject	3.0	<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>
2. Object and Motivation	2.0	<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>
3. 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 medical researcher experience with both research and using EHR data for research, the query analyst assigns the medical researcher as novice or expert. This helps establish an expectation for the time and effort needed for the needs negotiation.</i>
4. Relationship of inquiry with file organization	4.3.1	<i>I 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 I 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>
5. 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>

## **5.4 Discussion**

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

### *5.4.1 Implications*

Our study established that a cognitive task analysis could be used to better define and understand a data needs negotiation. That is, I was 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 [201]. By doing so, the greatest contribution from this effort is making explicit content that has been largely the implicit. The model represents an amalgamated perspective of the tasks used to facilitate BQM and the knowledge needed to complete those tasks. I 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. I believe this representation could provide query analysts a knowledge resource to better document and describe the work they perform. It may serve the novice query analyst as a guide to navigate the BQM process and for expert query analysts, may provide strategies for best serving specific medical researcher clients. This knowledge provides a framework for introducing process re-design to increase BQM efficiency and ultimately improve EHR data access for medical researchers.

## 5.4.2 *Expected findings*

### 5.4.2.1 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 the study sample of query analysts, and may be due to the small sample size. Nevertheless, my model included all tasks that any of the query analysts may perform during BQM and, as such, some tasks may have been unfamiliar to the majority of the participants.

### 5.4.2.2 Related models of information seeking

My analysis found that most query 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 the query analysts utilized task 3.0 – “Clarify project concept and methods” to understand the context of the project. This allowed the query analysts to minimize assumptions and frame the data in the context of the intended application. All query analysts performed task 4.0 and established the index/associated phenotype. The iterative nature of this task allowed the query analyst and researcher to focus on a clear definition of the EHR data need. This is analogous to other information seeking studies [202] and models, the ASK hypothesis [2] and Berrypicking [28].

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 [203]. 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 query analyst and researcher. Their findings are consistent with my model. For example, the iterative loop associated with task 4.0 – “Establishing the index/associated phenotype(s)” mirrors the iterative discussion between the query analyst and researcher observed by Hoxha et al.

Finally, I posited that the reference interview serves as a gold standard for the elicitation of an information need from an information seeker. I aligned the goals of the tasks within the BQM model with those of the five components of the reference interview. Table 5-4 compares and contrasts the five components of the reference interview with tasks from my 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 query analyst to include all the corresponding tasks of the reference into their respective process.

#### *5.4.3 Unexpected Findings*

Participants were able to identify only a small number of tasks that could be automated. This may represent their perception of the complex decisions they make for each task. It may also reflect their reluctance to admit they could be replaced by an application. Participants identified task 4.2 – “If known, explain EHR data limitations”, as semi-valid. This surprised me, as my previous understanding of data quality concerns were resided by the medical researcher and not the query analyst. This is promising, as I believe BQM presents an opportunity to share knowledge about EHR data limitations between query analysts and researchers. Furthermore, Weiskopf et al. developed a guideline for data quality assessment that can potentially be used to facilitate data exploration and data quality awareness during the information needs negotiation

[204]. Task 4.2 compliments the data quality assessment guidelines and in fact could serve as a point for incorporating these guidelines into the BQM process.

#### *5.4.4 Limitations*

The work in this chapter has three limitations.

First, the semi-structured interviews may have contained biases. Specifically, the process workflow representations (“representation”) generated for the process used by each of the 11 participating query analysts may be biased toward my initial description of BQM [191]. In an effort to avoid this, each participant underwent several rounds of review and acceptance, both during and after the interview to confirm a fair representation of each query analyst’s process.

Second, I only used one coder to annotate the interview transcripts and generate new concepts. As such, results may be skewed to the interpretation of the coder. To address this issue, I implemented two instances of member checks to ensure the internally validity of my interpretations from the interviews aligned with the ideas held by the query analysts interviewed.

Third, I only studied one stakeholder of the BQM process, the query analyst, without exploring the tasks and challenges involved on the biomedical researcher side.

### **5.5 Conclusion**

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

## **Chapter 6. Data-driven Concept Schema for Defining Clinical Research Data Needs**

### ***6.1 Introduction***

The rich data made available by EHRs represents a promising resource for accelerating clinical and translational research [23]. However, medical researchers face significant barriers to accessing EHR data including

- (1) articulation often abstract and vague data needs
- (2) poor understanding of data details
- (3) inability to map these needs to fine-grained, contextual lower-level data representations

Common data elements (CDE) [205-207] serve as a bridge to map medical researcher's data needs to EHR data representations. CDEs are developed for standardizing research data collection and retrieval. 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 a query analyst and engage in BQM [151, 191]. 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 a 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 commensurate 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 [63]. 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 [189]. 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 [141]. Furthermore, templates are used to standardize the information collecting process, thereby increasing the quality and the efficiency for specifying information dense summaries [208]. The best template example in the medical domain is the PICO framework [89], where P stands 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 [208] and improves information retrieval accuracy [105, 112]. The PICO framework has been shown to be effective at improving the resolution of information needs for medical literature [110, 141]. 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 [133] 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

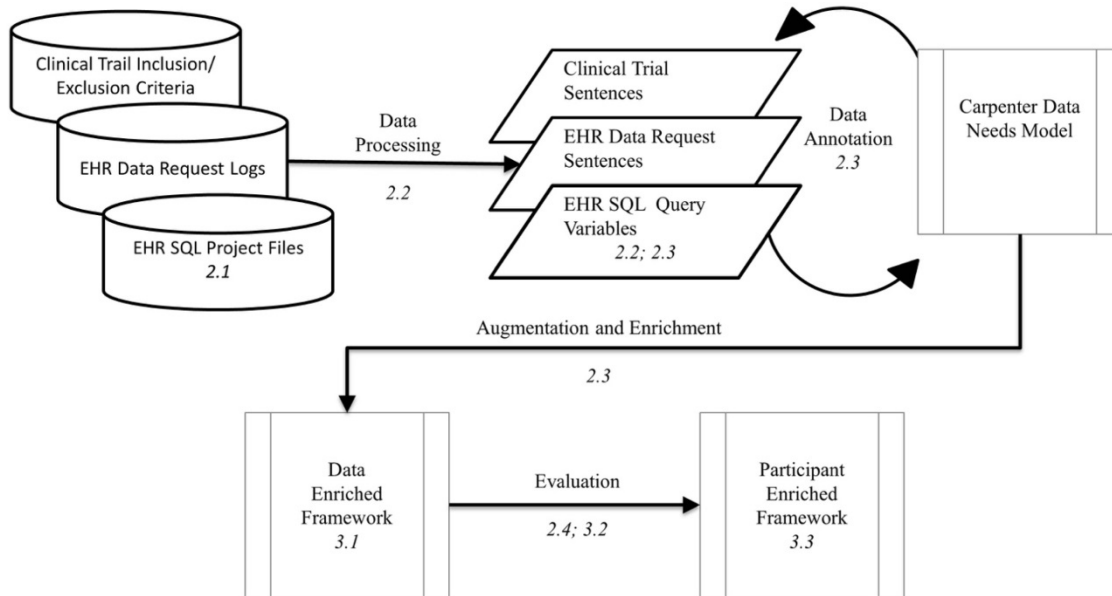


their temporal and association relations. The framework also represents the iterative nature of the cancer care continuum [133]. The framework provides a semi-granular representation of data needs yet remains compact enough to achieve an efficient representation of a complex information space. If able to extend beyond cancer, this framework may serve as a template for defining data requests for medical research in general.

This chapter will focus on the use of 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. In this study, I have validated and extended the Carpenter framework utilizing three data sources that represent researchers' data needs in disparate medical domains.

## **6.2 *Methods***

**Figure 6-1** illustrates the study design. Three data sources were processed and analyzed to identify discrete variables for specifying research data needs. I used the Carpenter framework as the starting point for data annotation and iterative schema enrichment. I performed an evaluation with eight multidisciplinary medical researchers and refined the resulting class schema for representing generic clinical research data needs accordingly. This study received approval from Columbia University Institutional Review Board.



**Figure 6-1. Research Design. The corresponding section from both the Methods and Results sections are noted with an *italicized number*.**

### 6.2.1 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 Columbia University Medical Center’s 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). I conclude that these three data sources provide a rich and complementary representation of medical researchers’ data needs. Table 6-1 provides a detailed description of the datasets used for this project. The next section will discuss the sampling strategy for each data source.

**Table 6-1. Datasets used in this study and their characteristics**

<b>Data Source</b>	<b>Source Quantity</b>	<b>Annotation Quantity</b>	<b>Medical Domain Representativeness</b>	<b>Use of Data</b>
<b>Clinical Trial Inclusion/Exclusion Criteria</b>	181,356 Studies	1000 Sentences	No domain selection	Cohort identification
<b>EHR Data Request Logs</b>	432 Requests	897 Sentences	No domain selection	Cohort identification and dataset generation
<b>EHR SQL Queries</b>	204 Projects	1,445 Variables	Urology domain	Dataset generation for retrospective CER

### 6.2.2 Data Sampling

To obtain a representative sample of sentences from the clinical trial eligibility criteria, I extracted 2,729,525 sentences from 181,356 Clinical Trials downloaded from the public Clinicaltrials.gov on 2/12/2015. I annotated the concepts in these sentences with UMLS semantic types using a previously published method [209]. Using the K-means clustering algorithm [210], I divided all the enriched sentences into 27 classes. To cover sentences from these classes evenly, I sampled 1000 sentences evenly from these clusters for further annotation. For the EHR data requests logs, I randomly sampled 432/1200 data requests submitted to 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, I used the SQL transact code associated with the 204 research projects performed at Columbia University’s Department of Urology over the course of five years (2008-2012). For each project SQL code, I selected the “SELECT\* FROM\* WHERE\*” statements and isolated the “SELECT \*” clause for annotation.

### 6.2.3 *Dataset Annotation and Analysis*

I annotated the datasets. I have 10 years of experience conducting research and 6 years of experience resolving medical researchers' data requests. I did not ask independent annotators to annotate the datasets and measure inter-rater agreement for the following reasons. First, my 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 my 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 [211] . 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, I elected to utilize a data-driven approach rather than a human-driven approach to achieve my goal. Therefore, the 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 terminology can be UMLS for clinical research eligibility criteria or ICD-9 codes for EHR SQL queries. Therefore, the classification step performed by the annotator was informed by the rich semantic information in the UMLS concepts, including UMLS semantic types and concept definitions, rather a completely subjective process. Third, this annotator strictly followed a transparent

systematic process to perform the annotation, as suggested by the following article on improving the rigor of qualitative study [212]:

1. Recognize all the concepts in the sentences/SQL variables and map each concept to a class in the Carpenter framework semi-automatically using a previously published method.
2. Tag the sentence/SQL variables with the class(es) identified from the Carpenter framework.
3. If a concept within the sentence/SQL variables is unable to be tagged with a class from the carpenter framework, label that sentence/SQL variables with “new class.”
4. Group all “new class” sentences/SQL variables and perform a thematic review to name the “new class”.
5. Review the Carpenter framework and insert new concept classes in the right positions in the hierarchy.
6. Repeat steps 1-5 until no new classes can be identified or relocated in the hierarchy.

I augmented the Carpenter framework by editing a preexisting class, adding a new class, deleting an unused class, or moving a class in the hierarchy. For example, the original class, *Comorbidities*, was expanded with the following subclasses: *Medical/Disease History*; *Medical/Surgical/Radiation Treatment History*; *Medical Device Implant*; *Current Medications*; and *Current Treatment/Experimental Trials*. **Appendix A, section 1.1** provides the details of the augmentation.

#### 6.2.4 Evaluation

I assessed the enriched schema using selected measures proposed by Mehmood et al.: concept class coverage, schema generalizability, class preservation, understandability, and structural correctness [213]. Each evaluation metric is further described in Table 6-2.

**Table 6-2. Evaluation metrics and their definitions**

<i>Metric</i>	<i>Definition</i>
<i>Class coverage</i>	The percent of concept classes representing clinical research data needs included
<i>Schema generalizability</i>	The median percentage of class coverage across disease domains of our evaluators
<i>Class preservation</i>	The percent of classes from the original framework included in the enriched schema
<i>Understandability</i>	Evaluator's assessment of the clarity of the classes within the enriched schema
<i>Structural correctness</i>	The validity of the semantic relations and hierarchical relations among classes

The evaluation consisted of two parts. The first part evaluated class preservation through a direct comparison of the enriched schema to the original. The second assessed the metrics of concept class coverage, schema generalizability, understandability, and structural correctness through a semi-structured one-on-one interview with eight clinical researchers (Table 6-3) identified through a convenience sample. Each interviewee was consented for participation and the interviews were recorded. The semi-structured interview was conducted in three blocks. **Appendix B, section 2.1** contains the interview material used for this evaluation. First, an introduction section designed to establish the researcher's area of research, their cumulative experience conducting research, and the number of data request they submit in a year. Next, I presented each participant with a recent study from his or her lab and asked the participant to list the major types of data needed to conduct the study. Then I introduced the enriched schema to the participant and asked them to map the concepts they listed to the classes in the enriched schema. For example, if the participant listed 10 major types of data needed to conduct the study and they were only able to map these data to seven of the concept classes, and then I would

calculate class coverage for this participant at 70% (7/10). To evaluate schema generalizability, I calculated the median of eight participants' class coverage. During the concept mapping exercise, I instructed the participants to “think-aloud” their actions and decision-making processes. I followed this with a set of questions addressing difficulties they may have had during the mapping process. I used the transcripts from the think-aloud process and follow-up responses to assess the evaluation metric, understandability. In the third block of questions, I evaluated the metric, structural correctness. Member checking was performed to confirm my interpretation of the evaluation results with each participant. Moreover, augmentations to the enriched schema were made to accommodate constructive feedback I received during the evaluation process.

**Table 6-3. Evaluator characteristics**

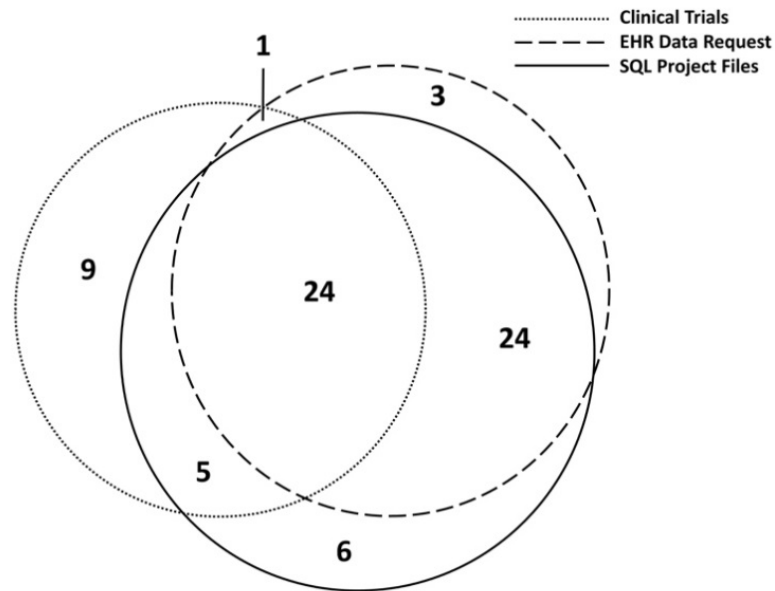
<b>Participant</b>	<b>Department</b>	<b>Title</b>	<b>Research Expertise</b>	<b>Years of Research Experience</b>	<b>Number of data requests submitted/year</b>
1	Hematology Oncology	Fellow	Quality Improvement	3	3
2	Emergency Department	Emergency Medicine Director	EHR health practice research	10+	5
3	Pediatrics; Infectious Disease	Professor of Clinical Pediatrics	Observational Epidemiology	28	5
4	Medicine; Behavioral Cardiovascular Health	Assistant Professor of Medicine	Prospective and Retrospective Studies	4	3
5	Medicine; Digestive and Liver Diseases	Assistant Professor of Medicine	Retrospective, Epidemiology	16	5-10
6	Urology Department	Professor and Chair	Prospective and Retrospective	15	52+
7	Medicine; Naomi Berrie Diabetes Center	Professor of Clinical Diabetes, Medicine and Pediatrics	Clinical Trials and Retrospective	16	20+
8	Division of Colorectal Surgery	Chief	Retrospective Outcomes Research	14	52

## 6.3 Results

### 6.3.1 Data-Enriched Schema

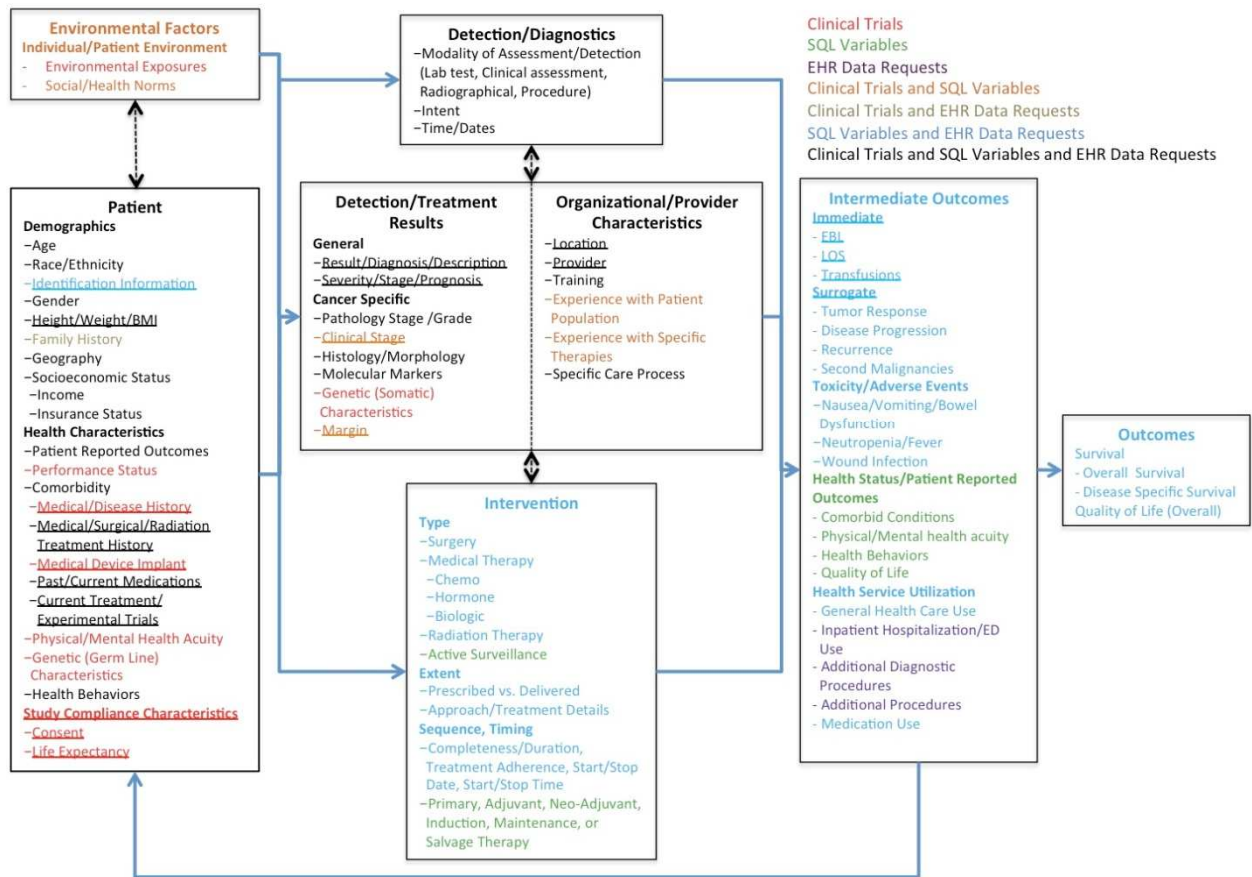
I identified 1064, 1970, and 1892 concepts from the clinical trial eligibility criteria, the clinical research data requests, and the SQL statements, respectively. These concepts were mapped to 72 classes in the enriched schema.

