MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190457

Dialogue Analysis for Clinical Data Query Mediation

Chunhua Weng[a], Amy K. Mir[a,b], David Hanauer[c], James Cimino[a,d]

[a] Columbia University, New York, NY, USA
[b] Harvard College, Cambridge, MA, USA
[c] University of Michigan, Ann Arbor, MI, USA
[d] University of Alabama, Birmingham, AL, USA

Abstract

Efficient access to clinical data by investigators is critical for data-driven biomedical research. Mediated query is commonly adopted to facilitate data access for investigators. However, the query mediation process remains a black-box. This study analyzed the email-based dialogues between investigators and clinical data query mediators from three institutions. We identify discussion topics and their frequencies, model task flows, and analyze user needs for query mediation support revealed from the dialogues. While the datasets between different institutions are distinct in some notable respects, we find that together they provide common insights for streamlining data access. From our findings, we conclude an intelligent dialogue-based query support model is feasible to automate human-mediated clinical data access for investigators and stakeholders.

Keywords:

Information storage and retrieval; Interdisciplinary communication; Needs assessment.

Introduction

Big data in healthcare offers immense research potential. In practice, much of the data access is mediated by query analysts who have the adequate technical knowledge, e.g. structured query languate (SQL), to pull information from complex databases. Clinical data access for biomedical researchers is limited not only by patient privacy and data governance but also ineffective communication between clinical researchers and query analysts. In addition to logistic barriers such as lengthy institutional paperwork for approving data requests, there are interdisciplinary communication barriers due to vocabulary differences, conceptualization discrepancies, and knowledge gaps impeding efficient delivery of accurate data sets. With proliferating data, human-mediated data queries are difficult to scale.

The underlying biomedical query mediation (BQM) process is an iterative question-answering process between investigators and query analysts. BQM centers on the investigators' definition of a research statement followed by more precise specifications of the clinical process in question so that query analysts are able to locate the appropriate data elements. In the process of BQM, contextual data constraints often guide revision of the researchers' data queries [1]. The more complex or granular databases are, the more complex BQM may inherently be due to iterative refinements and understanding of the precise information needed from both parties.

Currently, there is minimal literature detailing BQM and much of the underlying communication space exists in a "black-box" in which the actual needs of researchers and the methodology of query analysts' work are not always transparent to each other, which is reflected by the iterative nature of BQM.

Increasing the transparency of the BQM communication space is thus an important first step into understanding how to better streamline data access, as understanding the processes underlying information retrieval can lead to informed redesigns of communication and information flow [2].The biomedical query negotiation can generally be seen as a continuum of automated information retrieval and human-centered communication (e.g. conversations or email exchanges). Depending on the information task, automation can be substituted for in-person communication [3], which is by direct conversation or email, is not always efficient.

In this paper, we investigate email communications during BQM since email exchanges are an excellent candidate for a computational decision-support paradigm in which elements of information flow can be automated. In our analysis of the email communications from three institutions, we attempt to understand how researchers and query analysts negotiate data needs via email and identify knowledge gaps between biomedical researchers and query analysts. We use insights from task and temporal communication patterns, categorization of questions, sender-recipient networks, and the results of predictive text models to better understand email-based BQM in evaluating whether we can feasibly design an automated dialogue-based query support system.

Data and Methods

Email communication for 20 self-reported critical incidents for BQM totaling to 307 emails at an average of 15 emails per case with a standard deviation of 19 were collected from 3 query analysts from three institutions. These cases were exchanges characterized by logistical inefficiencies and communication gaps. Most of the variability in case size comes from Site 3, which also makes up 80% of the email messages. Cases from Sites 1 and 2 average 5 to 6 emails per case with a standard deviation of 2 emails. Cases from Site 3 average to 23 emails per case with a standard deviation of 22. Case 3's variability comes from an outlier case containing 93 emails (Figure 1).

Data De-identification and Structuring

De-identification of sensitive information was done manually by replacing sender names, including titles (e.g. Dr., MD), and associated email addresses with generic codes. In addition to signatures, institutional footers were removed from messages.

