Characterization of the Biomedical Query Mediation Process

Gregory W. Hruby, MA[1], Mary Regina Boland, MA[1], James J. Cimino, MD[1,2], Junfeng Gao, PhD[1], Adam B. Wilcox, PhD[1], Julia Hirschberg, PhD[3], Chunhua Weng, PhD[1]
1. Department of Biomedical Informatics, Columbia University, New York, NY 10032
2. NIH Clinical Center, Bethesda, MD 20892
3. Department of Computer Science, Columbia University, New York, NY 10027

Abstract

To most medical researchers, databases are obscure black boxes. Query analysts are often indispensable guides aiding researchers to perform mediated data queries. However, this approach does not scale up and is time-consuming and expensive. We analyzed query mediation dialogues to inform future designs of intelligent query mediation systems. Thirty-one mediated query sessions for 22 research projects were recorded and transcribed. We analyzed 10 of these to develop an annotation schema for dialogue acts through iterative refinement. Three coders independently annotated all 3160 dialogue acts. We assessed the inter-rater agreement and resolved disagreement by group consensus. This study contributes early knowledge of the query negotiation space for medical research. We conclude that research data query formulation is not a straightforward translation from researcher data needs to database queries, but rather iterative, process-oriented needs assessment and refinement.

Introduction

The accelerated adoption of electronic health records (EHRs) systems nationwide, fueled in part by administrative initiatives, has made an unprecedented amount of data available.[1,2] Appropriate reuse of such "big data" from healthcare processes is indispensible for achieving comparative effectiveness research (CER).[3] 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. Unfortunately, this approach is not scalable, partly because many requests contain vague or nonspecific concepts that are hard to understand for query analysts, who usually have limited medical domain knowledge. Typically, query analysts have to consult researchers many times via emails or phone calls to clarify the query details.

To relieve the burden on query analysts, a variety of data query tools were developed.[4-8] Notable ones include Informatics for Integrating Biology and Bedside (i2b2),[9-11] the Visual Aggregator and Explorer (VISAGE),[12] and the Stanford Translational Research Integrated Database Environment (STRIDE).[13] I2b2 enables users to drag and drop concepts to construct queries. The modified Web version, SHRINE, also enables federated queries across multiple databases.[9] 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 autonomously. One approach to achieve this goal is to support computer-based reference interviews with biomedical researchers. "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".[14] Query analysts, like librarians, often use a negotiation process to comprehend the needs of the researcher.[15,16] 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 our analysis of the query mediation dialogues between a query analyst and medical researchers and our findings of the characteristics of the biomedical query mediation process. This study extends the work of a previous poster presented at the 2012 AMIA Fall Symposium entitled, "Analysis of Query Negotiation between a Researcher and a Query Expert."[17]

Data and Methods

1. Data: Between July 2011 and January 2012, we recorded and transcribed 31 discussions for 22 medical research projects between one query expert (QE) and eight medical researchers (MRs) at the Columbia University Department of Urology. The Columbia University Medical Center Institutional Review Board approved this study

(**IRB-AAAJ8850**). **Figure 1** shows 5 example dialogue acts. In the context of this paper, a dialogue act is one exchange of speech. We arrived at 3160 dialog acts for the 31 query mediation sessions.

QE: Alright. So we're going to be talking about your study so I guess briefly describe to me what you want to do.

MR: 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

QE: Uh huh

MR: 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

QE: So we're looking at the effect of urethral involvement, urethral or ureteral?

Figure 1. Example Dialogue Acts

2. Annotation Schema Development: We used the dialogue acts from 10 randomly selected projects to develop a dialogue act classification schema. We first derived the common tasks of dialogue acts, such as understanding the clinical process, identifying available data, and explaining data characteristics. Then, we 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. We decided to classify dialogue acts along the "Stages of Mediation" aspect in order to see the temporal patterns of dialogue acts. We iteratively designed and tested a classification schema on sample transcripts and finalized the schema with group consensus among three independent raters.

3. Dialogue Act Annotation: Three raters (GH, MB, JG) independently annotated all the 3160 dialogue acts. Each dialogue act was annotated with at least one classification code. We assessed inter-rater agreement with the kappa statistic. For dialogue acts with inter-rater disagreement, we reached consensus by accepting the pair-wise consensus between GH and MB first, GH and JG next, and MB and JG last. We resolved the remaining disagreements of the clinical content of dialogue acts with GH codes.