**Figure 6-2** is a Venn diagram displaying the union and intersections for the 72 classes across the three data sets. **Figure 6-3** displays the data enriched schema. The notable structural change was to associate “Organizational/Provider Characteristics” with “Detection/Diagnosis” and “Intervention” instead of the “Patient” section. In **Appendix B section 3.1**, I provide definitions and examples for the 72 classes presented in **Figure 6-3**.



**Figure 6-2.** The Venn diagram displays how the concepts from the three respective data sources mapped to the classes within the data enriched schema. The three datasets share coverage for 33% (24/72) classes represented in the data enriched schema.





**Figure 6-3. The data enriched schema. The blue directed edges represent the temporal process as the patient moves through the care continuum. The cyclical nature of this graph implies the patient can re-enter the care cycle. The bi-directional edges indicate an association between the sections. New additions to the schema are underlined, and color-coded classes correspond to the dataset that contains the class.**

### 6.3.2 Evaluation

With regard to class coverage, the schema contains 89% (73/82) of the concept classes used by the participants. For generalizability, the schema accurately identified concept classes from diverse medical domains with a median accuracy rate of 95% (60-100%). For the metric of preservation, Table 6-4 displays the schema's preservation of the entities from the Carpenter framework. Overall, 79% (70/89) of the entities within the enriched schema originated from the original Carpenter framework. **Table 6-5** shows the participant breakdown of concept

preservation. The participant from Pediatrics, infectious disease reported the lowest class coverage (60%).

**Table 6-4. Class Preservation. This table compares the number of sections, classes and edges from the original framework to the data enriched schema. I calculate degree of preservation as the ratio of preserved entities over the total number of entities from the data enriched schema. Both major elements of the Carpenter framework, sections and the directed edges were maintained. However, the enriched schema deviated from the granular details of the original framework.**

<b>Elements</b>	<b>Carpenter Framework</b>	<b>Data-enriched Schema</b>	<b>Preserved Elements</b>	<b>Degree of Preservation</b>
Sections	8	7	6	86%
Classes	63	72	57	79%
Directed Edges	8	7	6	86%
Bi-directional edges	4	3	1	33%
<b>Total</b>	<b>83</b>	<b>89</b>	<b>70</b>	<b>79%</b>

**Table 6-5. Participant breakdown for generalizability and class coverage**

Participant	Department	Concepts Identified	Concepts Mapped	Class Coverage	Participant comments for concepts not mapped to classes within the data-enriched schema
1	Hematology Oncology	10	9	90%	<i>No class described the cost associated with tests</i>
2	Emergency Department	11	8	72%	<i>No class covered "Diet Status" for patients; The one concept existed as classes, but the participant didn't map the concept ("Provider Behavior"; the last concept was a complex concept assessing if an order was part of a larger set of orders.</i>
3	Pediatrics; Infectious Disease	10	6	60%	<i>This participant provided concepts from a study assessing secondary preventative options for a primary treatment (e.g. The success of peri-op prophylaxis for patients undergoing cardiac treatment). The schema did not provide a class that described other health service interactions on a patient treatment regimen. This case highlights a theme of studies our schema would be unable to adequately represent.</i>
4	Medicine; Behavioral Cardiovascular Health	9	9	100%	<i>NA</i>
5	Medicine; Digestive and Liver Diseases	11	11	100%	<i>NA</i>
6	Urology Department	11	10	90%	<i>The concept listed was a set of lab tests, pre and post-operative treatment Creatinine values, that do not represent a disease status, but a health status measuring collateral damage of a primary treatment choice. While this potentially could be mapped to some classes within our schema, the association could be considered vague.</i>
7	Medicine; Naomi Berrie Diabetes Center	10	10	100%	<i>NA</i>
8	Division of Colorectal Surgery	10	10	100%	<i>NA</i>
Generalizability				95%	



Table 6-6 presents the subjective metrics evaluated. For each metric, I identified themes derived from the interviews. I organized themes into quotes that support or oppose the data-enriched schema and provided counts for the number of times at which those themes occurred. In addition, **Table 6-6** provides representative quotes for each theme. For the metric of understandability, the majority of the positive sentiments surrounded the organization of the classes and the schema's effect to stimulate additional medical concepts needed for research. However, the participants found significant ambiguity in the enriched schema; they described the enriched schema containing overlaps between classes from different sections. Even though the participants were able to map 89% (73/82) of the concepts they identified, they still noted missing classes. For structure, the majority found the temporal and interaction relationships between the sections of the enriched schema to be sound, with the exception of the temporal edge conveying the iterative nature of the care continuum.

**Table 6-6. The subjective metrics of understandability and structure. Within each metric, I ordered themes based on occurrence in the interview transcripts. Additionally, I provide a definition and contextualized quote for each.**

Metric	Dimension	Definition	Sentiment	Sentiment Freq.	Example Quote
<b>Understandability</b>	Ambiguity	Difficulty in differentiating classes from different sections	Oppose	21	<i>“My main question is, I feel like the middle part represents what you are studying, such as like a diagnostic test, that’s fine, but there is some overlap conceptually between what is the test you are studying versus the test result and I think that is what informs the eligibility.”</i>
	Precision	Applicability of the concept schema to data needs	Oppose	16	<i>“Interventions as two different ways, a risk factor or as a management, like the way people were randomly assigned. I think this is great, but very specific to cancer”</i>
			Support	3	<i>“Prescribed vs Delivered, well that’s what I was getting at, ordered vs. Delivered, So prescribed is the provider order and delivered is the administration record”</i>
	Organization	Alignment of the concept organization with user conceptualization	Support	11	<i>Organizational/Provider Characteristics, oh, location is there, I found it.”</i>
			Oppose	3	<i>“So, first I was a little confused as to where to look first, cause the first thing to hit my eye was the ‘Environmental Factors’, and I was looking for the patient stuff, but I found it.”</i>
Generalizability	Generalizability of the concept schema to real experience	Support	4	<i>“I think it’s great by the way, congratulations, I think that everything I could do could go into these buckets, but I think it makes a lot of sense, there is nothing loco here.”</i>	
<b>Structural Correctness</b>	Temporality	The temporal relationships among concept classes	Support	5	<i>“The overarching flow, is what you would predict as we are all time orientated, I started as a patient and now I am dead”</i>
			Oppose	3	<i>“Well at first glance I have no idea, there is directionality of the arrows...yeah not clear”</i>
	Association	The bi-directional relationships among concept classes	Support	5	<i>“So the dotted lines seem like they’re more of an interaction between the blocks, it is not so systematic in that it must flow in one direction...”</i>
			Oppose	2	<i>“I don’t really get why organization/provider characteristics are here, paired with results as opposed to anywhere else, ‘Organizational/Provider’ could be paired with patient, the intervention, I don’t really see why it has to be attached to ‘detection/treatment.’”</i>
	Subsumption	The hierarchical relationships among the classes and the sections were they are located.	Support	3	<i>“The rest of the parent child-relationships seem fine.”</i>
			Oppose	3	<i>“So the ‘Study Compliance Characteristics’ Yes you have consent, but ‘Life Expectancy’ how do you, I just don’t understand, how are you getting that... how is that grouped with consent, I think that is the only one that doesn’t really make any sense.”</i>



Table 6-7 presents interesting quotes that speak to the broader issues surrounding the participants' experiences with defining data needs for research projects. Researchers are aware of the difficulties of gaining access to high quality EHR data.

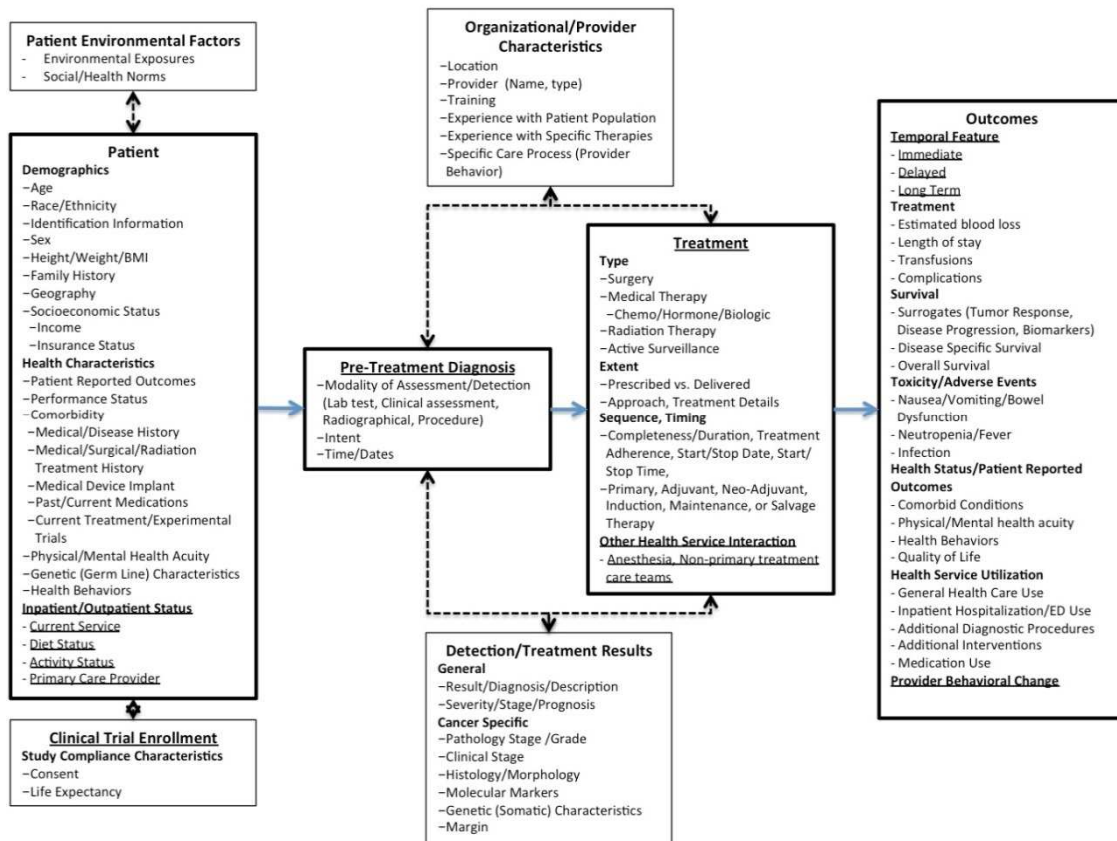


**Table 6-7. Notable quotes that speak to the difficulties of conveying medical information needs.**

ID	Quote	Why it was notable
4	“Data that the researcher has access to are different from data that the researcher needs to request access for...It might be helpful when presenting these elements even in a color coded way, to let the researcher know you do not have access to X, but you do have access to Y.”	This is a theme that continues to present itself. Researchers do not know what data is actually available, and if it is, do they have access to it? This is a major source of anxiety for researchers when they are expressing their information/data needs
5	“It would be helpful to have a separate box, cause you saw I struggled a little bit with [Quality of Bowel Prep] and [Cecum Reached], those aren’t quite findings, but quality related outcomes, and with the increased interest and focus on quality and benchmarks, were all these necessary steps fulfilled to say this was a good exam, I wonder if this should even be a box.”	Quality related outcomes seem to be a new line of research as a direct result of the required reporting of quality metrics due to the current political environment. Furthermore, these concepts are not granular data elements but derivations of existing data elements, or elements not routinely collected to produce abstract measures.
5	“I think ultimately you are on the right track, I just think that you can’t infer perfection just because of the nature of that, the researcher is going to have to reconcile that there is going to be a fixed template and they have to look in those specific spots [for what they need]”	A static representation is flawed. The context of the data need is highly relevant for how the medical concepts are organized conceptually.
6	“Here is a classic one, Preceding interventions in other health systems. This deserves a box... This falls on the unobtainable data but probably should be a category of information that exists, and it could exist pre and post, other health systems with all the same boxes but we don’t have that information, at least to recognize that is a huge piece of the conversation. The pre care and post care world, it would be all this [the model] in each of these boxes, recognizing that is profoundly helpful on day one”	Interestingly, many patients have a limited interaction with individual healthcare systems, meaning the majority of their healthcare data is contained outside of any one institution and as such, their data is unavailable. The assumption that this data is easily accessible or that these types of patients are outliers creates issues during the data needs negotiation process.
7	“The surgeons and the people like that have always had more support and have a long history of having registries. It was more available to them. That’s new to other fields [our needs are poorly represented].”	The difficulty of building a model that represents the collective needs for all researchers is inherently biased towards those who had the resources.

### 6.3.3 *Participant-Enriched Schema*

**Figure 6-4** is the participant-enriched schema based on my evaluation. This representation is a significant departure from the original Carpenter framework. I will first describe the major structural changes followed by granular class changes and their justifications, respectively. First, many evaluators expressed confusion with the directed temporal edge that made the conceptual graph cyclic. I removed this edge to simplify the intended temporal information conveyed by the directed edges. Second, many participants expressed difficulty following the temporal pattern. The original framework presented many sections connected in a parallel temporal circuit. While, this representation is probably more accurate of the clinical process, I decided to serialize the major sections in an attempt to better illustrate the temporal pathway a patient follows. Additionally, I increased the border thickness for the major sections of this temporal process: Patient, Pre-Treatment Diagnosis, Treatment, and Outcomes.



**Figure 6-4. Participant enriched schema. The major sections are aligned and highlight as boxes with thicker borders. The sections are connected in series with blue directed edges to simplify the implications of a temporal flow. The associated sections are connected with dashed, undirected edges. The participants added 9 additional classes to the enriched schema. These are underlined within the sections. These classes were not found in the original framework. Additionally, section names that were changed are also underlined.**

Furthermore, I renamed the major sections to better align with clinical terminology. For example, I changed the sections “Detection/Diagnostics” and “Intervention” to “Pre-Treatment Diagnosis” and “Treatment” as this better reflects clinical care documentation. This alteration is a direct change based on the following quote,

*“If you want to be more generic and applicable to screening procedures in general, one heading that preceded the EMR, back when it was all on paper, operative notes had a ‘Pre-procedure diagnosis’. So, I wonder if data elements would be better organized that way...*

*That would guide, the clinician would immediately know which box to go to for those two things.”*

The traditional language used to describe the clinical course of a patient is a key component. The language used by physicians to describe the clinical course is best used to represent the sections of the schema. The original framework is based on the cancer care continuum and as such probably over emphasizes the survival outcomes from the cancer domain. Non-malignant disease researchers were confused by the focus on survival outcomes. I felt that both survival and non-survival outcomes were both classes under the section “Outcomes” and as such are represented in one section. Finally, I created the section “Clinical Trial Enrollment” as multiple participants felt it did not belong to the set of classes in the “Patient” section. I added the following classes to the “Patient” section: Inpatient/Outpatient status (Current Service, Diet Status, Activity Status, and Primary Care Provider). Multiple participants described this as an integral class aiding cohort identification.

I added the following classes to the “Treatment” Section: Other Health Service Interaction (Anesthesia, Non-primary treatment care teams) based on an inference observed by participants 2, 3, and 7, in that many of the interventions in their studies are secondary treatments or care processes to a primary intervention the patient is receiving. This class was also of interest to participant 6, as this subject was concerned with what effect this may have on major outcomes of interest.

## **6.4 Discussion**

### *6.4.1 Implications of Results*

I posit the way medical researchers organize medical concepts may aid the efficient elicitation of data needs, and may provide an easier interface for query analysts to map CDE or EHR data elements to medical concepts described in the data needs. The Carpenter framework is representative for how researchers conceptually organize cancer research data needs. I hypothesized the Carpenter framework was a well-organized and comprehensive representation of concepts used in comparative effectiveness research (CER) for cancer and that it could be extended with new classes identified through real-world data to represent data needs for various medical domains. My enrichment of the Carpenter framework utilizing three datasets provides some interesting findings. First, I confirmed that the Carpenter framework is a well-organized and comprehensive representation of medical concepts used in CER for cancer. This was observed through the high preservation of the original classes in the data-enriched schema. 79% of concepts were preserved in the data-enriched schema. Furthermore, 86% of the sections and 86% of the directed edges were preserved, suggesting the conceptual organization was persevered. Additionally, the data-enriched schema extends the breadth of classes represented for other medical domains and research approaches.

Finally, the evaluation of the data-enriched schema provided significant insight regarding the understandability of the schema. Specifically, the reorganization of the core sections in line with the directed edge representing a temporal sequence was a major adjustment intended to convey a focus on the sections across a timeline. Additionally, my intended use of the enriched schema as an aid for the specification of data needs showed initial promise. During the course of the evaluation, specifically the mapping component, the data-enriched schema stimulated many

participants to describe additional medical concepts they required to complete their research. Many saw the enriched schema as a mechanism to help aid the specification of their needs, and others saw it as a tool to be used during a data needs negotiation with a query analyst. I expand on this idea in the next section.