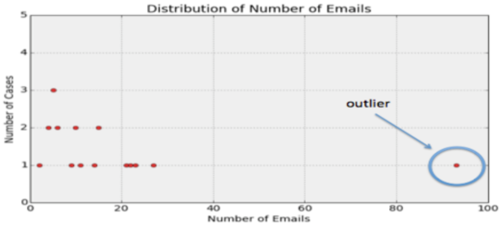
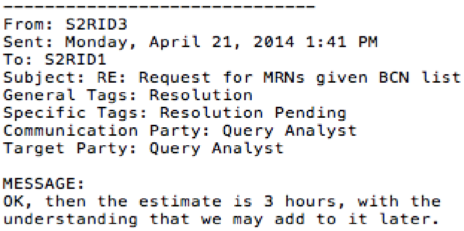


Figure 1 - Distribution of cases by number of emails

Telephone numbers, addresses and locations, passwords, and file names of attached documents were deleted from the text. To ensure further data privacy, the delivery context of emails, e.g. 'Sent from my iphone', were removed from all emails. The structure of all emails was standardized to allow easier parsing of the text. Junk delimiters, excluding message separators, such as > and Original Message identifiers were cleaned from the dataset. Non-alphanumeric text was deleted unless it was related to the data or query commands in an email. Any Date delimiters were changed to Sent. Times were truncated from hour-minute-second format to hour-minute format to ensure uniform precision in temporal analysis of emails. Completed recipient information could not be imputed in all emails in cases where copied (cc'd) recipients were not explicitly denoted in abbreviated email headers. Actual email text was demarcated with the MESSAGE tag and additional tags were added to identify task domains and communication parties.



```
------------------------------
From: S2RID3
Sent: Monday, April 21, 2014 1:41 PM
To: S2RID1
Subject: RE: Request for MRNs given BCN list
General Tags: Resolution
Specific Tags: Resolution Pending
Communication Party: Query Analyst
Target Party: Query Analyst

MESSAGE:
OK, then the estimate is 3 hours, with the
understanding that we may add to it later.
```

Figure 2 - An example de-identified email message

Coding Book for Hierarchical Task Analysis

We defined email tasks, i.e. email events, as what the sender wants to accomplish with the primary recipient. All emails were manually tagged with domains and sub-domains using this coding book (Table 1).

We append an additional identifier Communication Party, an indicator variable of whether a sender is a query analyst retrieving data or biomedical researcher requesting data. Because a sender does not necessarily always communicate with the opposite party, an additional tag Target Party was created to account for researcher-to-researcher and analyst-to-analyst communications in addition to the communications between biomedical researchers and query analysts.

We ran basic email counts across sites, computing min, max median, mode, and standard deviation in the number of emails in a case to get a basic sense of communication volume. We constructed a basic box-plot view of email counts across sites. For each critical incident, we looked at which communication party (researcher, query analyst, or other/unknown) was initiating the thread of the critical incident, as a percentage of the number of cases in each site. We determined the number of emails sent by different communication parties as a percentage of the total number of emails in each site. We additionally determined the number of party-to-party (i.e. researcher to query analyst, query analyst to researcher, researcher to researcher, and query analyst to query analyst) communication flow as a percentage of the total number of emails across sites. We furthermore broke down task domains as well as task subdomains by researcher and by query analyst across sites.

Table 1 - Domains and Sub-domains for Task Analysis

Domain	Subdomain	Definition
Inquiries		
	Task Request	the recipient is asked to assist with a data request by sending information, modifying the structure of the data, supplying additional information and detail, retrying a solution with new information, or confirming feasibility
	Clarification	the recipient is asked to clarify detail on information sent or confirm and make sure of whatever information was previously provided
	Follow-up	a check-in for updates on progress from recipient
Meetings		
	scheduling	Schedule a meeting
	Meeting confirmed	Confirm the schedule
	Meeting completed	Close the conversation
Resolution		
	Resolution Pending	a sender states that progress will be made or is currently being made. This can include mention of failed attempts at a solution or data request acknowledgements by a data team processing the request
	Resolution Offered	the sender offers at least one part of a larger solution or notes which information is already readily available
Other		
	Notice	a general information email, generated without regard to a particular data request
	Forward	an email in which the purpose of the message is to forward a communication
	Introduction	An email in which a new researcher or query analyst introduces himself or is being introduced into the email thread
	Acknowledgment	an email in which the sole purpose of the message is to acknowledge the resolution offered or the progress being made