4. Data Analysis: We used the consensus annotation results for further dialogue flow analysis. We 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. We aggregated the annotated content of the 22 projects into one representation of the negotiation space. We used descriptive statistics and graphs to visualize this space.

Results

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 we 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.

1.0 State the Problem (e.g., *"Alright, So we're going to be talking about your study so I guess briefly describe to me what you want to do.")*
2.0 Explain the Clinical Process (e.g., "*And then they are diagnosed with cancer after the image?*")
    2.1 Patients Demographics
    2.2 Temporal Aspect of clinical process
        2.2.1 Initial Diagnosis of Disease
        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 Radiograhical studies
    2.5 Clinical findings
        2.5.1 Disease Confounders and Comorbidities
        2.5.2 Social History
        2.5.3 Family History
        2.5.4 Clinical Stage/Risk assessment/Disease Status of Diagnosis
        2.5.5 Survival: Disease Specific, and Overall survival
    2.6 Surgical Procedure
    2.7 Pathology
    2.8 Medical Therapy
    2.9 Radiation Therapy
    2.10 Other Treatment
    2.11 Treatment Toxicities, Complications, and Adverse events
3.0 Locate Data Elements in EHRs (e.g., *"You will have to look in the operative note."*)
4.0 Discuss Study Design (e.g., *"Because we want to exclude any disease that could potentially have an effect on the GFR."*)
5.0 Clarify Research Workflow (e.g., *"It's gonna be rare. So you're probably gonna have to update it as well."*)
6.0 Explain Data Results to Researchers (e.g., *" So follow-up is last time known alive. So this is corresponding to overall survival information."*)
7.0 Review IRB and Privacy Policies (e.g., *"It is expedited because it is de-identified."*)
8.0 Confirm Completed Process (e.g., *"Alright. I think we have enough information."*)

Figure 2. The Classification schema for Dialogue Acts in Query Mediation

These served as the basis for our coding book (**Figure 2**). Tasks were iteratively organized into a hierarchical structure (**Figure 2**) to be used to describe the dialogue acts of the mediation process between the QE and MR. **Figure 2** shows the coding schema we used to annotate the dialogue acts of the query negotiation space. Our inter-rater kappa score over all the dialogue acts was 0.61.

2. Temporal Distribution of Dialogue Act Classes

Figure 3 illustrates the broad variety of issues discussed between a QE and MR. This figure also represent 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 QE and the MR discuss IRB and privacy policy.

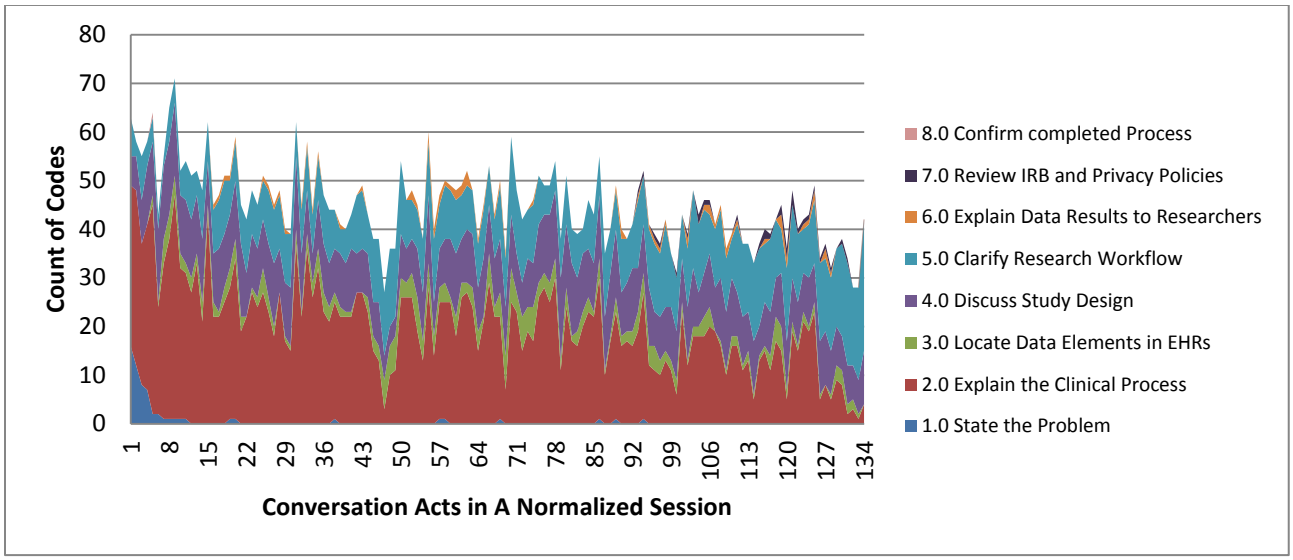

Figure 3. Temporal Distribution of Dialogue Acts in a Normalized Mediated Query Conversation Session

3. A closer look on the Discussion of the Clinical Process

Figure 4 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.