#### *6.4.2 Intended Use Case*

Our final schema presented in **Figure 6-4** may serve as a bridge between the medical researcher and the query analyst. Both stakeholders may use this schema to specify and elicit key medical concepts needed for a research project. I envision the employment of this schema in three scenarios. The first would be to refine a data request by providing a template through which the medical researcher could specify their data need initially. The representation may stimulate the researcher to define their data need with increased granularity and clarity. The second would provide a concept schema through which a query analyst could orient themselves to the mental model of researchers, allowing them to better engage and elicit additional criteria related to the initial data request. The schema may facilitate the negotiation between the researcher and query analyst by supplying a checklist through which the data need can be defined. The third would serve as a metadata schema for indexing and reusing data requests. The concept schema can provide a compact list of codes for annotating the data requests.

#### *6.4.3 Limitations*

Our study has several limitations. First, as the evaluation confirmed, the data enriched schema does contain ambiguity. The abstraction of granular medical concepts introduces ambiguity. However, the more positively reviewed aspect of the data enriched schema was its conceptual organization of medical concepts used in research. Second, each dataset I chose contains an inherent bias. Clinical Trials represent the current state of research as influenced by

major health concerns, for example cardiovascular disease, metabolic disease, and cancer. As such, this dataset may overemphasize these medical domains affecting my ability to generalize the results to other domains. Similarly, the institutional data request logs are also a representation of the research priorities at Columbia University and as such may skew the results toward those domains. Thirdly, the EHR SQL query dataset is from one domain of medicine and hence may not cover variables outside Urology.

### **6.5 Conclusion**

I used a data-driven approach to develop a conceptual schema for defining clinical research data needs. My evaluation confirms the satisfactory concept class coverage of this schema and its generalizability across disease domains. This schema has the potential to facilitate communication between researchers and query analysts, or to serve as a metadata schema for indexing, organizing data requests thereby empowering knowledge reuse among researchers. Future studies are warranted to test these potentials.

## Chapter 7. Summary and Conclusions

In this chapter, I will summarize my findings from investigations of BQM and my enriched concept class schema (Section 7.1), discuss my contributions to the biomedical literature (Section 7.2), review the limitations of my work (Section 7.3), and discuss new opportunities for continued investigations of BQM enabled by this dissertation research (Section 7.4).

### 7.1 *The gestalt view of biomedical query mediation*

This dissertation provides a detailed understanding of how BQM facilitates EHR data access for medical researchers. Guided by my dissertation blueprint displayed in **Figure 1-4**, I (1) conducted a thorough review of the literature, (2) developed an in-depth understanding of BQM and (3) created an enriched EHR data needs conceptual model. In the following paragraphs, I will review the major findings produced in my effort for each aim.

In AIM I, I provided a thorough review of the information science literature and identified several knowledge gaps in the context of BQM: (1) BQM lacks a method to measure the complexity of the medical researchers' information need, (2) the literature presents scarce understanding of medical researchers' cognitive styles effect on their information seeking process, (3) BQM lacks a formal structure for medical researchers to express an information need, and (4) BQM while poorly understood may share similarities to the librarian reference interview.

In AIM II, I studied three expressions of BQM. First, I examined the content of ten Clinical and Translational Science Awards supported institutions' data request forms to understand how medical researchers begin the BQM process. My analysis found that these forms contain an



overabundance of regulatory elements, and while the need to ensure the medical researcher is compliant with the procedures established by the institution, extensive regulatory compliance may be out of scope for individuals providing EHR data access and is probably best served by the institutional review board. More importantly, data request forms contain simple form elements guiding the specification of the medical researcher's EHR data need. This finding complemented results obtained from the literature review. Data request forms serve as the initial point of contact for medical researchers to access EHR data. It may benefit the whole process to ensure these forms focus the researcher using form elements that direct the elicitation of non-vague descriptions of their need. Second, I conducted a study using content analysis to research the BQM conversation space between medical researchers and query analysts. My results showed that a large portion of this conversation space focuses on the discussion of the clinical process. Additionally, the context of the conversation oscillates between study design and research workflow, suggesting an iterative nature to the data needs negotiation process until both parties reach consensus. This work provided the preliminary knowledge needed to conduct a cognitive task analysis of BQM. Third, I generated a generalized hierarchical task model representing an amalgamation of multiple query analysts approaches to BQM. As discussed in **section 5.3.3**, this model demonstrates that BQM shares many characteristics with the reference interview used by librarians to elicit a clear definition of a patron's information need. Key similarities can be seen in the content analysis of the BQM conversation space. BQM and the "reference interview" share core attributes. For example, the reference interview task, "Understanding the object and motivation" of the information need is similar to the BQM task "Clarify project type." Both elements attempt to elicit the big picture with the goal of

contextualizing the information need providing a more effective foundation to discuss the particulars of the information need.

Two of the studies confirmed the need for a formal structure through which the medical researcher can propose their information need with clarity. To address this need, I identified an existing framework that categorized data needs for cancer comparative effectiveness research across the cancer care continuum. I enriched this framework using a data driven approach combined with user assessment to arrive at an enriched concept class schema designed to represent medical researcher data needs. This approach to representing EHR data needs through real world information needs has not been previously attempted. The schema presents a novel query template designed to improve query formulation of EHR data needs for medical researchers. In addition, this schema may serve both medical researcher and query analyst as a bridge to expedite and enable efficient communication during the data needs negotiation process. Finally, this schema may serve query analysts as an indexing tool to help organize data requests and facilitate knowledge reuse from analyst to analyst.

## ***7.2 Contributions to Clinical Research Informatics***

This dissertation produced several contributions to the Clinical Research Informatics knowledge base. Specifically, the dissertation contributes a detailed understanding of the decisions and tasks used to formulate an EHR data need providing a roadmap for the development of future cognitive computing applications facilitating the complex decision making process of BQM. These contributions include the following:

- (1) The identification of complementary knowledge from outside clinical research informatics literature that aided this study of users accessing EHR data

(2) An understanding of the complex sociotechnical process used by medical researchers and query analyst to formulate and translate EHR data queries into executable EHR database queries

(3) An enriched concept class schema representing the EHR data needs of medical researchers

In the biomedical literature, the process medical researchers use to access EHR data is rarely visible and poorly studied. Although some attempts have been made to bridge the barriers to EHR data access, the complex sociotechnical aspects involved with formulating an information need and translating that need into an EHR data representation remain daunting. The detailed BQM process I illustrated may provide others with a better understanding of the BQM process as a whole, and allow informaticians to better target future studies and interventions.

Self-service query tools represent initial attempts to provide EHR data access to researchers and have been successful in resolving the majority of simple EHR data needs. Though these tools are efficient at resolving the medical researcher's simple EHR data needs, currently they are incapable of resolving complex EHR data needs. They fail to provide cognitive support and ignore the socio-technical components of query formulation and query execution. This dissertation provides a clear understanding of BQM and correlates the tasks of BQM with elements of other models facilitating a clear, non-vague description of a user's information need, specifically, the reference interview. I identified key tasks of BQM that contribute to query formulation and aligned them with established goals of the reference interview, suggesting several BQM tasks are critical to understanding the information needs of medical researchers.

Additionally, this finding suggests that an interactive BQM occurring between a medical researcher and query analyst is most appropriate for supporting effective query formulation.

Finally, to enable the communication between two experts of different domains, medical and technical, I enriched a concept class schema designed to represent EHR data needs of medical researchers. The schema represents concept classes in core groups organized in a temporal flow designed to represent how a medical researcher organizes concepts along the care continuum of the patient. This organization may better facilitate a clear specification of a medical researcher's data needs by presenting concepts in a framework reflective of the medical researcher's mental model. This organization may also align the query analyst's mental model to that of the medical researcher providing a shared framework to engage in BQM. The evaluation confirms the schema's usefulness across multiple medical domains. Additionally, the schema showed initial promise during the evaluation as a mechanism to elicit additional details from medical researchers who may be struggling to describe an EHR data need. This is promising for the incorporation of the schema as a query template for the medical researcher engaged in a BQM process.

### ***7.3 Limitations***

My research contains several limitations. First, my work may not be generalizable to other types of requests as my studies focused on medical research data requests. How users seek information is significantly influenced by the context of their need. Additionally, my study of the BQM process was largely from the perspective of the query analyst. I did not address the perspectives of all key stakeholders across each of the aims of this dissertation. In the following sections, I will analyze the limitations of using a single use case scenario and a single stakeholder approach to investigate BQM.

### *7.3.1 Limitations of using single use case scenario to understand BQM*

Medical researchers represent a small subset of the population of users seeking and accessing EHR data. Through my investigation of the literature in Chapter 2, I discovered that the context of the user's need influences the process of obtaining data. Medical researchers often approach information seeking from the mindset of a clinical scenario. The expression of their need represents a clinical progression. My research of BQM is limited to this setting. For example, it could be inappropriate to use this understanding of BQM and apply it to the setting of a hospital administrator seeking EHR data assessing resource use within the hospital. Although similar concepts may be needed, the administrator most likely would approach their information need from a different mindset than that of the medical researcher. It is possible the query analyst would use a different process to formulate the query of the hospital administrator.

### *7.3.2 Limitations of using a single stakeholder analysis to understand BQM*

AIM II and III address two different research questions from the perspective of one stakeholder. First, AIM II investigates the process and tasks used by query analysts to conduct BQM. I investigated BQM from the perspective of the query analyst. I decided to approach this problem with the following assumption: the query analyst conducts BQM. This suggests the medical researcher has a more passive role during the information seeking process. However, my analysis found the medical researcher to provide an active role in the process, educating the query analyst on the data collection process during the clinical workflow, and linking clinical concepts with EHR data elements and representations. This one example may suggest there are other latent tasks central to the medical researcher that must occur during BQM that I failed to identify because of this largely query analyst-driven approach.

Second, AIM III built a conceptual model for medical researchers' data needs. I used a data-driven approach to enrich an existing model and evaluated the enriched model using a diverse group of medical researchers to evaluate and further enrich the model. This work only considered the perspective of the medical researcher. I believe this model could enable the elicitation of non-vague EHR data needs from medical researchers. Additionally, I suggest the model could facilitate the data needs negotiation between medical researchers and query analysts by providing an organizational framework to guide their discussion. This conclusion assumes the enriched class schema is easily understood by and would be helpful to the query analyst. However, my evaluations did not assess this statement.

#### ***7.4 Future Work***

This dissertation provides previously unavailable insight into the complexities of BQM. It highlights several pathways to improve cognitive support for medical researchers seeking EHR data. To begin, medical researchers' information seeking process is limited by Self-service query tools. My work suggests future Self-service query tools for EHR data access may benefit from improvements supporting query formulation. Second, the generalized task model of BQM can serve as a knowledge source for the building of a process management application to facilitate the organization of EHR data requests an institution receives. In addition, the task model provides insight enabling the development of cognitive computing applications.

As well, this study of BQM highlights a tangential lead for future work. Query analysts and medical researchers leverage their respective knowledge to build EHR data phenotypes for various medical concepts. BQM, as a knowledge source, is ripe for the study of EHR phenotyping. The following sections will discuss these ideas.

#### *7.4.1 Supporting cognition for BQM*

Query formulation for medical researchers is not a trivial process. Current Self-service query tools allow medical researchers to navigate through exhaustive medical terminologies and select appropriate EHR data elements for their query. It is incumbent on the medical researcher to identify and organize EHR data elements into buckets representing various medical concepts. However, Self-service query tools do not provide a mechanism to first define and organize the buckets representing medical concepts for their information need. The concept class schema presented in Chapter 6 may serve future self-service query tool design facilitating query formulation. A tool that guides the medical researcher's information seeking process to first define the information need and then select EHR data elements commensurate with that need, may produce Self-service query tools capable of resolving complex EHR data needs.

The scope of this dissertation focused on acquiring detailed BQM knowledge. It uncovered a complex socio-technical process that aids in the formulation of a conceptual EHR information need and then translating the need into EHR data representations. As more users request access to EHR data for medical research and other use cases, a process management application is needed. Such a tool may better facilitate the tracking and delivery of EHR data to users. The knowledge generated in this dissertation provides a framework for the development of a workflow management application. This application would be particularly useful to EHR data warehouse managers and their teams of EHR query analyst. It could enable managers to organize EHR data requests and monitor the life cycle of multiple EHR data requests occurring across multiple discrete BQM processes. The application may also provide future researchers with an enhanced representation of BQM for continued study. For example, such an application would track time to completion for BQM tasks providing researchers with a surrogate marker to

measure the complexity of tasks within BQM. In addition, the enhanced BQM representation may also serve an end goal for the development of an automated agent to facilitate Biomedical Query Mediation with medical researchers and other users seeking EHR data for secondary use.

#### *7.4.2 BQM as a data source for EHR Phenotyping*

In Chapter 5, Figure 5-2 highlights the process flow for the BQM task model. The task, “Establishing the Index and Associated Phenotype(s)”, highlights the most complex task detailing an iterative process where a medical researcher and query analyst produce an EHR data representation from a medical concept. Medical researchers and query analysts leverage their respective knowledge of medicine and the information system to develop EHR data representations or an EHR phenotype. Current approaches to producing phenotype definitions and EHR data representations include expert derived, semi-automated, and automated approaches. However, we have yet to exploit the content generated during BQM for phenotype knowledge acquisition. The descriptive knowledge generated from the BQM process may be shown to be of significant value as it provides a mechanism for refining phenotype definitions and EHR data representations that are in synch with the evolving EHR systems and medical practices.

### **7.5 Conclusions**

There exists a correlation between the types of information collected in an EHR and the complexity of EHR data requests for research use submitted by medical researchers. Over time both the types of information and the complexity of EHR data requests will increase, placing a greater strain on limited resources facilitating EHR data access. To enable these resources to accommodate the new demands asked of them, I developed a deep understanding of the socio-technical process, BQM, used to facilitate complex EHR data requests. The generalized



hierarchical task model may serve as a reference for query analysts engaging in BQM. It will encourage the efficient expression and formulation of information needs from users, helping remove ambiguity and thereby producing accurate EHR data representations of medical concepts. I have carried out an extensive series of studies that document this process and present a generalized view across many institutions. Furthermore, based on recommendations for increasing specificity of an information need, I developed an enriched concept schema representing EHR data needs organized in agreement with medical researchers' conceptual organization of medical concepts. This dissertation brings transparency to a complex process that facilitates EHR data access for medical researchers. This understanding provides a path for further studies and BQM process refinements that will result in more time efficient and accurate EHR data representations of the medical researcher's information needs, thus contributing to improved medical research results.

## References

1. Kuhlthau, C.C., *Inside the search process: Information seeking from the user's perspective*. JASIS, 1991. **42**(5): p. 361-371.
2. Belkin, N.J., R.N. Oddy, and H.M. Brooks, *ASK for information retrieval: Part I. Background and theory*. Journal of documentation, 1982. **38**(2): p. 61-71.
3. Li, Y. and N.J. Belkin, *A faceted approach to conceptualizing tasks in information seeking*. Information Processing & Management, 2008. **44**(6): p. 1822-1837.
4. Ford, N., T. Wilson, A. Foster, D. Ellis, and A. Spink, *Information seeking and mediated searching. Part 4. Cognitive styles in information seeking*. Journal of the American Society for Information Science and Technology, 2002. **53**(9): p. 728-735.
5. Ruthven, I., *Interactive information retrieval*. Annual review of information science and technology, 2008. **42**(1): p. 43-91.
6. Hansen, P. and K. Järvelin, *Collaborative information retrieval in an information-intensive domain*. Information Processing & Management, 2005. **41**(5): p. 1101-1119.
7. Robins, D., *Shifts of focus on various aspects of user information problems during interactive information retrieval*. Journal of the American Society for Information Science, 2000. **51**(10): p. 913-928.
8. Spink, A. and T. Wilson. *Toward a Theoretical Framework for Information Retrieval (IR) Evaluation in an Information Seeking Context*. in *Mira*. 1999. Taylor Graham Publishing. p. 21-34
9. Wildemuth, B.M., *The effects of domain knowledge on search tactic formulation*. Journal of the American Society for Information Science and Technology, 2003. **55**(3): p. 246-258.
10. Vakkari, P., *Task-based information searching*. Annual review of information science and technology, 2005. **37**(1): p. 413-464.
11. Murphy, S.N., V. Gainer, and H.C. Chueh. *A visual interface designed for novice users to find research patient cohorts in a large biomedical database*. in *AMIA Annual Symposium Proceedings*. 2003. American Medical Informatics Association. p. 489
12. Plaisant, C., S. Lam, B. Shneiderman, M.S. Smith, D. Roseman, G. Marchand, M. Gillam, C. Feied, J. Handler, and H. Rappaport, *Searching electronic health records for temporal patterns in patient histories: a case study with microsoft amalga*. AMIA Annu Symp Proc, 2008: p. 601-5.
13. Murphy, S.N., M.E. Mendis, D.A. Berkowitz, I. Kohane, and H.C. Chueh, *Integration of clinical and genetic data in the i2b2 architecture*. AMIA Annu Symp Proc, 2006: p. 1040.
14. Weber, G.M., S.N. Murphy, A.J. McMurry, D. Macfadden, D.J. Nigrin, S. Churchill, and I.S. Kohane, *The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories*. J Am Med Inform Assoc, 2009. **16**(5): p. 624-30.
15. McMurry, A.J., S.N. Murphy, D. Macfadden, G. Weber, W.W. Simons, J. Orechia, J. Bickel, N. Wattanasin, C. Gilbert, P. Trevvett, S. Churchill, and I.S. Kohane, *SHRINE: Enabling Nationally Scalable Multi-Site Disease Studies*. PLoS One, 2013. **8**(3): p. e55811.