The interrogative pronouns *'who', 'what', 'which', the adverbs 'where', 'when'*, and the conjunctions 'if', 'whether' characterize straightforward questions can be automated by a database. The interrogative adverbs 'how', 'why' may be more open-ended and less likely to be candidates for the computational paradigm in communication. Better understanding of these grammatical identifiers helps us better understand the nature of questions in email-based BQM. In a representation based on these grammatical terms, we categorize our questions as Identification questions involving 'who', 'what', 'where', 'when', 'which' inquiries, Choice questions involving queries of 'whether', 'if', 'or', Quantification questions involving 'how many', 'how much', 'how long', 'how far', 'how often', 'how high', 'how long', and Discussion-oriented inquiries represented by 'why' and other forms of 'how' like 'how do' and 'how would'. We filtered on emails whose task domains include inquiries and whose specific sub-domains exclude stand-alone follow-ups. In the filtering, we were thus able to focus more on questions related to the data request at hand. Our results are represented by researcher and by query analyst.

The bag-of-words model is popular for document classification. Given that emails are labeled with the task domain, task sub-domain, communication party, and target party, we approach task and party identification as a supervised learning problem. Representing email messages as multi-sets of words that only take into account word multiplicity but disregard word order and grammatical structure, word frequencies in the multi-sets are used as a feature in training a multinomial naive Bayes classifier. Multinomial naive Bayes assumes independence between features for a multinomial event model, which are our task and party classifications.

In task identification, we first predicted general domains. Because multiple domains can occur together in an email, we separately ran the model for each general domain and predicted the binary indicator of that particular domain. We then performed a nested prediction, in which we predict specific task sub-domains conditional on the corresponding general domain, extracting only emails for which that domain is the only domain tag. We applied the same algorithm in predicting communication and target parties. For each prediction set, we performed randomized cross-validation in 1000 iterations and noted the range of accuracies with the middle 800 values, i.e. accuracies at the 10th and 90th percentiles.

We developed a simpler classification algorithm premised on the concept of a codebook; that is, a set of key terms and phrases representative of an aspect of the conversation such an email task domain. Unlike bag-of-words, our codebook-based classification gives consideration to word order since it extracts one to four-word phrases in common among at least two messages of a certain category. For domains, we looked first at matches between single-domain messages and then matches of those single-domain messages with multi-domain messages. For sub-domains, we again conditioned on the relevant broader domain first, as in our bag-of-words-based prediction. For domains, sub-domains, and communication parties, we ensured that the codebooks constructed do not overlap between their different classifications so that the set of codes is unique to that particular classification.

The construction of codebooks allows us to understand the language surrounding an email task or communication party. We hence made another probe into the language of biomedical researchers and query analysts using part-of-speech (POS) tagging to determine the most frequent nouns and verbs, and their overlap, among different communication parties. We adapted a script for noun-phrase extraction. Frequent topics are the functional foundation of automated dialogue-based query support and we hence computed the frequencies of our extracted topics across sites.

We examined how tasks follow from one another across emails. We computed a ranked moving average of the number of task sub-domains across emails. We looked at the intra-message and inter-message relationship between tasks, respectively, in computing the most frequents tasks concurring with a specific task and the most frequent tasks preceding or following that specific task. We constructed network visualizations specifically for clarifications and task requests to better understand the events surrounding the iterative question and answer process. For cases without any missing timestamps we visualized the time lengths between conversations in a box-and-whisker plot and analyzed task domains and sub-domain at peaks, i.e lags, in response times. To understand how involved the communication networks are, we plotted a color-coded network visualization of senders, primary recipients, and secondary (cc'd) recipients who are neither senders nor primary recipients. We then generalized the communication interactions to four different network models.

Results

Communication Volume and Communication Networks

In the breakdown of thread origination across sites, we look at which communication party, i.e. researchers or query analysts, are initiating the conversation. While most cases begin with contact by a query analyst with institutions 2 (100%) and 3 (55%), with institution 2 being entirely analyst-initiated, institution 1 is predominantly researcher-initiated (Figure 3a). The researcher to query analyst communication volume among individual email messages is about the same as the query analyst to researcher volume for Sites 1 (40%) and 2 (38%), and is slightly higher with Site 3 (45%) (see Figure 3b).