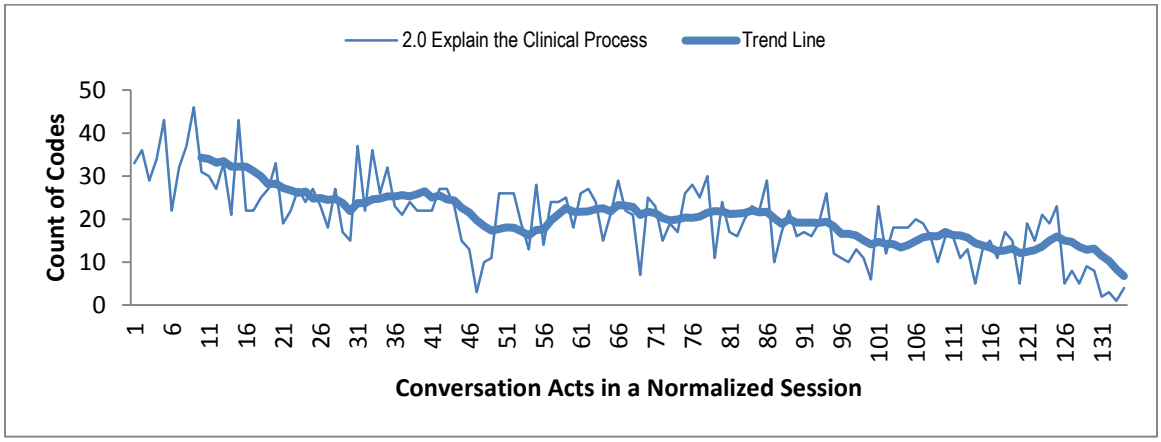

Figure 4. Discussion of the Clinical Process over the Course of a Normalized Conversation Session*4. Temporal Distribution of Study Design and Research Workflow Discussions*

Figure 5 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.

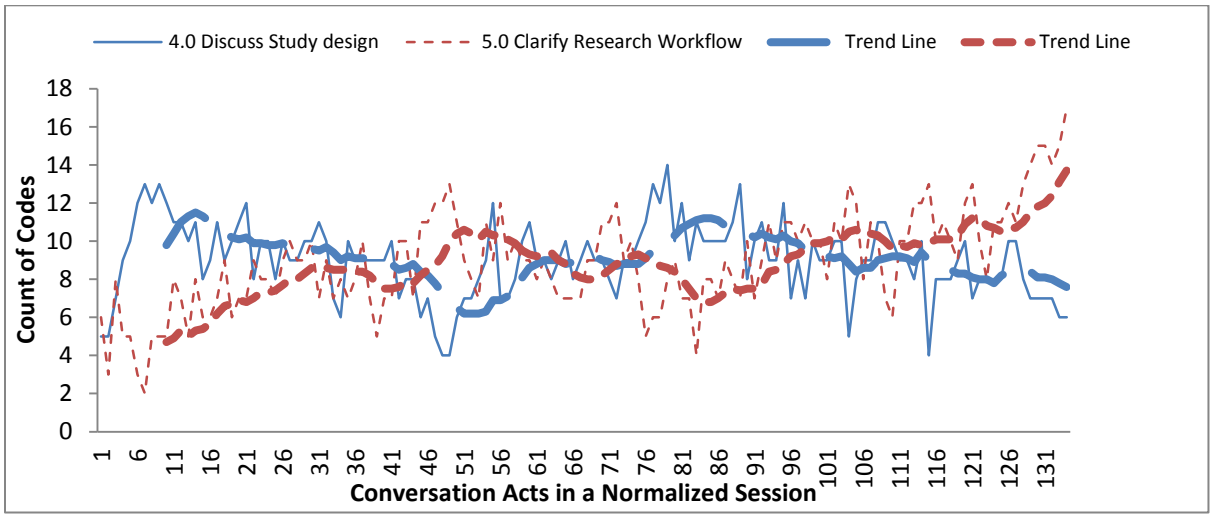

Figure 5. Discussion about Study Design and Workflow Issues throughout a Normalized Conversation

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 those 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 QE and the MR. Zhang et al. briefly describe this process in their "data access paradigm model".[7]

We were able to identify 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 our annotation schema for dialogue acts for mediated queries. Although, we had an inter-rater kappa score of 0.61, we 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 QE and MR.