16. Zhang, G.Q., T. Siegler, P. Saxman, N. Sandberg, R. Mueller, N. Johnson, D. Hunscher, and S. Arabandi, *VISAGE: A Query Interface for Clinical Research*. AMIA Summits Transl Sci Proc, 2010. **2010**: p. 76-80.
17. Lowe, H.J., T.A. Ferris, P.M. Hernandez, and S.C. Weber, *STRIDE--An integrated standards-based translational research informatics platform*. AMIA Annu Symp Proc, 2009. **2009**: p. 391-5.
18. Hripcsak, G., J. Duke, N. Shah, C. Reich, V. Huser, M. Schuemie, M. Suchard, R. Park, I. Wong, and P. Rijnbeek, *Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers*. Studies in health technology and informatics, 2014. **216**: p. 574-578.
19. Holve, E., C. Segal, and M. Hamilton Lopez, *Opportunities and challenges for comparative effectiveness research (CER) with Electronic Clinical Data: a perspective from the EDM forum*. Med Care, 2012. **50 Suppl**: p. S11-8.
20. Deshmukh, V., S. Meystre, and J. Mitchell, *Evaluating the informatics for integrating biology and the bedside system for clinical research*. BMC medical research methodology, 2009. **9**(1): p. 70.
21. Natarajan, K., A.B. Wilcox, N. Sobhani, and A. Boyer. *Analyzing Requests for Clinical Data for Self-Service Penetration*. in *AMIA*. 2013. p. 1049
22. Ammenwerth, E., C. Iller, and C. Mahler, *IT-adoption and the interaction of task, technology and individuals: a fit framework and a case study*. BMC Medical Informatics and Decision Making, 2006. **6**(1): p. 3.
23. Hruby, G.W., J. McKiernan, S. Bakken, and C. Weng, *A centralized research data repository enhances retrospective outcomes research capacity: a case report*. J Am Med Inform Assoc, 2013. **20**(3): p. 563-567.
24. Suchman, L., *Making work visible*. Commun. ACM, 1995. **38**(9): p. 56-ff.
25. Weir, C.R., J.J. Nebeker, B.L. Hicken, R. Campo, F. Drews, and B. LeBar, *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): p. 65-75.
26. Taylor, R.S., *Question-negotiation and information seeking in libraries*. College & research libraries, 1967. **29**(3): p. 178-194.
27. Spink, A., *Study of interactive feedback during mediated information retrieval*. Journal of the American Society for Information Science, 1997. **48**(5): p. 382-394.
28. Bates, M.J., *The design of browsing and berrypicking techniques for the online search interface*. Online Information Review, 1989. **13**(5): p. 407-424.
29. Wang, D., D.R. Kaufman, E.A. Mendonca, Y.H. Seol, S.B. Johnson, and J.J. Cimino, *The cognitive demands of an innovative query user interface*. Proc AMIA Symp, 2002: p. 850-4.
30. Johnson, S.B., G. Hripcsak, J. Chen, and P. Clayton, *Accessing the Columbia Clinical Repository*. Proc Annu Symp Comput Appl Med Care, 1994: p. 281-5.
31. Zheng, K., Q. Mei, and D.A. Hanauer, *Collaborative search in electronic health records*. J Am Med Inform Assoc, 2011. **18**(3): p. 282-91.
32. DesRoches, C.M., D. Charles, M.F. Furukawa, M.S. Joshi, P. Kralovec, F. Mostashari, C. Worzala, and A.K. Jha, *Adoption of electronic health records grows rapidly, but fewer than half of US hospitals had at least a basic system in 2012*. Health Affairs, 2013: p. 10.1377/hlthaff.2013.0308.

33. Hersh, W.R., *Adding value to the electronic health record through secondary use of data for quality assurance, research, and surveillance*. Am J Manag Care, 2007. **81**: p. 126-28.
34. D'Avolio, L.W., W.R. Farwell, and L.D. Fiore, *Comparative effectiveness research and medical informatics*. The American Journal of Medicine, 2010. **123**(12): p. e32-e37.
35. Holve, E., C. Segal, M.H. Lopez, A. Rein, and B.H. Johnson, *The Electronic Data Methods (EDM) forum for comparative effectiveness research (CER)*. Med Care, 2012. **50 Suppl**: p. S7-10.
36. Miriovsky, B.J., L.N. Shulman, and A.P. Abernethy, *Importance of health information technology, electronic health records, and continuously aggregating data to comparative effectiveness research and learning health care*. J Clin Oncol, 2012. **30**(34): p. 4243-8.
37. Hoffman, S. and A. Podgurski, *Big Bad Data: Law, Public Health, and Biomedical Databases*. Journal of Law, Medicine and Ethics, 2013. **41**(s1): p. 56-60.
38. Safran, C., M. Bloomrosen, W.E. Hammond, S. Labkoff, S. Markel-Fox, P.C. Tang, and D.E. Detmer, *Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper*. Journal of the American Medical Informatics Association, 2007. **14**(1): p. 1-9.
39. Hripcsak, G., C. Knirsch, L. Zhou, A. Wilcox, and G.B. Melton, *Bias associated with mining electronic health records*. Journal of biomedical discovery and collaboration, 2011. **6**: p. 48.
40. Hripcsak, G. and D.J. Albers, *Next-generation phenotyping of electronic health records*. Journal of the American Medical Informatics Association, 2012. **20**(1): p. 117-121.
41. Zerhouni, E.A. and B. Alving, *Clinical and translational science awards: a framework for a national research agenda*. Translational research : the journal of laboratory and clinical medicine, 2006. **148**(1): p. 4-5.
42. Selby, J.V., A.C. Beal, and L. Frank, *The Patient-Centered Outcomes Research Institute (PCORI) national priorities for research and initial research agenda*. JAMA, 2012. **307**(15): p. 1583-1584.
43. Friedman, C. and M. Rigby, *Conceptualising and creating a global learning health system*. International journal of medical informatics, 2013. **82**(4): p. e63-e71.
44. Friedman, C.P., A.K. Wong, and D. Blumenthal, *Achieving a nationwide learning health system*. Science translational medicine, 2010. **2**(57): p. 57cm29-57cm29.
45. Blumenthal, D. and M. Tavenner, *The "meaningful use" regulation for electronic health records*. N Engl J Med, 2010. **363**(6): p. 501-4.
46. Read, K.B., A. Surkis, C. Larson, A. McCrillis, A. Graff, J. Nicholson, and J. Xu, *Starting the data conversation: informing data services at an academic health sciences library*. Journal of the Medical Library Association: JMLA, 2015. **103**(3): p. 131.
47. Rein, A., *Finding Value in Volume: An Exploration of Data Access and Quality Challenges*. AcademyHealth: Briefs and Reports, 2012: p. 9.
48. Patel, V.L., J.F. Arocha, and D.R. Kaufman, *A primer on aspects of cognition for medical informatics*. Journal of the American Medical Informatics Association, 2001. **8**(4): p. 324-343.
49. Byström, K. and K. Järvelin, *Task complexity affects information seeking and use*. Information Processing & Management, 1995. **31**(2): p. 191-213.

50. Weiskopf, N.G. and C. Weng, *Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research*. J Am Med Inform Assoc, 2013. **20**(1): p. 144-51.
51. Apgar, V., *A proposal for a new method of evaluation of the newborn infant*. Curr Res Anesth Analg, 1953. **32**(4): p. 260-7.
52. Goldman, L., D.L. Caldera, S.R. Nussbaum, F.S. Southwick, D. Krogstad, B. Murray, D.S. Burke, T.A. O'Malley, A.H. Goroll, C.H. Caplan, J. Nolan, B. Carabello, and E.E. Slater, *Multifactorial index of cardiac risk in noncardiac surgical procedures*. N Engl J Med, 1977. **297**(16): p. 845-50.
53. Adler-Milstein, J., C.M. DesRoches, P. Kralovec, G. Foster, C. Worzala, D. Charles, T. Searcy, and A.K. Jha, *Electronic Health Record Adoption In US Hospitals: Progress Continues, But Challenges Persist*. Health Affairs, 2015: p. 10.1377/hlthaff.2015.0992.
54. Borycki, E., D. Newsham, and D. Bates, *eHealth in North America*. Yearbook of medical informatics, 2012. **8**(1): p. 3.
55. Hersh, W.R., M.G. Weiner, P.J. Embi, J.R. Logan, P.R.O. Payne, E.V. Bernstam, H.P. Lehmann, G. Hripcsak, T.H. Hartzog, J.J. Cimino, and J.H. Saltz, *Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research*. Medical Care, 2013. **51**(8): p. S30-S37.
56. Wade, T.D., R.C. Hum, and J.R. Murphy, *A Dimensional Bus model for integrating clinical and research data*. Journal of the American Medical Informatics Association, 2011. **18**(Suppl 1): p. i96-i102.
57. Yip, Y., *Unlocking the potential of electronic health records for translational research. Findings from the section on bioinformatics and translational informatics*. Yearbook of medical informatics, 2012. **7**(1): p. 135.
58. Arzberger, P., P. Schroeder, A. Beaulieu, G. Bowker, K. Casey, L. Laaksonen, D. Moorman, P. Uhler, and P. Wouters, *Promoting access to public research data for scientific, economic, and social development*. Data Science Journal, 2004. **3**(0): p. 135-152.
59. Kumpulainen, S. and K. Järvelin, *Barriers to task-based information access in molecular medicine*. Journal of the American Society for Information Science and Technology, 2012. **63**(1): p. 86-97.
60. Collins, F.S., K.L. Hudson, J.P. Briggs, and M.S. Lauer, *PCORnet: turning a dream into reality*. Journal of the American Medical Informatics Association, 2014. **21**(4): p. 576-577.
61. Meystre, S.M., V.G. Deshmukh, and J. Mitchell, *A clinical use case to evaluate the i2b2 Hive: predicting asthma exacerbations*, in *AMIA Annual Symposium Proceedings*. 2009, American Medical Informatics Association. p. 442.
62. Borgman, C.L., *Why are online catalogs still hard to use?* JASIS, 1996. **47**(7): p. 493-503.
63. Vakkari, P., *Task complexity, problem structure and information actions: integrating studies on information seeking and retrieval*. Information Processing & Management, 1999. **35**(6): p. 819-837.
64. Dervin, B., *From the mind's eye of the user: the sense-making qualitative-quantitative methodology*. Qualitative research in information management, 1992. **9**: p. 61-84.
65. Dervin, B., *Sense-making theory and practice: an overview of user interests in knowledge seeking and use*. Journal of knowledge management, 1998. **2**(2): p. 36-46.

66. Feltovich, P.J., R.R. Hoffman, D. Woods, and A. Roesler, *Keeping it too simple: How the reductive tendency affects cognitive engineering*. Intelligent Systems, IEEE, 2004. **19**(3): p. 90-94.
67. Klein, G., B. Moon, and R.R. Hoffman, *Making sense of sensemaking 2: A macrocognitive model*. Intelligent Systems, IEEE, 2006. **21**(5): p. 88-92.
68. Klein, G., B. Moon, and R.R. Hoffman, *Making sense of sensemaking 1: Alternative perspectives*. Intelligent Systems, IEEE, 2006. **21**(4): p. 70-73.
69. Klein, G., J.K. Phillips, E.L. Rall, and D.A. Peluso, *A data-frame theory of sensemaking. Expertise out of context*, 2007: p. 113-155.
70. Olsson, M.R., *Re-thinking our concept of users*. Australian Academic & Research Libraries, 2009. **40**(1): p. 22-35.
71. Blandford, A. and S. Attfield, *Interacting with information*. Synthesis Lectures on Human-Centered Informatics, 2010. **3**(1): p. 1-99.
72. Ford, N. and R. Ford, *Towards a cognitive theory of information accessing: an empirical study*. Information Processing & Management, 1993. **29**(5): p. 569-585.
73. Wildemuth, B.M., *Post-positivist research: two examples of methodological pluralism*. The Library Quarterly, 1993: p. 450-468.
74. Warner, H.R. and J.D. Morgan, *High-density medical data management by computer*. Computers and Biomedical Research, 1970. **3**(5): p. 464-476.
75. Warner, H.R. *Knowledge sectors for logical processing of patient data in the HELP system*. in *Proceedings of the Annual Symposium on Computer Application in Medical Care*. 1978. American Medical Informatics Association. p. 401
76. Jarke, M., J. Tuner, E.A. Stohr, Y. Vassiliou, N.H. White, and K. Michielsen, *A field evaluation of natural language for data retrieval*. Software Engineering, IEEE Transactions on, 1985(1): p. 97-114.
77. Dervin, B. and P. Dewdney, *Neutral questioning: A new approach to the reference interview*. RQ, 1986: p. 506-513.
78. Borgman, C.L., N.J. Belkin, W.B. Croft, M.E. Lesk, and T.K. Landauer. *Retrieval systems for the information seeker: can the role of the intermediary be automated?* in *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1988. ACM. p. 51-53
79. Merz, R.B., C. Cimino, G.O. Barnett, D.R. Blewett, J.A. Gnassi, R. Grundmeier, and L. Hassan, *Q & A: a query formulation assistant*. Proc Annu Symp Comput Appl Med Care, 1992: p. 498-502.
80. Schoening, P.A., C.A. Abrams, and M.G. Kahn. *An object model for uniform access to heterogeneous databases*. in *Proceedings of the Annual Symposium on Computer Application in Medical Care*. 1993. American Medical Informatics Association. p. 502
81. Cimino, J.J., A. Aguirre, S.B. Johnson, and P. Peng, *Generic queries for meeting clinical information needs*. Bull Med Libr Assoc, 1993. **81**(2): p. 195-206.
82. Lindberg, D.A., B.L. Humphreys, and A.T. McCray, *The Unified Medical Language System*. Methods of information in medicine, 1993. **32**(4): p. 281-291.
83. Richardson, W.S. and A.L. Murphy, *Ask, and ye shall retrieve*. Evidence Based Medicine, 1998. **3**(4): p. 100-101.
84. Hripesak, G., B. Allen, J.J. Cimino, and R. Lee, *Access to data: comparing AccessMed with Query by Review*. J Am Med Inform Assoc, 1996. **3**(4): p. 288-99.

85. Murphy, S.N., M.M. Morgan, G.O. Barnett, and H.C. Chueh. *Optimizing healthcare research data warehouse design through past COSTAR query analysis*. in *Proceedings of the AMIA Symposium*. 1999. American Medical Informatics Association. p. 892
86. Mendonça, E.A. and J.J. Cimino, *Building a knowledge base to support a digital library*. *Studies in health technology and informatics*, 2001(1): p. 221-225.
87. Mendonça, E.A., J.J. Cimino, S.B. Johnson, and Y.-H. Seol, *Accessing heterogeneous sources of evidence to answer clinical questions*. *Journal of biomedical informatics*, 2001. **34**(2): p. 85-98.
88. Wu, M.M. and Y.H. Liu, *Intermediary's information seeking, inquiring minds, and elicitation styles*. *Journal of the American Society for Information Science and Technology*, 2003. **54**(12): p. 1117-1133.
89. Schardt, C., M.B. Adams, T. Owens, S. Keitz, and P. Fontelo, *Utilization of the PICO framework to improve searching PubMed for clinical questions*. *BMC medical informatics and decision making*, 2007. **7**(1): p. 16.
90. Hung, P.W., S.B. Johnson, D.R. Kaufman, and E.A. Mendonça, *A multi-level model of information seeking in the clinical domain*. *Journal of biomedical informatics*, 2008. **41**(2): p. 357-370.
91. Rankin, J.A., S.F. Grefsheim, and C.C. Canto, *The emerging informationist specialty: a systematic review of the literature*. *Journal of the Medical Library Association: JMLA*, 2008. **96**(3): p. 194.
92. Newton, K.M., P.L. Peissig, A.N. Kho, S.J. Bielinski, R.L. Berg, V. Choudhary, M. Basford, C.G. Chute, I.J. Kullo, and R. Li, *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): p. e147-e154.
93. Kelly, G.A., *A theory of personality: The psychology of personal constructs*. New York: Norton, 1963.
94. King, G.B., *Open & Closed Questions: The Reference Interview*. RQ, 1972. **12**(2): p. 157-160.
95. Knapp, S.D., *The reference interview in the computer-based setting*. RQ, 1978. **17**(4): p. 320-324.
96. Lynch, M.J., *Reference interviews in public libraries*. *The Library Quarterly*, 1978: p. 119-142.
97. White, M.D., *The dimensions of the reference interview*. RQ, 1981: p. 373-381.
98. White, M.D., *Evaluation of the reference interview*. RQ, 1985: p. 76-84.
99. Kahn, M.G., *The desktop database dilemma*. *Academic Medicine*, 1993. **68**(1): p. 34-37.
100. Kahn, M.G., *Clinical databases and critical care research*. *Critical care clinics*, 1994. **10**(1): p. 37.
101. Richardson, W.S., M.C. Wilson, J. Nishikawa, and R.S. Hayward, *The well-built clinical question: a key to evidence-based decisions*. *ACP J Club*, 1995. **123**(3): p. A12-3.
102. Steib, S., R. Reichley, S. McMullin, K. Marrs, T.C. Bailey, W.C. Dunagan, and M. Kahn. *Supporting ad-hoc queries in an integrated clinical database*. in *Proceedings of the Annual Symposium on Computer Application in Medical Care*. 1995. American Medical Informatics Association. p. 62
103. Dewdney, P. and G. Michell, *Asking "why" questions in the reference interview: A theoretical justification*. *The Library Quarterly*, 1997: p. 50-71.