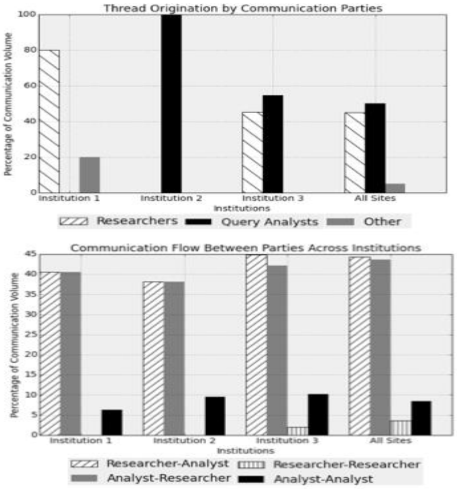


Figure 3 - Breakdown of pairwise communication

Interestingly, exchanges are not simply between opposite parties. Across all three institutions a non-negligible portion of the communication volume occurs between query analysts themselves, and for Site 3, there are also some smaller number of exchanges from researcher to researcher (see Figure 3b).

Communication networks represent the complexity of conversations between senders and primary recipients for an entire critical incident. Two-node interactions are the most common and simplest network type in our data, with nine critical incident cases following this prototype. Linear three-node networks are represented in four cases, where the additional third node can be any party. In linear three-node structures, the two relationships could be bidirectional relationships among the same parties and between different parties. Our cases also include the instance of both bidirectional and unidirectional interactions among the two pairs of different parties. Triangular three-node structures, represented by four critical incidents, can have one to three bi-directional graph edges and, in our data, roughly the same number of cases involves the majority of the nodes being of either party. Complex networks are the rarest case, represented by only three cases, all within Site 3. The first three network models all exist in exchanges from each of the three institutions.

Task domains are split almost evenly between researchers and query analysts, with insignificant variations between specific domains and sites (see Figure 4a). In the small set of cases from Site 2, there are no emails requesting, confirming, or referencing meetings. Within the inquiry domain, most questions are straightforward identification or choice questions as opposed to quantification questions and more involved, discussion-oriented inquiries (see Figure 4b).

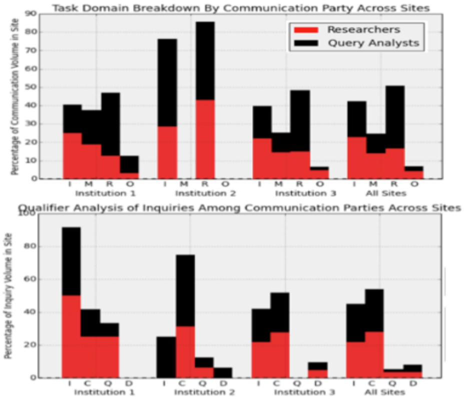


Figure 4 - Breakdown of tasks and classification of inquiries. (a) Task domain breakdown by communication party across all sites. I = Inquires, M = Meetings, R = Resolutions, O = Other. (b) Question breakdown by communication party across all sites. I = Identification, C = Choice Between Options, Q = Quantification, D = Discussion

Predictions of Tasks and Communication Parties

The bag-of-words algorithm with naive bayes assumptions on our multinomial event model for domains and sub-domains has generally good predictive performance. For the middle 80 percent of values within the 10th and 90th percentile of prediction accuracies, we find that single-domain prediction of inquiries, meetings, and resolutions range from 61-74%, 73-84%, and 60-71%, respectively. Because we approached this as binary a prediction for each domain, these results are better than a 50% random prediction.

For the nested prediction in which we condition in the presence of a particular single-domain, we again have better-than-random results. For inquiries, we find that task requests, clarifications, and follow-ups have 80 percent of their accuracies at 50-78%, 50-78%, and 61-83%, respectively. The similar and slightly lower range of prediction accuracies for task requests (i.e. information requests, and clarifications) reflect the similarity in these two sub-domains as they are both inquiries specifically regarding the data at hand, as opposed to questions following up with progress or requesting meetings.