The initial illustration of the negotiation space (**Figure 3**) is a clear representation of the complexity that exists. Furthermore, we interpret this result as a clear refutation of the idea that data needs assessment is a simple and easy process. A significant amount of QE and MR 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. We interpret the clinical content illustration (**Figure 4**) to represent a clear presentation of potential clinical variables 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 5** 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.

Limitations

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

Conclusion

To the best of our knowledge, this study represents the first attempt to understand the mediated query dialogues between a query expert and a medical researcher. Our 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 what those needs are. Query mediation represents a process-based needs assessment and clarification. The results of this study prepare us for our 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.

Acknowledgements

This research was supported by grants **R01LM009886,** and **5T15LM007079** from the National Library of Medicine, and grant ***UL1 TR000040*** from the National Center for Advancing Translational Sciences. Dr. Cimino is supported by intramural research funds from the NIH Clinical Center and the National Library of Medicine.

References

1. Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. *N Engl J Med.* Aug 5 2010;363(6):501-504.
2. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *J Am Med Inform Assoc.* Sep 6 2012.
3. Holve E, Segal C, Lopez MH, Rein A, Johnson BH. The Electronic Data Methods (EDM) forum for comparative effectiveness research (CER). *Medical care.* Jul 2012;50 Suppl:S7-10.
4. Anderson N, Abend A, Mandel A, et al. Implementation of a deidentified federated data network for population-based cohort discovery. *J Am Med Inform Assoc.* Jun 1 2012;19(e1):e60-e67.
5. Kho AN, Pacheco JA, Peissig PL, et al. Electronic medical records for genetic research: results of the eMERGE consortium. *Science translational medicine.* Apr 20 2011;3(79):79re71.
6. Weber GM, Murphy SN, McMurry AJ, et al. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. *J Am Med Inform Assoc.* Sep-Oct 2009;16(5):624-630.
7. Zhang GQ, Siegler T, Saxman P, et al. VISAGE: A Query Interface for Clinical Research. *AMIA Summits on Translational Science proceedings AMIA Summit on Translational Science.* 2010;2010:76-80.
8. Cimino JJ, Aguirre A, Johnson SB, Peng P. Generic queries for meeting clinical information needs. *Bulletin of the Medical Library Association.* Apr 1993;81(2):195-206.
9. Weber GM, Murphy SN, McMurry AJ, et al. The Shared Health Research Information Network (SHRINE): A Prototype Federated Query Tool for Clinical Data Repositories. *Journal of the American Medical Informatics Association.* 2009;16(5):624-630.
10. Murphy SN, Weber G, Mendis M, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association.* March 1, 2010 2010;17(2):124-130.
11. Murphy S, Mendis M, Hackett K, et al. Architecture of the open-source clinical research chart from Informatics for Integrating Biology and the Bedside. *AMIA Annual Symposium proceedings.* 2007:548-552.
12. Zhang GQ, Siegler T, Saxman P, et al. VISAGE: A Query Interface for Clinical Research. *AMIA Summits on Translational Science proceedings AMIA Summit on Translational Science.* 2010;2010:76-80.
13. Lowe H, Ferris T, Hernandez P, Weber S. STRIDE--An integrated standards-based translational research informatics platform. *AMIA Annu Symp Proc.* 2009:391-395.
14. Wikipedia for Reference Interview, http://en.wikipedia.org/wiki/Reference_interview, accessed on 1/14/2013.
15. Merz RB, Cimino C, Barnett GO, et al. Q & A: a query formulation assistant. *Proceedings / the ... Annual Symposium on Computer Application [sic] in Medical Care. Symposium on Computer Applications in Medical Care.* 1992:498-502.
16. Hripcsak G, Allen B, Cimino JJ, Lee R. Access to data: comparing AccessMed with Query by Review. *J Am Med Inform Assoc.* Jul-Aug 1996;3(4):288-299.
17. Hruby GW W, A, Weng C. Analysis of Query Negotiation between a Researcher and a Query Expert. Paper presented at: AMIA2012; Chicago.