104. Counsell, C., *Formulating questions and locating primary studies for inclusion in systematic reviews*. *Annals of Internal Medicine*, 1997. **127**(5): p. 380-387.
105. Snowball, R., *Using the clinical question to teach search strategy: fostering transferable conceptual skills in user education by active learning*. *Health Libraries Review*, 1997. **14**(3): p. 167-172.
106. Shahar, Y., *A framework for knowledge-based temporal abstraction*. *Artificial intelligence*, 1997. **90**(1): p. 79-133.
107. Johnson, S.B. and D. Chatziantoniou. *Extended SQL for manipulating clinical warehouse data*. in *Proceedings of the AMIA Symposium*. 1999. American Medical Informatics Association. p. 819
108. Nordlie, R. "User revealment"—*a comparison of initial queries and ensuing question development in online searching and in human reference interactions*. in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. 1999. ACM. p. 11-18
109. Booth, A., A.J. O'Rourke, and N.J. Ford, *Structuring the pre-search reference interview: a useful technique for handling clinical questions*. *Bull Med Libr Assoc*, 2000. **88**(3): p. 239.
110. Villanueva, E.V., E.A. Burrows, P.A. Fennessy, M. Rajendran, and J.N. Anderson, *Improving question formulation for use in evidence appraisal in a tertiary care setting: a randomised controlled trial [ISRCTN66375463]*. *BMC medical informatics and decision making*, 2001. **1**(1): p. 4.
111. Murphy, S.N., G.O. Barnett, and H.C. Chueh. *Visual query tool for finding patient cohorts from a clinical data warehouse of the partners HealthCare system*. in *Proceedings of the AMIA Symposium*. 2000. American Medical Informatics Association. p. 1174
112. Ely, J.W., J.A. Osheroff, M.H. Ebell, M.L. Chambliss, D.C. Vinson, J.J. Stevermer, and E.A. Pifer, *Obstacles to answering doctors' questions about patient care with evidence: qualitative study*. *BMJ: British Medical Journal*, 2002. **324**(7339): p. 710.
113. Spink, A., T.D. Wilson, N. Ford, A. Foster, and D. Ellis, *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): p. 695-703.
114. Janes, J., *Question Negotiation in an Electronic Age*. *The Digital Reference Research Agenda*, 2003: p. 48-60.
115. Small, S., N. Shimizu, T. Strzalkowski, and T. Liu. *HITIQA: A Data Driven Approach to Interactive Question Answering: A Preliminary Report*. in *New Directions in Question Answering*. 2003. p. 94-104
116. Wilcox, A.B. and G. Hripcsak, *The role of domain knowledge in automating medical text report classification*. *J Am Med Inform Assoc*, 2003. **10**(4): p. 330-8.
117. Diekema, A.R., O. Yilmazel, J. Chen, S. Harwell, L. He, and E.D. Liddy, *Finding answers to complex questions*. 2004: p. 143.
118. Goren-Bar, D., Y. Shahar, M. Galperin-Aizenberg, D. Boaz, and G. Tahan. *KNAVE II: the definition and implementation of an intelligent tool for visualization and exploration of time-oriented clinical data*. in *Proceedings of the working conference on Advanced visual interfaces*. 2004. ACM. p. 171-174
119. Lankes, R.D., *The Digital Reference Research Agenda*. *Journal of the American Society for Information Science & Technology*, 2004. **55**(4): p. 301-311.



120. McCracken, N.J., A.R. Diekema, G. Ingersoll, S.C. Harwell, E.E. Allen, O. Yilmazel, and E.D. Liddy. *Modeling reference interviews as a basis for improving automatic QA systems*. in *Proceedings of the Interactive Question Answering Workshop at HLT-NAACL 2006*. 2006. Association for Computational Linguistics. p. 17-24
121. Post, A.R., A.N. Sovarel, and J.H. Harrison Jr. *Abstraction-based temporal data retrieval for a Clinical Data Repository*. in *AMIA Annual Symposium Proceedings*. 2007. American Medical Informatics Association. p. 603
122. Post, A.R. and J.H. Harrison, *Protempa: A method for specifying and identifying temporal sequences in retrospective data for patient selection*. *Journal of the American Medical Informatics Association*, 2007. **14**(5): p. 674-683.
123. Wang, T.D., C. Plaisant, A.J. Quinn, R. Stanchak, S. Murphy, and B. Shneiderman. *Aligning temporal data by sentinel events: discovering patterns in electronic health records*. in *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2008. ACM. p. 457-466
124. Lin, J., P. Wu, and E. Abels, *Toward automatic facet analysis and need negotiation: Lessons from mediated search*. *ACM Transactions on Information Systems (TOIS)*, 2008. **27**(1): p. 6.
125. Kahn, M.G., D. Batson, and L.M. Schilling, *Data model considerations for clinical effectiveness researchers*. *Med Care*, 2012. **50**: p. S60-S67.
126. Kahn, M.G., L.M. Schilling, B.M. Kwan, A. Bunting, C. Uhrich, and C. Singleton. *Preparing Electronic Health Records Data for Comparative Effectiveness Studies*. in *Healthcare Informatics, Imaging and Systems Biology (HISB), 2012 IEEE Second International Conference on*. 2012. IEEE. p. 2-2
127. Tao, C., G. Jiang, T.A. Oniki, R.R. Freimuth, Q. Zhu, D. Sharma, J. Pathak, S.M. Huff, and C.G. Chute, *A semantic-web oriented representation of the clinical element model for secondary use of electronic health records data*. *Journal of the American Medical Informatics Association*, 2013. **20**(3): p. 554-562.
128. Chute, C.G. *(1) Obstacles and options for big-data applications in biomedicine: The role of standards and normalizations*. in *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on*. 2012. IEEE. p. 1-1
129. Zhao, L., S.N.L.C. Keung, A. Taweel, E. Tyler, I. Ogunsina, J. Rossiter, B.C. Delaney, K.A. Peterson, F.R. Hobbs, and T.N. Arvanitis, *A Loosely Coupled Framework for Terminology Controlled Distributed EHR Search for Patient Cohort Identification in Clinical Research*. *Studies in health technology and informatics*, 2012. **180**: p. 519.
130. Rea, S., J. Pathak, G. Savova, T.A. Oniki, L. Westberg, C.E. Beebe, C. Tao, C.G. Parker, P.J. Haug, and S.M. Huff, *Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: The SHARPn project*. *Journal of biomedical informatics*, 2012. **45**(4): p. 763-771.
131. Stang, P.E., P.B. Ryan, J.A. Racoosin, J.M. Overhage, A.G. Hartzema, C. Reich, E. Welebob, T. Scarnecchia, and J. Woodcock, *Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership*. *Annals of internal medicine*, 2010. **153**(9): p. 600-606.
132. Dowdy, D., C. Dye, and T. Cohen, *Data needs for evidence-based decisions: a tuberculosis modeler's wish list [Review article]*. *The International Journal of Tuberculosis and Lung Disease*, 2013. **17**(7): p. 866-877.

133. Carpenter, W.R., A.-M. Meyer, A.P. Abernethy, T. Stürmer, and M.R. Kosorok, *A framework for understanding cancer comparative effectiveness research data needs*. Journal of Clinical Epidemiology, 2012. **65**(11): p. 1150-1158.
134. Cimino, J.J., E.J. Ayres, A. Beri, R. Freedman, E. Oberholtzer, and S. Rath, *Developing a Self-Service Query Interface for Re-Using De-Identified Electronic Health Record Data*. Studies in health technology and informatics, 2013. **192**: p. 632.
135. Edinger, T., A.M. Cohen, S. Bedrick, K. Ambert, and W. Hersh. *Barriers to retrieving patient information from electronic health record data: failure analysis from the TREC Medical Records Track*. in *AMIA Annual Symposium Proceedings*. 2012. American Medical Informatics Association. p. 180
136. Wilcox, A.B., D.K. Vawdrey, Y.-H. Chen, B. Forman, and G. Hripcsak. *The evolving use of a clinical data repository: facilitating data access within an electronic medical record*. in *AMIA Annual Symposium Proceedings*. 2009. American Medical Informatics Association. p. 701
137. Kho, A.N., J.A. Pacheco, P.L. Peissig, L. Rasmussen, K.M. Newton, N. Weston, P.K. Crane, J. Pathak, C.G. Chute, S.J. Bielski, I.J. Kullo, R. Li, T.A. Manolio, R.L. Chisholm, and J.C. Denny, *Electronic medical records for genetic research: results of the eMERGE consortium*. Sci Transl Med, 2011. **3**(79): p. 79re1.
138. Price, R.C., D. Huth, J. Smith, S. Harper, W. Pace, G. Pulver, M.G. Kahn, L.M. Schilling, and J.C. Facelli, *Federated Queries for Comparative Effectiveness Research: Performance Analysis*. Studies in health technology and informatics, 2012. **175**: p. 9.
139. Sittig, D.F., B.L. Hazlehurst, J. Brown, S. Murphy, M. Rosenman, P. Tarczy-Hornoch, and A.B. Wilcox, *A survey of informatics platforms that enable distributed comparative effectiveness research using multi-institutional heterogeneous clinical data*. Med Care, 2012. **50**: p. S49-S59.
140. Bayley, K.B., T. Belnap, L. Savitz, A.L. Masica, N. Shah, and N.S. Fleming, *Challenges in Using Electronic Health Record Data for CER Experience of 4 Learning Organizations and Solutions Applied*. Medical Care, 2013. **51**(8): p. S80-S86.
141. Vechtomova, O. and H. Zhang, *Articulating complex information needs using query templates*. Journal of Information Science, 2009. **35**(4): p. 439-452.
142. Hurdle, J.F., S.C. Haroldsen, A. Hammer, C. Spigle, A.M. Fraser, G.P. Mineau, and S.J. Courdy, *Identifying clinical/translational research cohorts: ascertainment via querying an integrated multi-source database*. Journal of the American Medical Informatics Association, 2013. **20**(1): p. 164-171.
143. Anderson, N., A. Abend, A. Mandel, E. Geraghty, D. Gabriel, R. Wynden, M. Kamerick, K. Anderson, J. Rainwater, and P. Tarczy-Hornoch, *Implementation of a deidentified federated data network for population-based cohort discovery*. J Am Med Inform Assoc, 2012. **19**(e1): p. e60-e67.
144. Horvath, M.M., S. Winfield, S. Evans, S. Slopek, H. Shang, and J. Ferranti, *The DEDUCE Guided Query tool: providing simplified access to clinical data for research and quality improvement*. Journal of biomedical informatics, 2011. **44**(2): p. 266-276.
145. Dörk, M., C. Williamson, and S. Carpendale, *Navigating tomorrow's web: From searching and browsing to visual exploration*. ACM Transactions on the Web (TWEB), 2012. **6**(3): p. 13.

146. Jin, J. and P. Szekely. *QueryMarvel: A visual query language for temporal patterns using comic strips*. in *Visual Languages and Human-Centric Computing, 2009. VL/HCC 2009. IEEE Symposium on*. 2009. IEEE. p. 207-214
147. Wongsuphasawat, K., C. Plaisant, M. Taieb-Maimon, and B. Shneiderman, *Querying event sequences by exact match or similarity search: Design and empirical evaluation*. *Interacting with computers*, 2012. **24**(2): p. 55-68.
148. Monroe, M., R. Lan, H. Lee, C. Plaisant, and B. Shneiderman, *Temporal Event Sequence Simplification*. *Visualization and Computer Graphics*, IEEE Transactions on, 2013. **19**(12): p. 2227-2236.
149. Lan, R., H. Lee, A. Fong, M. Monroe, C. Plaisant, and B. Shneiderman, *Temporal search and replace: An interactive tool for the analysis of temporal event sequences*. 2013, Technical Report HCIL-2013-TBD, HCIL, University of Maryland, College Park, Maryland.
150. Olvera-Lobo, M.D. and J. Gutiérrez-Artacho, *Question-answering systems as efficient sources of terminological information: an evaluation*. *Health Information & Libraries Journal*, 2010. **27**(4): p. 268-276.
151. Hruby GW, B.M., Cimino JJ, Gao J, Wilcox AB, Hirschberg J, Weng C, *Characterization of the Biomedical Query Mediation Process*, in *AMIA Summits on Translational Science Proceedings*. 2013: San Francisco. p. 89-93.
152. Conway, M., R.L. Berg, D. Carrell, J.C. Denny, A.N. Kho, I.J. Kullo, J.G. Linneman, J.A. Pacheco, P. Peissig, and L. Rasmussen, *Analyzing the heterogeneity and complexity of Electronic Health Record oriented phenotyping algorithms*, in *AMIA Annual Symposium Proceedings*. 2011, American Medical Informatics Association. p. 274-283.
153. Combi, C., G. Pozzi, and R. Rossato, *Querying temporal clinical databases on granular trends*. *Journal of biomedical informatics*, 2012. **45**(2): p. 273-291.
154. Moskovitch, R. and Y. Shahar. *Medical temporal-knowledge discovery via temporal abstraction*. in *AMIA*. 2009. p. 453-456
155. Program, S.L.D.S.G., *Centralized vs. Federated: State Approaches to P-20W Data Systems*, I.o.E. Sciences, Editor. 2012. p. 6.
156. Wilcox, A., G. Randhawa, P. Embi, H. Cao, and G.J. Kuperman, *Sustainability considerations for health research and analytic data infrastructures*. *Egems*, 2014. **2**(2).
157. Fleurence, R.L., L.H. Curtis, R.M. Califf, R. Platt, J.V. Selby, and J.S. Brown, *Launching PCORnet, a national patient-centered clinical research network*. *Journal of the American Medical Informatics Association*, 2014. **21**(4): p. 578-582.
158. Christensen, T. and A. Grimsmo, *Instant availability of patient records, but diminished availability of patient information: a multi-method study of GP's use of electronic patient records*. *BMC medical informatics and decision making*, 2008. **8**(1): p. 12.
159. Chute, C.G., M. Ullman-Cullere, G.M. Wood, S.M. Lin, M. He, and J. Pathak, *Some experiences and opportunities for big data in translational research*. *Genetics in Medicine*, 2013. **15**(10): p. 802-809.
160. Yu, C., D. Hanauer, B.D. Athey, and H. Jagadish. *Simplifying access to a Clinical Data Repository using schema summarization*. in *AMIA... Annual Symposium proceedings/AMIA Symposium. AMIA Symposium*. 2006. p. 1163-1163
161. Sujansky, W., *Heterogeneous database integration in biomedicine*. *Journal of biomedical informatics*, 2001. **34**(4): p. 285-298.

162. Smith, S.W. and R. Koppel, *Healthcare information technology's relativity problems: a typology of how patients' physical reality, clinicians' mental models, and healthcare information technology differ*. Journal of the American Medical Informatics Association, 2014. **21**(1): p. 117-131.
163. Chute, C.G., S.A. Beck, T.B. Fisk, and D.N. Mohr, *The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data*. Journal of the American Medical Informatics Association, 2010. **17**(2): p. 131-135.
164. Greim, J., D. Housman, A. Turchin, B. Orlowitz, M. Eskin, A. Abend, J. Isikoff, and J. Einbinder. *The quality data warehouse: delivering answers on demand*. in *AMIA Annual Symposium Proceedings*. 2006. American Medical Informatics Association. p. 934
165. Kamal, J., J. Liu, M. Ostrander, J. Santangelo, R. Dyta, P. Rogers, and H.S. Mekhjian. *Information warehouse—a comprehensive informatics platform for business, clinical, and research applications*. in *AMIA Annual Symposium Proceedings*. 2010. American Medical Informatics Association. p. 452
166. Lyman, J.A., K. Scully, and J.H. Harrison, *The development of health care data warehouses to support data mining*. Clinics in Laboratory Medicine, 2008. **28**(1): p. 55-71.
167. Wiesenauer, M., C. Johner, and R. Röhrig, *Secondary use of clinical data in healthcare providers—an overview on research, regulatory and ethical requirements*. Stud Health Technol Inform, 2012. **180**: p. 614-8.
168. Zerhouni, E.A., *Translational and clinical science—time for a new vision*. New England Journal of Medicine, 2005. **353**(15): p. 1621-1623.
169. Shurin, S.B., *Clinical Translational Science Awards: Opportunities and Challenges*. Clinical and translational science, 2008. **1**(1): p. 4-4.
170. Cimino, J.J. and E.J. Ayres, *The clinical research data repository of the US National Institutes of Health*. Studies in health technology and informatics, 2010. **160**(Pt 2): p. 1299.
171. Del Rio, S. and D.R. Setzer, *High yield purification of active transcription factor IIIA expressed in E. coli*. Nucleic acids research, 1991. **19**(22): p. 6197-6203.
172. Murphy, S.N., G. Weber, M. Mendis, V. Gainer, H.C. Chueh, S. Churchill, and I. Kohane, *Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2)*. Journal of the American Medical Informatics Association, 2010. **17**(2): p. 124-130.
173. Pennington, J.W., B. Ruth, M.J. Italia, J. Miller, S. Wrazien, J.G. Loutrel, E.B. Crenshaw, and P.S. White, *Harvest: an open platform for developing web-based biomedical data discovery and reporting applications*. Journal of the American Medical Informatics Association, 2014. **21**(2): p. 379-383.
174. Danford, C.P., M.M. Horvath, W.E. Hammond, and J.M. Ferranti. *Does access modality matter? Evaluation of validity in reusing clinical care data*. in *AMIA Annual Symposium Proceedings*. 2013. American Medical Informatics Association. p. 278
175. Hruby, G.W., M.R. Boland, J.J. Cimino, J. Gao, A.B. Wilcox, J. Hirschberg, and C. Weng, *Characterization of the biomedical query mediation process*. AMIA Summits on Translational Science Proceedings, 2013. **2013**: p. 89.
176. Brown, P.J. and V. Warmington, *Data quality probes—exploiting and improving the quality of electronic patient record data and patient care*. International journal of medical informatics, 2002. **68**(1): p. 91-98.