In the conditional predictions with single-domain meetings, we find that meeting scheduling, meeting confirmations, and completed meetings have most of their accuracies at 62-92%, 62-92%, and 100%, respectively. In conditional prediction with single-domain resolutions, we find that pending resolutions and definitively offered resolutions can be mostly predicted within 65-87% and 61-87% accuracies, respectively. Looking at party predictions, we find the party originating the email message can be predicted mostly within 64-77% accuracy for both biomedical researchers and query analysts. The party to which the message is targeted can be accurately predicted within 58-70% and 57-70% accuracy for researchers and query analysts, respectively. The general domains of inquiries, meetings, and resolutions are accurately predicted with 42%, 75%, and 61% accuracy, respectively. Within inquiries, task requests, clarifications, and follow-ups are correctly predicted 20%, 43%, and 85% of the time. Within meetings, meeting scheduling, meeting confirmations, and completed meetings are respectively determined with 54%, 95%, and 96% accuracies. Within resolutions, pending solutions and definitively offered solutions are respectively determined with 33% and 38% accuracy. Communication parties are determined with 50% accuracy.

Topic Extraction

The codebook approach provides a basis for assessing the common language used in dialogue-based query mediation. Looking at the nouns and verbs, the basic elements of topics, we see some overlap among researchers and query analysts and because of the different sizes of word sets, we measure overlap with respect to the smaller word set. Looking specifically at noun topics, we find that between researchers and analysts, there are 43%, 39%, and 58% overlaps across Sites 1, 2, and 3 respectively. Between researchers these overlaps are 15%, 41%, 39%, and between query analysts these are 29%, 34%, 37%. The smaller overlaps within the same parties make sense as each critical incident represents a distinct data request. The more significant overlaps in noun topics between opposite parties is reasonable since a common language is needed to mediate information requests. Looking at the most frequent topics in our subsequent noun-phrase extraction, we see that important common topics across all cases include logistic terms such "MRN," "codes," "data," "warehouse," and "ICD9" as well as more medically-focused concepts such as "adenovirus," "hemoglobin," "ulcer," and "creatinine." Common to all sites are terms such as "lab," "dates," "IRB," "gender," and "diagnosis." Common to Sites 1 and 2 are terms such as "admissions" and "months." Sites 1 and 3 share terms like "control," "age," "medication," "reaction," "demographics," "ICD9," and "MRI." For Sites 2 and 3 we see common terms such as "chronic" and "surgery." These frequent topics serve as the core of dialogue-based query mediation.

Task Flow and Temporal Analysis

In determining the most common tasks preceding, concurring with, and following particular task sub-domains across sites, we found that the email tasks surrounding clarifications and task requests, i.e. information requests, are quite similar across sites. Within the ongoing process of BQM, both task request, i.e. information request, and clarification events tend to be preceded by clarifications, task requests, and definitively offered resolutions, and they both tend to be followed by clarifications and definitively offered solutions. Moreover, these two email events tend to coincide with a clarification or task request alongside an offered resolution.

Whereas meetings will attempt to be scheduled following a clarification, these task events tend to coincide with or precede a task request. Completed meetings also coincide with task requests. Task requests can also commonly occur in response to forwarded messages, whereas clarifications can occur in response to follow-ups suggesting that clarifications intuitively highlight bottlenecks in the information retrieval process.

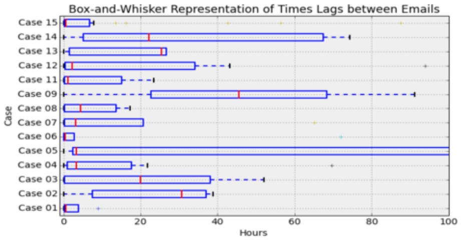


Figure 5 - Distribution of times between emails within a critical incident. Box-plot representations with median and quartiles shown for cases with complete time data. Case 5 has significant variation in email response times with its box-plot extending outside the timescale range used in this comparison chart

There is significant variation in the time durations of our critical incidents (see Figure 5). Analyzing the temporal lags, we find that most frequently preceding time peaks, query analysts have completed a meeting, have offered a resolution, and are currently working on another solution. At the opposite end of the time lag researchers typically contact query analysts with information requests. During the course of exchange, there seems to be no regular pattern among institutions in how the number of distinct task sub-domains change, whether they are generally increasing, decreasing, or peaking midway, emphasizing BQM to be a non-linear process.