177. Wakefield, D.S., K. Clements, B.J. Wakefield, J. Burns, and K. Hahn-Cover, *A framework for analyzing data from the electronic health record: verbal orders as a case in point*. The Joint Commission Journal on Quality and Patient Safety, 2012. **38**(10): p. 444-451.
178. Gallagher, S.A., A.B. Smith, J.E. Matthews, C.W. Potter, M.E. Woods, M. Raynor, E.M. Wallen, W.K. Rathmell, Y.E. Whang, and W.Y. Kim. *Roadmap for the development of the University of North Carolina at Chapel Hill Genitourinary OncoLogic Database—UNC GOLD*. in *Urologic Oncology: Seminars and Original Investigations*. 2014. Elsevier. p. 32. e1-32. e9
179. Durand-Zaleski, I., F. Roudot-Thoraval, J. Rymer, J. Rosa, and J. Revuz, *Reducing unnecessary laboratory use with new test request form: example of tumour markers*. The Lancet, 1993. **342**(8864): p. 150-153.
180. Durieux, P., P. Ravaud, R. Porcher, Y. Fulla, C.-S. Manet, and S. Chaussade, *Long-term impact of a restrictive laboratory test ordering form on tumor marker prescriptions*. International journal of technology assessment in health care, 2003. **19**(01): p. 106-113.
181. Henderson, A., *The test request form: a neglected route for communication between the physician and the clinical chemist?* Journal of clinical pathology, 1982. **35**(9): p. 986-998.
182. Dattani, N., P. Hardelid, J. Davey, R. Gilbert, N. Modi, J. Kurinczuk, A. McMahon, J. Thain, J. Vohra, and A. Macfarlane, *Accessing electronic administrative health data for research takes time*. Archives of disease in childhood, 2013. **98**(5): p. 391-392.
183. Jackson, J.H., B. Gutierrez, O.E. Lunacsek, and S. Ramachandran, *Better asthma management with advanced technology: creation of an Asthma Utilization Rx Analyzer (AURA) Tool*. Pharmacy and Therapeutics, 2009. **34**(2): p. 80.
184. Xiaoli Huang, M., J. Lin, and D. Demner-Fushman. *Evaluation of PICO as a Knowledge Representation for Clinical Questions*. in *American Medical Informatics Association*. 2006. Washington, D.C. p. 359-363
185. Loke, Y.K., *Use of databases for clinical research*. Archives of disease in childhood, 2014. **99**(6): p. 587-589.
186. Murphy, S.N., M. Mendis, K. Hackett, R. Kuttan, W. Pan, L. Phillips, V. Gainer, D. Berkowicz, J.P. Glaser, and I.S. Kohane. *Architecture of the open-source clinical research chart from Informatics for Integrating Biology and the Bedside*. in *AMIA*. 2007. p. 548-552
187. Hruby GW, W., A, Weng C, *Analysis of Query Negotiation between a Researcher and a Query Expert*, in *AMIA*. 2012: Chicago. p. 1780.
188. Rasmussen, L.V., *The electronic health record for translational research*. Journal of cardiovascular translational research, 2014. **7**(6): p. 607-614.
189. Hanauer, D.A., G.W. Hruby, D.G. Fort, L.V. Rasmussen, E.A. Mendonça, and C. Weng, *What Is Asked in Clinical Data Request Forms? A Multi-site Thematic Analysis of Forms Towards Better Data Access Support*, in *AMIA Annual Symposium Proceedings*. 2014, American Medical Informatics Association. p. 616-625.
190. Bernstein, C.N., J.F. Blanchard, P. Rawsthorne, and A. Wajda, *Epidemiology of Crohn's disease and ulcerative colitis in a central Canadian province: a population-based study*. American journal of epidemiology, 1999. **149**(10): p. 916-924.
191. Hruby GW, C.J., Patel VL, Weng C, *Toward a Cognitive Task Analysis for Biomedical Query Mediation*, in *2014 Summit on Clinical Research Informatics*. 2014. p. 218-222.

192. Clark, R.E. and F. Estes, *Cognitive task analysis for training*. International Journal of Educational Research, 1996. **25**(5): p. 403-417.
193. Chipman, S.F., J.M. Schraagen, and V.L. Shalin, *Introduction to cognitive task analysis*. Cognitive task analysis, 2000: p. 3-23.
194. Fosså, S.D., Y. Nilssen, R. Kvåle, E. Hernes, K. Axcrona, and B. Møller, *Treatment and 5-year survival in patients with nonmetastatic prostate cancer: the norwegian experience*. Urology, 2014. **83**(1): p. 146-153.
195. Padwal, R., M. Lin, M. Etminan, and D.T. Eurich, *Comparative Effectiveness of Olmesartan and Other Angiotensin Receptor Blockers in Diabetes Mellitus Retrospective Cohort Study*. Hypertension, 2014. **63**(5): p. 977-983.
196. Rodrigues, G., A. Warner, J. Zindler, B. Slotman, and F. Lagerwaard, *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): p. 52-58.
197. Hoffart, N., *A member check procedure to enhance rigor in naturalistic research*. Western Journal of Nursing Research, 1991. **13**(4): p. 522-534.
198. Weston, C., T. Gandell, J. Beauchamp, L. McAlpine, C. Wiseman, and C. Beauchamp, *Analyzing interview data: The development and evolution of a coding system*. Qualitative sociology, 2001. **24**(3): p. 381-400.
199. Lawshe, C.H., *A quantitative approach to content validity I*. Personnel psychology, 1975. **28**(4): p. 563-575.
200. Wilson, F.R., W. Pan, and D.A. Schumsky, *Recalculation of the critical values for Lawshe's content validity ratio*. Measurement and Evaluation in Counseling and Development, 2012. **45**(3): p. 197-210.
201. Shortliffe, E.H. and V.L. Patel, *Human-Intensive Techniques*, in *Clinical Decision Support: the Road Ahead*. 2007. p. 207-26.
202. Kannampallil, T.G., A. Franklin, R. Mishra, K.F. Almoosa, T. Cohen, and V.L. Patel, *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): p. 21-29.
203. Hoxha, J., P. Chandar, Z. He, J. Cimino, D. Hanauer, and C. Weng, *DREAM: Classification scheme for dialog acts in clinical research query mediation*. Journal of biomedical informatics, 2016. **59**: p. 89-101.
204. Weiskopf, N.G., *Enabling the Reuse of Electronic Health Record Data through Data Quality Assessment and Transparency*. 2015 (Unpublished), Columbia University: Graduate School of Arts and Sciences.
205. Shenvi, E.C., D. Meeker, and A.A. Boxwala, *Understanding data requirements of retrospective studies*. International journal of medical informatics, 2014. **84**(1): p. 76-84.
206. von Eschenbach, A.C. and K. Buetow, *Cancer informatics vision: caBIG™*. Cancer informatics, 2006. **2**: p. 22.
207. Covitz, P.A., F. Hartel, C. Schaefer, S. De Coronado, G. Fragoso, H. Sahni, S. Gustafson, and K.H. Buetow, *caCORE: a common infrastructure for cancer informatics*. Bioinformatics, 2003. **19**(18): p. 2404-2412.
208. Rao, P., A. Andrei, A. Fried, D. Gonzalez, and D. Shine, *Assessing quality and efficiency of discharge summaries*. American Journal of Medical Quality, 2005. **20**(6): p. 337-343.

209. Weng, C., X. Wu, Z. Luo, M.R. Boland, D. Theodoratos, and S.B. Johnson, *EliXR: an approach to eligibility criteria extraction and representation*. Journal of the American Medical Informatics Association, 2011. **18**(Suppl 1): p. i116-i124.
210. Hartigan, J.A. and M.A. Wong, *Algorithm AS 136: A k-means clustering algorithm*. Applied statistics, 1979: p. 100-108.
211. Lopetegui, M.A., S. Bai, P.-Y. Yen, A. Lai, P. Embi, and P.R. Payne. *Inter-Observer Reliability Assessments in Time Motion Studies: The Foundation for Meaningful Clinical Workflow Analysis*. in *AMIA Annual Symposium Proceedings*. 2013. American Medical Informatics Association. p. 889
212. Barbour, R.S., *Checklists for improving rigour in qualitative research: a case of the tail wagging the dog?* British medical journal, 2001. **322**(7294): p. 1115.
213. Mehmood, K. and S.S.-S. Cherfi, *Evaluating the functionality of conceptual models*, in *Advances in Conceptual Modeling-Challenging Perspectives*. 2009, Springer. p. 222-231.

## **Appendix A**

### **1.1 Semi-structured interview**

#### **a. Introduction**

What is your name?

What institution are you affiliated with?

What is today's date?

How long have you been a query analyst?

#### **b. 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.

What steps do you take?

What are the goals of the steps you use?

What information or knowledge do you use at each step? What is the source of this knowledge (e.g., expert, manager, or terminology dictionary)?

#### **c. 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[195]

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[194]

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[196]

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.

#### **d. Phase Three**

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:

Can you describe if this particular task may be related to or fit into the biomedical query mediation task list presented to you?

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:

Does this task represent a part of the process that you use? Why or why not?

Is there adequate information to describe this task?

Dose this task require additional sub-tasks?

Is the goal of this task accurate?

Is additional knowledge needed to perform this task?

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:

What are the sub-tasks used to complete this task?

What is the goal of this task?

What knowledge is needed to perform this task?

### **e. Figures and Tables for the biomedical query mediation task**

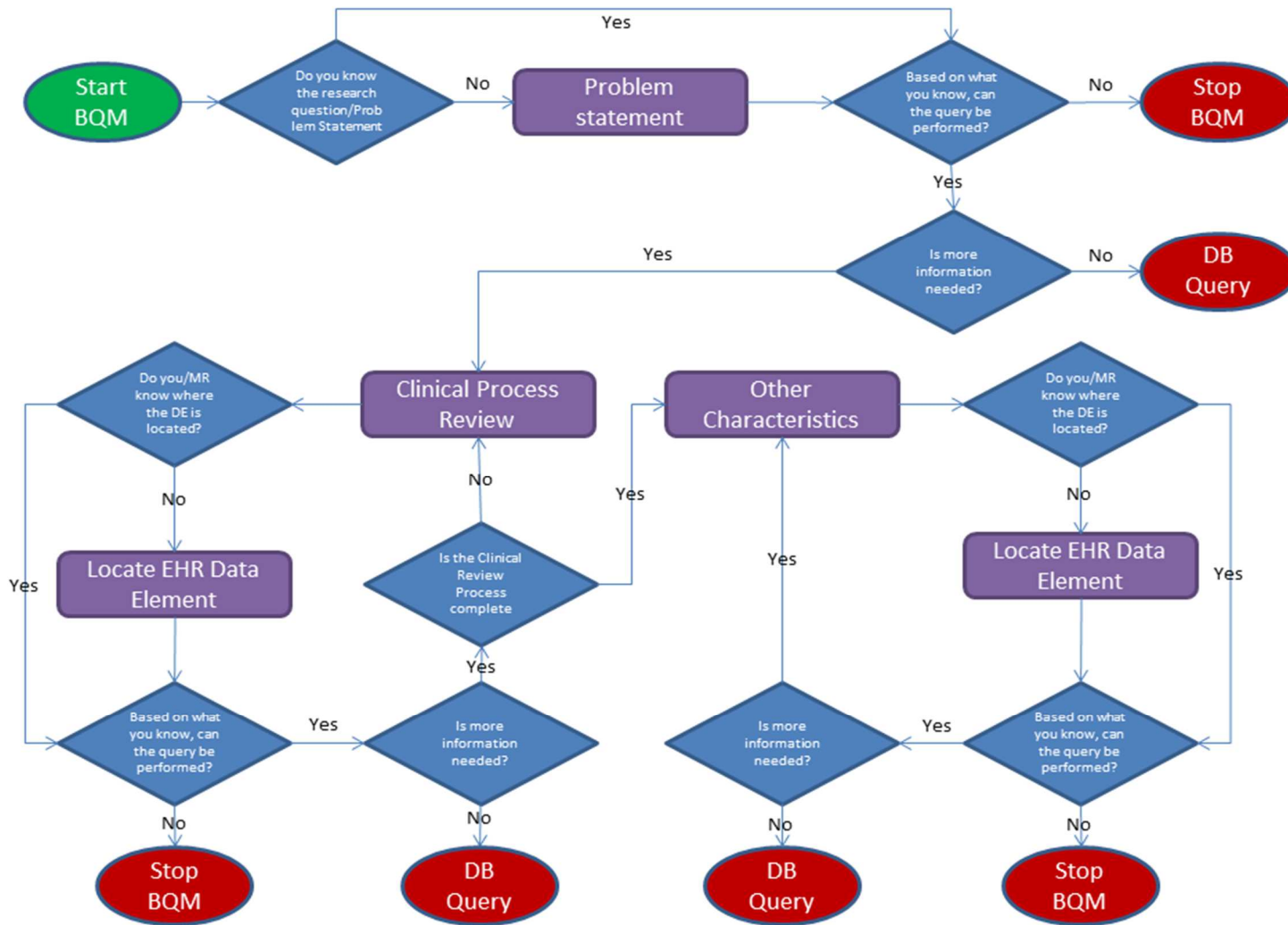
The figure and table presented herein is a biomedical query mediation task list derived from a single-institutions experience of biomedical query mediation. The task list underwent a face and content evaluation by seven subject matter experts. The task list was found to have face validity. 4/10 of the sub-tasks, sub-tasks 2.2, 2.4, 4.2, and 5.1, were judged to have content validity. All the sub-tasks were judged to have either valid or semi-valid content validity.

Appendix Table 1 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. Appendix Figure 0-1 organizes the tasks presented I Appendix Table 1 into a workflow the query analyst would follow during BQM.

Task	Sub-task	Goal	Knowledge Required	Example
<b>Define research statement</b>	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>re you looking at pre-treatment factors that affect the outcome measure?</i>
<b>Illustrate clinical process</b>	2.1 Elicit the clinical progression related to the information need	To establish the temporal order of abstract data elements	Medical domain knowledge	<i>Patients with disease <math>x</math> that undergo treatment <math>y</math>, can 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 <math>X</math> as <math>x</math>? 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 <math>X</math>? 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 <math>x</math> and <math>y</math> using this formula, <math>x/y + 5</math>.</i>
<b>Identify related data elements</b>	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>
<b>Locate EHR data elements</b>	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 <math>X</math>, 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 <math>X</math> is not collected in the EHR; Data element <math>Y</math> is available sporadically from patient to patient.</i>
<b>End mediation</b>	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>

**Appendix Table 1. Biomedical query mediation tasks and activities performed by the data analyst**

Appendix Figure 0-1. BQM Task Flowchart



## **2.1 Query Analyst Workflows used during Member checking**

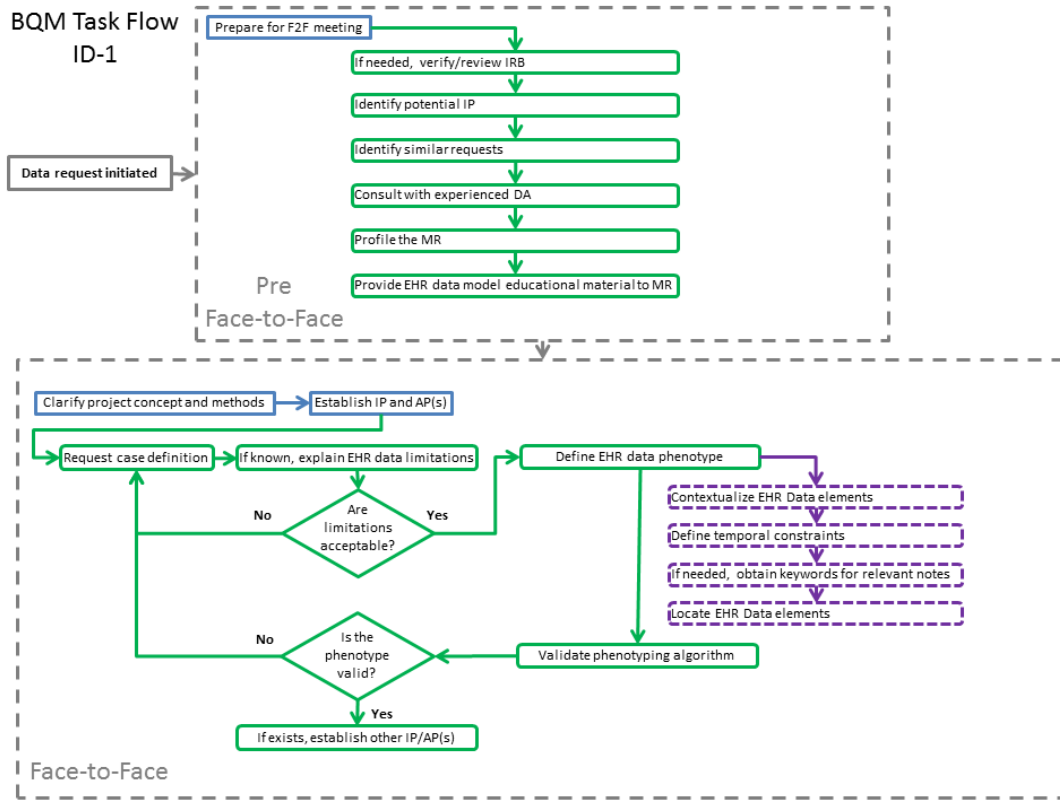
### **a. Definitions**

- i. MR – Medical Researcher
- ii. DA – Data Analysts
- iii. IP – Index Phenotype
- iv. AP – Associated Phenotypes
- v. EHR – Electronic Health Record

### **b. Validation Questions**

- i. Does this represent your process?
- ii. How would you change the language used to describe this process?
- iii. What is missing, what would you remove?

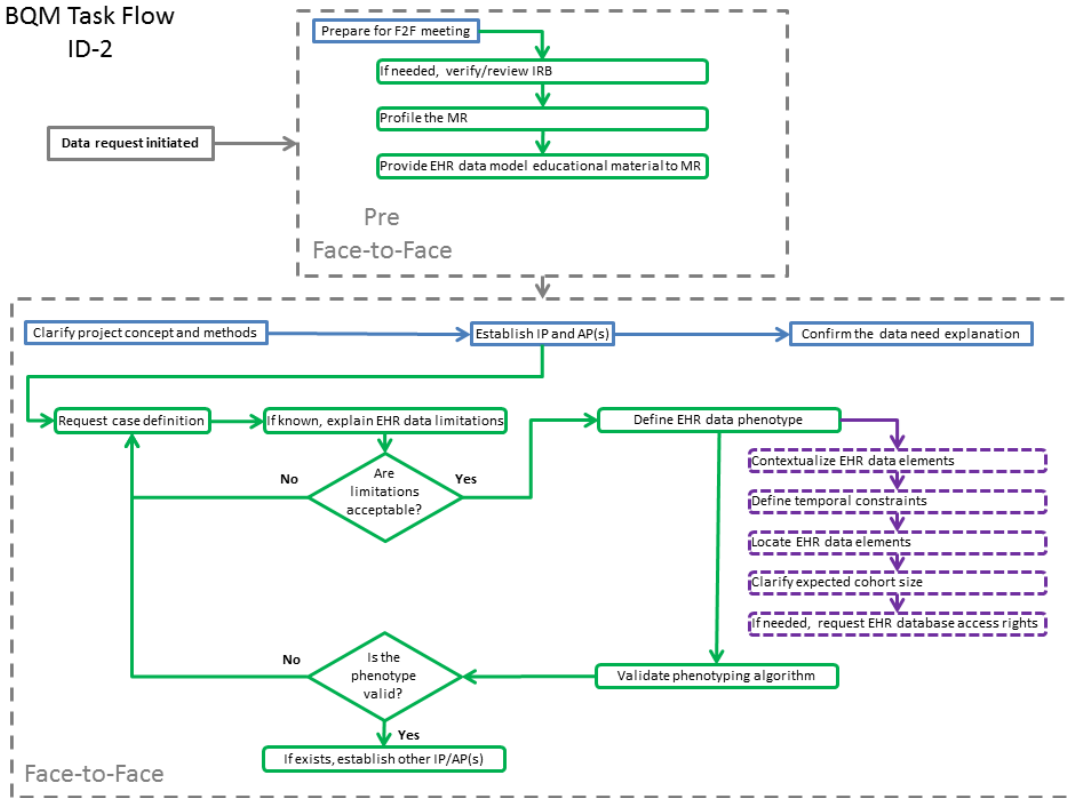
### c. Interview ID\_1



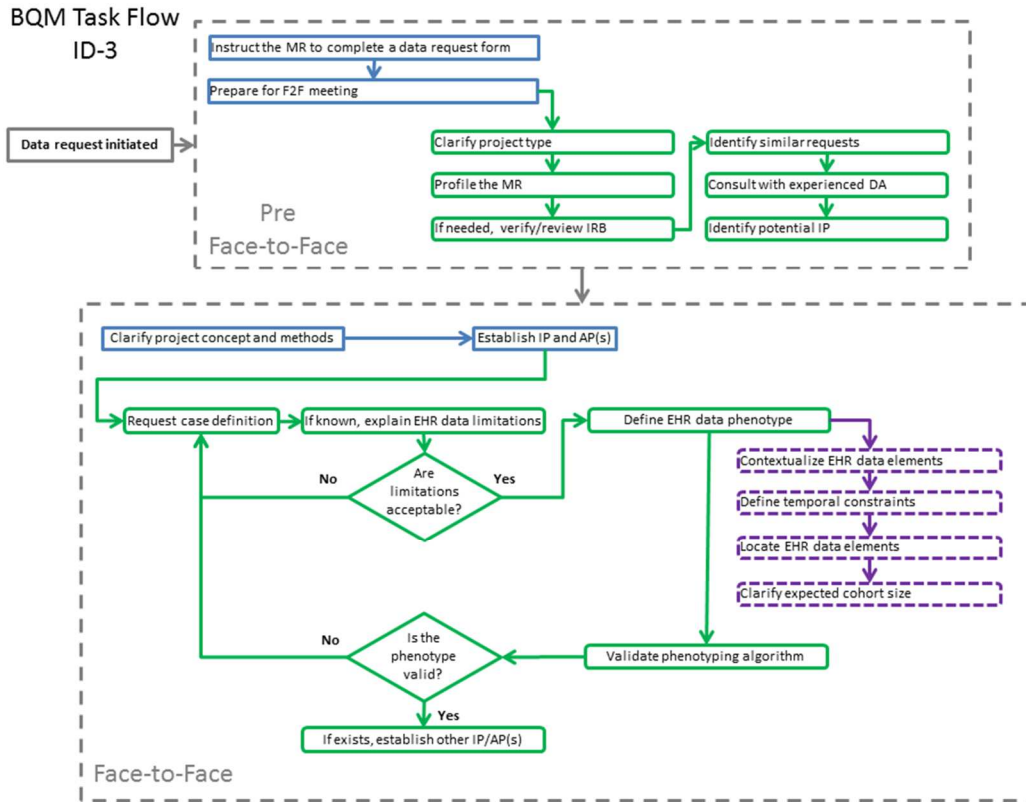


## d. Interview ID\_2

BQM Task Flow  
ID-2

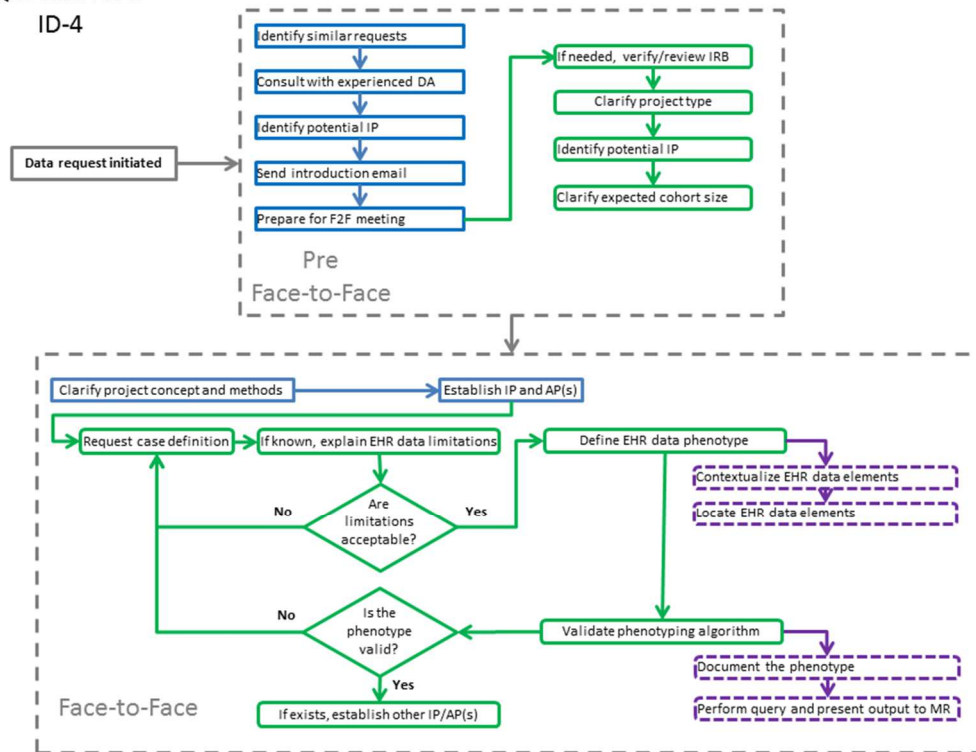


### e. Interview ID\_3



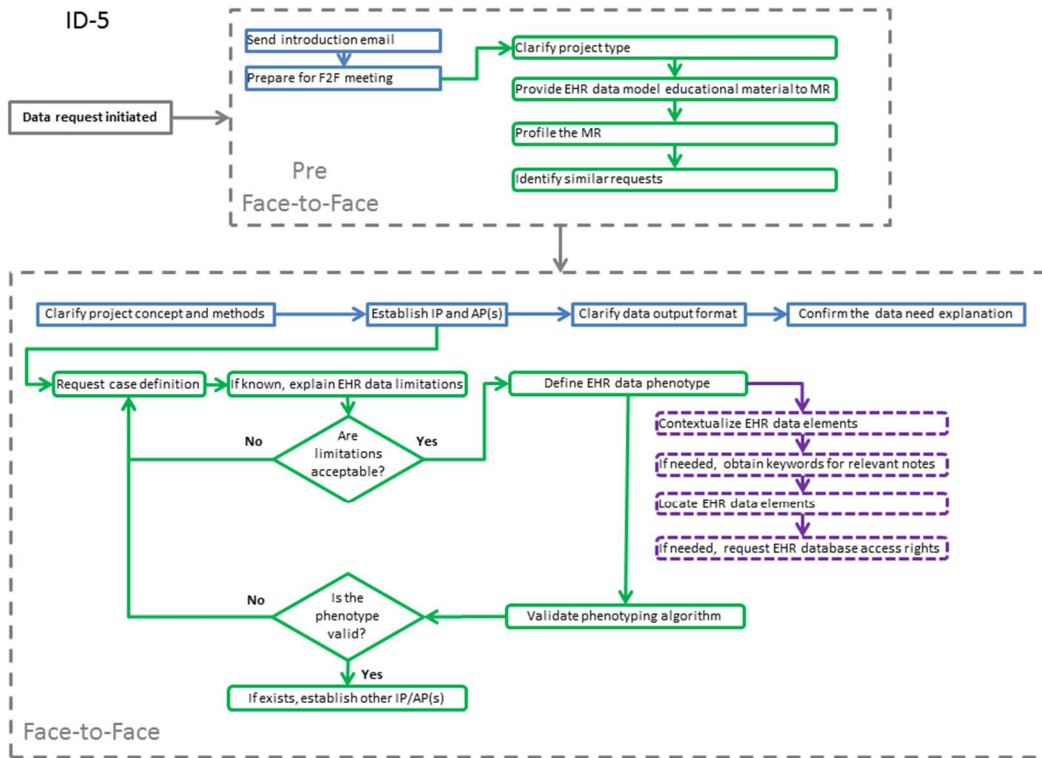
## f. Interview ID\_4

BQM Task Flow  
ID-4



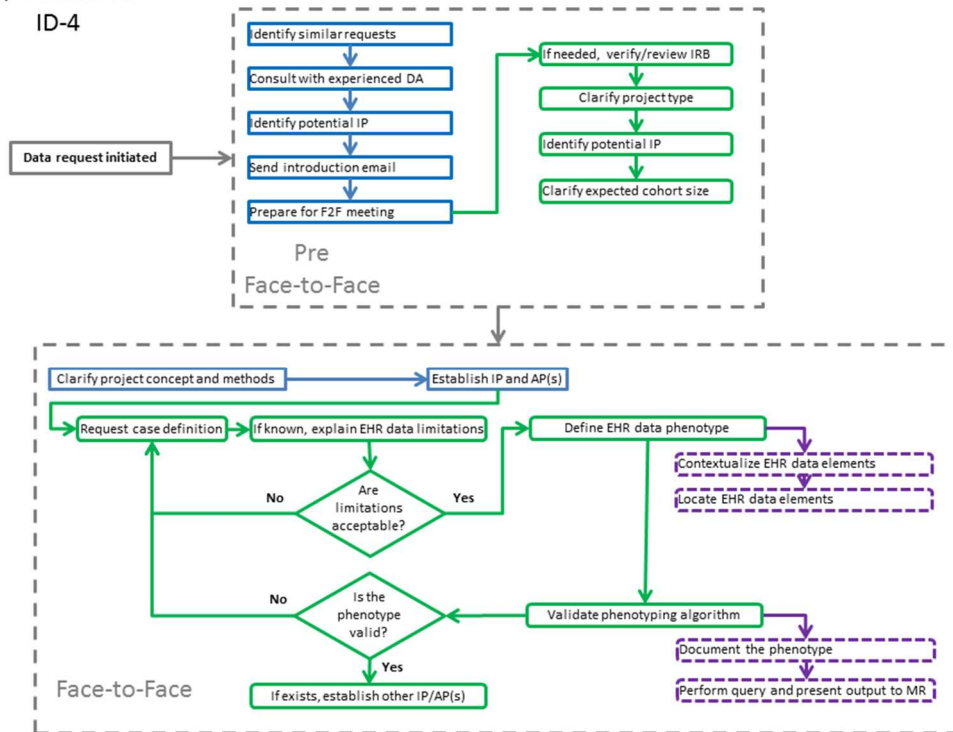
## g. Interview ID\_5

BQM Task Flow  
ID-5



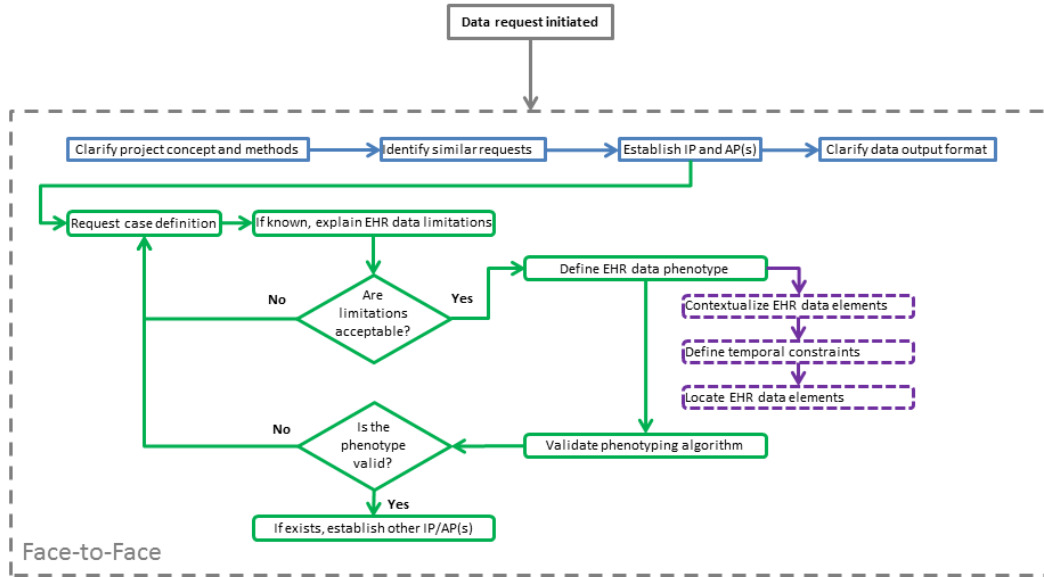
## h. Interview ID\_6

BQM Task Flow  
ID-4



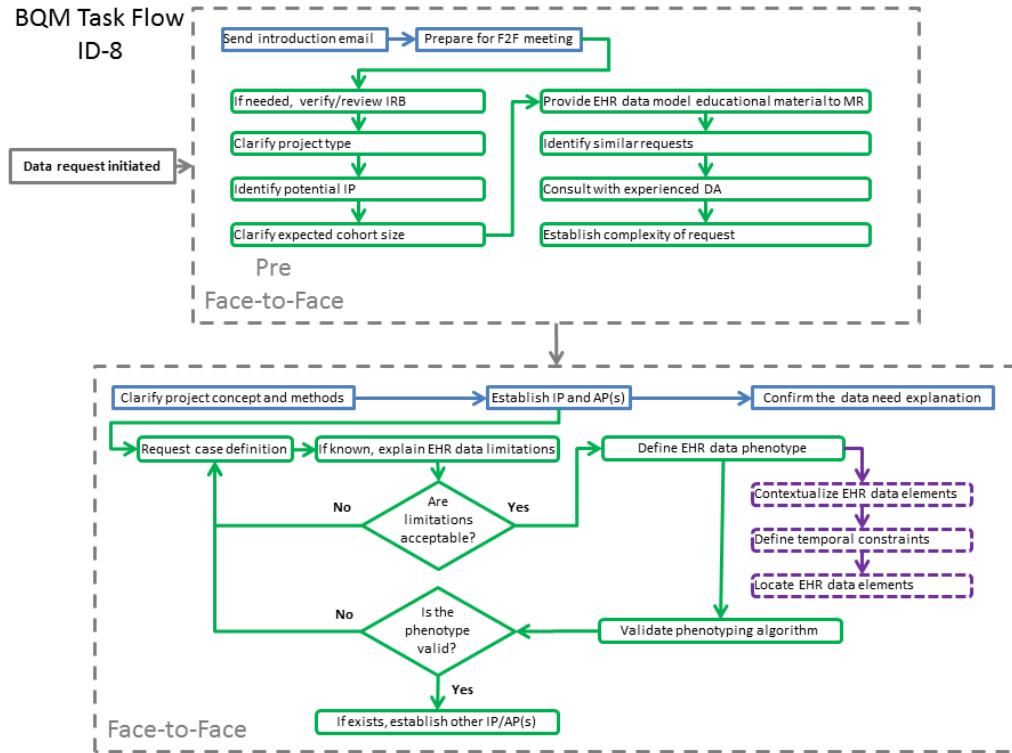
# i. Interview ID\_7

BQM Task Flow  
ID-7



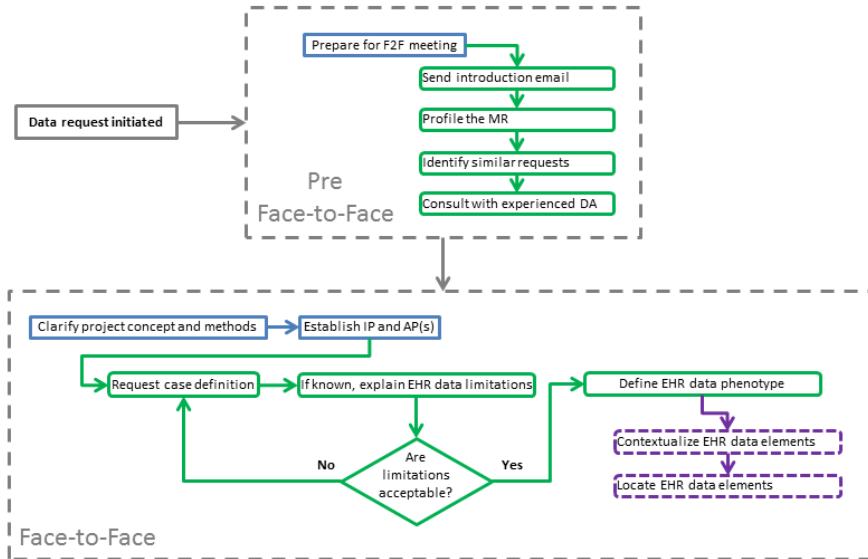
## j. Interview ID\_8

BQM Task Flow  
ID-8



## k. Interview ID\_9

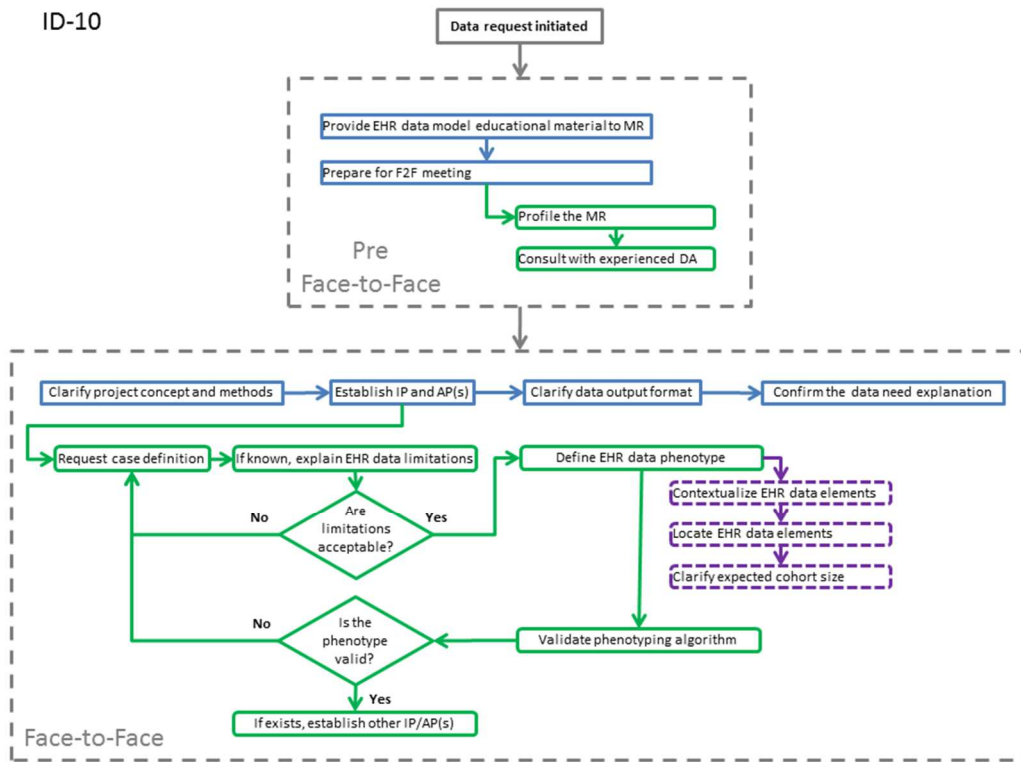
BQM Task Flow  
ID-9





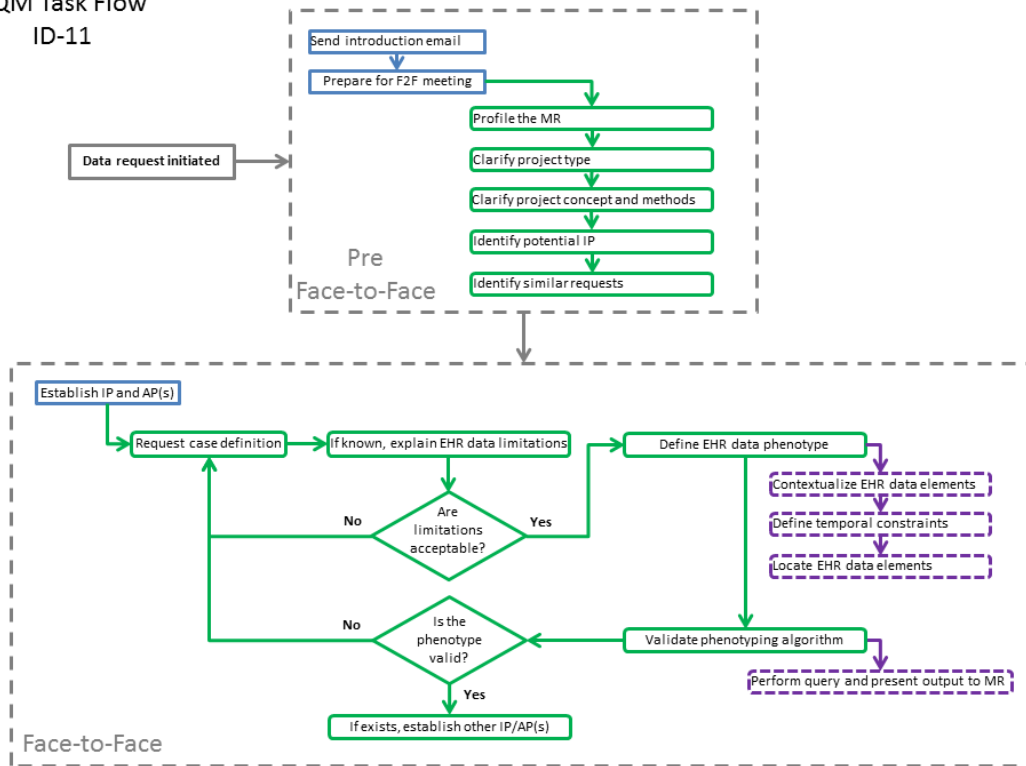
# I. Interview ID\_10

BQM Task Flow  
ID-10



## m. Interview ID\_11

BQM Task Flow  
ID-11



## **Appendix B**

### **1.1 Adaptations made to the original Carpenter Framework using each dataset**

Appendix Table 2 provides a detailed description for the augmentations to the sections and concepts that occurred during the annotation of the datasets. In summary, the clinical trials sentences dataset retained 23/63 classes from the original model and added 11 new classes. The EHR data request dataset retained 37/63 classes and added 13 new classes. The SQL project dataset retained 47 classes and added 11 new classes. Overall, The enriched concept schema retained 57/63 classes and added 15 new classes. We excluded the following six classes from the Carpenter model as no data elements from the three data sources were mapped to these classes: Health resources, Guidelines, Care systems and coordination, Other, Symptom/Side-effect management, and Economic outcome/burden.

	<b>Edit</b>	<b>Add</b>	<b>Delete</b>	<b>Move</b>
<b>Clinical Trial inclusion/Exclusion Criteria</b>				
<b>Sections</b>	Cancer Characteristics -> Detection/Treatment Results		Treatment; Intermediate Outcomes, Outcomes	
<b>Classes</b>		Height/Weight/BMI; Medical/Disease History; Medical/Surgical/Radiation Treatment History; Medical Device Implant; Past/Current Medications; Current Treatment/Experimental Trials; Study Compliance Characteristics; Consent; Life Expectancy; Result/Diagnosis/Description; Severity/Stage/Prognosis; Clinical Stage; Margin; Provider; Location;	Health Resources; Local disease burden; Genetic (Somatic) Characteristics; Training Guidelines; Care systems and coordination	
<b>EHR data request logs</b>				
<b>Sections</b>	Treatment -> Intervention; Cancer Characteristics -> Detection/Treatment Results		Environmental Factors	Organization al /Provider Characteristic s
<b>Classes</b>	Chemotherapy and Hormone Therapy -> Medical Therapy (Chemo/ Hormone/Biologic)	Identification Information; Height/Weight/BMI; Medical/Surgical/Radiation Treatment History; Past/Current Medications; Current Treatment/Experimental Trials; Result/Diagnosis/Description; Severity/Stage/Prognosis; Provider; Location; Provider; Location; Immediate; Surrogate	Genetic (Germ Line) Characteristics; Experience with Patient Population; Experience with Specific Therapies; Genetic (Somatic) Characteristics; Guidelines; Care systems and coordination; Active surveillance; Other; Economic outcome/burden; Symptom/side-effect management; Intent	
<b>EHR SQL Project Queries</b>				
<b>Sections</b>	Treatment -> Intervention; Cancer Characteristics -> Detection/Treatment Results			Organization al /Provider Characteristic s
<b>Classes</b>	Chemotherapy and Hormone Therapy -> Medical Therapy (Chemo/ Hormone/Biologic)	Identification Information; Height/Weight/BMI; Medical/Surgical/Radiation Treatment History; Past/Current Medications; Current Treatment/Experimental Trials; Result/Diagnosis/Description; Severity/Stage/Prognosis; Clinical Stage; Margin Status; Provider; Location; Immediate; Surrogate	Environmental Exposures; Health Resources; Family history; Genetic (Germ Line) Characteristics; Genetic (Somatic) Characteristics; Other; Economic outcome/burden; Symptom/side-effect management;	

**Appendix Table 2. Adaptations made to the original Carpenter framework during each data set annotation.**

## **2.1 Semi-Structured Interview Material**

### **a. Introduction**

Questions:

What is your area of research interest? E.g. Clinical Trials, Prospective, Retrospective research?

For how many years have you been conducting research?

How often do you submit data requests each year?

### **b. Concept Generation and Mapping**

In this block, I evaluate the completeness, generalizability, expressiveness, and understandability of my proposed conceptual model in two steps.

Step One

We present the researcher with a recently published comparative effectiveness research study from the participant's lab. We ask the participant to list at least ten medical concepts from the study.

Step Two

We introduce the model to the expert and have the expert map the concepts listed in step one to the nodes in the conceptual model. During this process, we will instruct the expert to think aloud their actions and explain their decisions.

After the expert has completed concept mapping, we ask a set of follow-up questions to gather additional information:

Concept mapping difficulty follow-up questions:

Would you add any additional granular nodes to the model?

Are we missing modules that would better articulate your data needs?

General Questions:

From your experience conducting research with EHR data, what other data needs are we missing?

Any module components?

Any Granular node within the modules?

### **c. Modeling Structure**

The third block evaluates the structure of the model through a series of questions, which assess the relationships among the modules, as well as the relatedness of the relationship between the parent-child nodes within the modules.

Questions:

For each module pairwise relationship, we ask the following questions:

How do you interpret the relationship between the module components?

Is the relationship (a) interesting (b) uninteresting or (c) I don't know

If interesting, please describe the relationship.

Within each module, we ask the following question:

Please identify ambiguous parent-child relationships?

How would you structure them otherwise?

### 3.1 Schema Class Definitions and Examples

Section	Class	Clinical Trials	EHR Data Requests	SQL Variables	Definition	Example Values
Environmental Factors (n=2)	Environmental Exposures	●			Incorporates concepts that describe a patients potential exposure to elements within the environment they live	Exposure to contaminated water; lead poisoning;
	Social/Health Norms	●		●	Defines the relative habits associated with the socioeconomic community a patient belongs to	The community's perception of healthcare; The recent vaccine opposition
Patient (n=21)	Age	●	●	●	Concepts that define a patients age	Date of birth
	Race/Ethnicity	●	●	●	Patients self-identified based on which they most closely identify	White; Latino; Jewish; African American; Asian
	Identification Information		●	●	Concepts used to identify patients	MRN; First and Last name; Home address; Email; Phone number
	Gender	●	●	●	Concepts that define the sex of the patient	Female; Male
	Height/Weight/BMI	●	●	●	Concepts that describe the body habitus of the patient	Height; weight;
	Family History	●	●		Concepts that describe the patient's family's social and medical history	Paternal cardiovascular disease; Maternal Breast cancer
	Geography	●	●	●	Concepts that define the geographical location the patient lives	Neighborhood; Zip code; State; Country
	Income	●	●	●	Concepts that define the wealth of the patient	Yearly salary
	Insurance Status	●	●	●	Concepts that define the level of health care insurance the patient has	Medicare; Medicaid; Private Insurance; Uninsured
	Patient Reported Outcomes	●	●	●	Concepts describing the patients perception of their heal status	Play golf; SF-36; Urinary function;
	Performance Status	●			Concepts used by physicians to objectively measure the general well-being and activities of daily	Karnofsky Score; Zubrod Score; Lansky Score



					life that may be used to augment potential treatments.	
	Medical/Disease History	●			Concepts defining the medical/disease history of the patient regardless of its association with the patient's current complaint	History of Diabetes