Discussion

Communications between query analysts in some of these critical incidents suggests that database querying is not a straightforward process that can always be directly handled by a single mediator. The somewhat smaller proportion of researcher-to-researcher communications suggests that an understanding of the clinical process can also be quite involved. However, these represent a minor proportion of the communication flow and for the most part, BQM is an interaction between opposite parties, mostly having simple communication network structures. That task domains are roughly broken down equally between parties suggests that there is as much of a need and refinement both for the specifications of the clinical process and for the data obtained.

This study complements earlier work in BQM focusing on in-person communication. In-person conversations that comprise non-electronic BQM are an important part of the information retrieval process [4]. Indeed, in 14 out 20 threads that set up meetings, 7 of those threads reference information from that conversation. However, while meetings are important, they are not critically necessary to BQM. Not only are meetings the smallest volume of task domains among our three major task domains, Site 2 does not involve meetings at all and six critical incidents across the entire dataset altogether exclude this task domain. This strongly suggests that BQM, or at least parts of it, can be taken to a computational model.

In fact, there is notable structure and predictability in BQM. On one hand, the volume of distinct events over the course of an exchange does not seem to follow any consistent patterns among our sites. On the other hand, our network analysis of task events surrounding information requests and clarifications confirms the iterative question-and-answer model of BQM discussed in earlier research. Our topic analysis is integral to understanding this iterative nature. Our findings of both logistically-oriented terms such as "codes," and "ICD9," in addition to the medical jargon, can guide the construction of an automated dialogue-based query support, as these codebooks can inform references within automated dialogue-based support that model the actual conversational support found in emails.

Intelligent query formulation relies on the ability to automatically understand the flow of a conversation and we have observed that much of the mediation process can be predicted at a high level. The content of email conversations moreover suggests that BQM is an ideal candidate for the computational design-support paradigm as many of the questions are straightforward inquiries, e.g. questions regarding basic identifying information and questions regarding choices between alternative sets of information. The relevant information for such inquiries can be autonomously accessed by researchers given the appropriate interface support [5; 6].

This study is limited by the small sample of emails from a small number of institutions for 20 critical incidents, thus not providing a complete representation of the entire BQM space. Nonetheless, we offer important initial insights into the dynamics of human-mediated query support, confirming earlier research in BQM as well as providing promising insights into the potential for automating BQM. Our codebook-based algorithm extracted a unique set of codes for a particular domain/sub-domain classification, codes common to two different domains are not used in the determining the classification of a message. This suggests that many domains may occur together, as we see in later analysis.

Conclusions

This study contributes an analysis method for understanding interdisciplinary communicaiton and early insights into the black-box biomedical query process. We conclude that Dialogue-based decision support is needed and feasible.

Acknowledgements

This publication was supported by Grants UL1TR001873 and R01LM009886 from National Institutes of Health of USA.

References

[1] G.W. Hruby, J.J. Cimino, V. Patel, and C. Weng, Toward a cognitive task analysis for biomedical query mediation, *AMIA Jt Summits Transl Sci Proc* **2014** (2014), 218-222.

[2] J.H. Gennari, C. Weng, J. Benedetti, and D.W. McDonald, Asynchronous communication among clinical researchers: a study for systems design, *Int J Med Inform* **74** (2005), 797-807.

[3] E. Coiera, When Conversation Is Better Than Computation, *Journal of the American Medical Informatics Association* **7** (2000), 277-286.

[4] G.W. Hruby, M.R. Boland, J.J. Cimino, J. Gao, A.B. Wilcox, J. Hirschberg, and C. Weng, Characterization of the biomedical query mediation process, *AMIA Jt Summits Transl Sci Proc* **2013** (2013), 89-93.

[5] H. Lieberman and C. Mason, Intelligent agent software for medicine, *Stud Health Technol Inform* **80** (2002), 99-109.

[6] L. Laranjo, A.G. Dunn, H.L. Tong, A.B. Kocaballi, J. Chen, R. Bashir, D. Surian, B. Gallego, F. Magrabi, A.Y.S. Lau, and E. Coiera, Conversational agents in healthcare: a systematic review, *J Am Med Inform Assoc* **25** (2018), 1248-1258.

Address for correspondence

Chunhua Weng, Department of Biomedical Informatics, Columbia University, 622 W 168th Street, PH-20 room 407, New York, NY 10032, USA. E-mail: chunhua@columbia.edu .