	Medical/Surgical/Radiation Treatment History	●	●	●	Concepts defining the treatment history of the patient regardless of its association with the patient's current complaint	Prior radiation treatment; Prior Chemotherapy
	Medical Device Implant	●			Concepts that describe in foreign devices the patient has	Pace Maker; Cardiac Stent; Artificial Hip; Urinary Catheter;
	Past/Current Medications	●	●	●	Concepts that describe any past or current medications the patient takes	Ibuprofen; Ritalin; Statins;
	Current Treatment/Experimental Trials	●	●	●	Concepts that describe ongoing treatments the patient may be receiving for a disease related to or unrelated to the patient's current complaint	Chemotherapy; Patient is enrolled in a Clinical Trial
	Physical/Mental Health Acuity	●			Concepts used by care providers to measure the sharpness of the mind and body	Cognitively impaired; Disassociation with reality;
	Genetic (Germ Line) Characteristics	●			Concepts that describe DNA mutations transmitted from the parents to the patient	BRCA1&2; SNPs
	Health Behaviors	●	●	●	Concepts that describe the patient's activities that influence their health status	Smoking; Exercise; Sleep; Diet; Alcohol
	Consent	●			Concepts that define the patient's consent for a particular treatment or clinical trial	Signed Consent Forms
	Life Expectancy	●			Concepts that define how long the patient has to live	Expected date of expiration
Detection/Diagnostics (n=3)	Modality of Assessment/Detection	●	●	●	Concepts that describe how a disease or health status was determined	Lab test, Clinical assessment, Radiographical, Procedure
	Intent	●	●	●	Concepts that describe why a diagnostic test was performed	Patient complaint; Suspicious clinical finding; Rule out disease

						X
	Time/Dates	●	●	●	Concepts that describe with the diagnostic procedure was performed	Dates the diagnostic test was performed
Detection/ Treatment Results (n=8)	Result/Diagnosis/Description	●	●	●	Concepts that describe the results of treatments or diagnostic procedures	Lab values; Path reports; Clinical Assessment
	Severity/Stage/Prognosis	●	●	●	Concepts that quantify the severity of the disease	Uncontrolled diabetic; Outcome prediction; Chronic Kidney Disease Stage
	Pathology Stage /Grade	●	●	●	Concepts that describe attributes of a pathology report	Pathologic TNM staging; Neoplastic Grading
	Clinical Stage	●		●	Concepts that stage the current gestalt view of the patient's disease.	Clinical TNM staging;
	Histology/Morphology	●	●	●	Concepts that describe the physical attributes of cells; Microscopic anatomy of cells	Renal Cell Carcinoma; Inflammatory Cells; Lymphocyte Invasion;
	Molecular Markers	●	●	●	Concepts that describe sites of heterozygosity for some type of silent DNA variation	Polymorphisms;
	Genetic (Somatic) Characteristics	●			Concepts that describe genetic mutations that occur in cells that are not inherited from the patient's parents	ATP1A1; p53; FGFR3
	Margin	●		●	Concepts that describe the extent of tumor removal	Positive or Negative Margin
Organizational/ Provider Characteristics (n=6)	Location	●	●	●	Concepts that describe the location of a treatment or diagnostic procedure	Community Hospital; Outside institution; Ambulatory Clinic
	Provider	●	●	●	Concepts that define the diagnosing or treating care provider	Care provider's Name
	Training	●	●	●	Concepts that define the training status of the health care provider	Medical Student; Resident; Fellow; Attending;
	Experience with Patient Population	●		●	Concepts that define the care providers experience treating patient with a particular disease	Care provider exclusively treats diabetic patients and treats 1000/year
	Experience with Specific Therapies	●		●	Concepts that define the care providers experience performing	Dr. X performs 100 cases a year while Dr. Y performs 20/year.

					particular treatment	
	Specific Care Process	●	●	●	Concepts that define the intended/actual care process used by the care provider	Care Guidelines; Guideline adherence
Intervention (n=10)	Surgery		●	●	Concepts that describe surgical therapies	Radical Prostatectomy; Open Heart Surgery; Cystoscopy
	Chemo		●	●	Concepts that describe chemotherapy treatments	MVAC; Cisplatin;
	Hormone		●	●	Concepts that describe Hormone treatments	Estrogen; Testosterone;
	Biologic		●	●	Concepts that describe biologic treatments	Immune therapy; interleukin-2; Colony-stimulating factors
	Radiation Therapy		●	●	Concepts that describe Radiation treatments	External Beam Radiation; Radioiodine ablation
	Active Surveillance			●	Concepts that describe a passive approach to disease treatment	Radiographical monitoring; Lab value monitoring; Tissue Biopsy Monitoring
	Prescribed vs. Delivered		●	●	Concepts that describe what was intended and what was performed	Planned laparoscopic procedure vs. Open procedure per
	Approach/Treatment Details		●	●	Concepts that describe attributes of the procedure	Ischemia time; sutures used; Surgical Equipment used
	Completeness/Duration, Treatment Adherence, Start/Stop Date, Start/Stop Time		●	●	Concepts that describe how long the treatment took	Treatment start and stop date/time
	Primary, Adjuvant, Neo-Adjuvant, Induction, Maintenance, or Salvage Therapy			●	Concepts that describe the treatment intent	Primary, curative; salvage;
Intermediate Outcomes (n=19)	EBL		●	●	Concepts that describing blood loss during a procedure	800cc EBL
	LOS		●	●	Concepts that describe how long the patient recovered in the hospital setting	8 days length of stay from treatment to discharge
	Transfusions		●	●	Concepts that describe if and how many blood transfusions occurred during or after the treatment	2 units packed red blood cells; 1 unit platelets; 1 unit plasma.

Tumor Response		●	●	Concepts that describe a tumors response to a medical therapy	Tumor size reduction demonstrated on CT scan
Disease Progression		●	●	Concepts that assess the progression of a disease to a more sever stage	
Recurrence		●	●	Concepts that describe the detectable recurrence of a disease	Positive lab tests; positive imaging;
Second Malignancies		●	●	Concepts that describe new malignancies identified after treatment	Incidental finding on checkup CT scan
Nausea/Vomiting/Bowel Dysfunction		●	●	Concepts that describe post treatment GI issues	Nausea; Vomiting
Neutropenia/Fever		●	●	Concepts that describe systemic conditions occurring after treatment	Fever; Sepsis
Wound Infection		●	●	Concepts that describe local infections related to the treatment site.	Wound infection
Comorbid Conditions			●	Concepts that describe comorbid conditions effected by the current treatment	Controlled to uncontrolled diabetic;
Physical/Mental health acuity			●	Concepts that describe the physical and mental acuity of the patient effected by the treatment	Cognitively impaired; Disassociation with reality;
Health Behaviors			●	Concepts that describe health behaviors effected by the treatment	Smoking; Exercise; Sleep; Diet; Alcohol
Quality of Life			●	Concepts that describe the patient's perception of their quality of life effected by the treatment	Play golf; SF-36; Urinary function;
General Health Care Use		●	●	Concepts that describe the patient's use of follow up health care after a procedure	Adherence to follow up procedure
Inpatient Hospitalization/ED Use		●		Concepts that describe the patient's use emergent health care settings	Emergency department; Hospital admission
Additional Diagnostic Procedures		●		Concepts that describe additional diagnostic procedures used to confirm positive results	Lab tests, Radiological Exams;
Additional Procedures		●		Concepts that describe additional procedures performed to correct	Procedure

					deficiency with the primary treatment	
	Medication Use		●	●	Concepts that describe any medication used to counter act symptoms caused by the current treatment.	Pain medications; Tylenol
Outcomes(n=3)	Overall Survival		●	●	Concepts that describe the survival status of the patient	Last known date alive; date of death
	Disease Specific Survival		●	●	Concepts that describe the current survival status of the patient related to the disease of interest	Last known date alive; Date of death from other causes; Date of death caused by the disease of interest
	Quality of Life (Overall)		●	●	Concepts that describe the patient's reported quality of life during the period after treatment to end of life	Play golf; SF-36; Urinary function; Functional ability