Characterizing Data Discovery and End-User Computing Needs in Clinical Translational Science

Parmit K. Chilana, University of Washington, USA

Elishema Fishman, University of Washington, USA

Estella M. Geraghty, University of California, Davis, USA

Peter Tarczy-Hornoch, University of Washington, USA

Fredric M. Wolf, University of Washington, USA

Nick R. Anderson, University of Washington, USA

ABSTRACT

In this paper, the authors present the results of a qualitative case-study seeking to characterize data discovery needs and barriers of principal investigators and research support staff in clinical translational science. Several implications for designing and implementing translational research systems have emerged through the authors' analysis. The results also illustrate the benefits of forming early partnerships with scientists to better understand their workflow processes and end-user computing practices in accessing data for research. The authors use this user-centered, iterative development approach to guide the implementation and extension of i2b2, a system they have adapted to support cross-institutional aggregate anonymized clinical data querying. With ongoing evaluation, the goal is to maximize the utility and extension of this system and develop an interface that appropriately fits the swiftly evolving needs of clinical translational scientists.

Keywords: *Biomedical Research, Clinical Data Discovery, Clinical Translational Science, End-User Scientific Computing, Federated Querying, Patient Information Systems, User Needs*

INTRODUCTION

Rapid advances in information technology are opening up new avenues for conducting research in biomedicine. The application of new technologies has enabled greater ability to generate, capture and analyze biological data for basic research, but there remain significant challenges in integrating and translating these data with clinical data into forms that can be used to improve clinical outcomes. Clinical translational science is an emerging interdisciplinary field that seeks to facilitate the translation of biomedical research advances from the laboratory to improve clinical and public health outcomes and vice-versa (Zerhouni, 2007).

DOI: 10.4018/joeuc.2011100102

Institutions that are part of the 46 member Clinical Translational Science Award (CTSA) consortium, sponsored by the National Institutes of Health (NIH) are beginning to develop, implement and support research on improving human health, both in local environments as well as collaboratively across sites. Individual CTSA sites are composed of researchers with expertise across the clinical research workflow, including biostatistics, informatics, bioethics, as well as community outreach and clinical translational research scientists. Within this novel consortia, there are challenges to consider in developing and incorporating information tools and methods that can catalyze and advance research both into as well as across the rapidly evolving and increasingly heterogeneous clinical research environments.

To better characterize the data discovery needs and end-user computing practices of clinical translational scientists and to steer the development of a novel cross-institutional clinical data discovery project, we conducted a pilot study involving semi-structured interviews with twelve principal investigators and research support staff working on a range of clinical research projects within the University of Washington's Medical Center, and affiliated with the UW CTSA site, the Institute of Translational Health Sciences (ITHS). We adopted the widely-used techniques of user and task analysis (Hackos & Redish, 1999) and focused on understanding workflows in translational research, data gathering methods, and previous experience in developing or customizing data discovery tools.

Our findings shed light on the diversity of user needs and expertise within clinical translational science and the potential barriers scientists face in accessing clinical data and using existing querying systems. The diversity exists not only in terms of the technical prowess of the scientists, but also in the range of their research questions, which still typically rely on and require customized data discovery and analysis features for domain specific work. The implications of these results are useful for supporting and enhancing our on-going project implementation as well as to other clinical data-discovery systems that seek to integrate and catalyze collaboration across complex and heterogeneous data domains.

At a higher level, we have found it critical to understand the context of use and the querying practices of clinical translational scientists early in the development process of our system, as the resulting computational tools become necessarily integrated components of the research enterprise, though in novel ways that test existing expectations of the end-user researchers. Developing query access to clinical data systems for research is particularly unique and challenging, as in addition to managing patient data privacy and data security, establishing methods to extract, analyze and compare highly heterogeneous clinical data challenges normative assumptions of how researchers understand and interact with operational clinical health environments. This challenge is enhanced when multiple institutions seek to build collaborative services of this form, and is demanding new roles and expertise to support characterization, deployment and support of these new systems. We plan to continue our partnership with end-user translational scientists by employing an iterative development process to refine use-cases and to better understand user perceptions, end-user computing needs and usability issues to maximize the overall utility of clinical research systems.

BACKGROUND AND MOTIVATION

We have implemented and are extending Informatics for Integrating Biology and the Bedside (*i2b2*), an interoperable open source software architecture that for our project facilitates the discovery of anonymized aggregate clinical data of patients who may meet eligibility criteria for clinical trial recruitment (e.g., counts). This Cross-Institutional Clinical Research Project (CICTR) effort is a generalization of the Harvard implementation of i2b2 and when fully deployed will provide query-level access to de-identified clinical data through a web interface across three different institutions: University of Washington (UW), University of California, Davis (UC Davis) and University of California, San Francisco (UCSF), with an estimated patient population in excess of 4 million individuals. By constructing a common technical interface and knowledge representation across these three sites, researchers will have the ability to create simultaneous federated queries to discover if sufficient study subjects for clinical trials are available either locally or more broadly within the consortium. The tool will further allow translational scientists to create queries to explore retrospective clinical data characteristics and modify and reuse query criteria to reflect ongoing research questions.

Although technical, semantic and governance issues are major challenges in building this collaborative information exchange environment, ensuring the overall usability and utility of the system to end-user researchers is crucial. Since clinical translational science is a newly emerging field, little prior work has specifically focused on the needs of scientists or software development within this domain.

Informatics and human-computer interaction (HCI) studies related to biomedical research have largely focused on biologists in lab settings. Some have focused on understanding information tasks and workflows (MacMullen & Denn, 2005; Tran et al., 2004; Bartlett & Toms, 2005), while others have investigated larger socio-technical issues of biomedical research environments (Anderson et al., 2007; Ash et al., 2008). Other works have focused on facilitating end-user programming needs in bioinformatics (Massar et al., 2005; Letondal, 2005) and understanding software development practices (Umarji & Seaman, 2008; Chilana et al., 2009). Such studies have not considered the issues that scientists face in clinical translational science where the focus is on patient data rather than molecular data.

On the other hand, studies related to patient data have largely revolved around clinicians and use of patient information systems. For example, some studies have outlined barriers and solutions in the use of electronic medical records (EMRs) (Miller & Sim, 2004), while others have focused specifically on usability issues (Rose et al., 2005). Others have focused on understanding errors in patient information systems (Ash et al., 2004). Research on doing formal design, verification and prototyping of patient information systems has also been explored (Mathe et al., 2007). Although these works are clearly important for clinical practice, they have not looked at how scientists query and interpret patient data for secondary research purposes, which has been the primary goal of our study.

The insights gained from our study are useful not only for developing and extending *i2b2*, but also serve as a case study of data discovery and end-user computing needs in clinical translational science. In this regard, this paper also complements other case studies of scientific software development (Carver et al., 2007; Segal, 2005, 2009) but sheds new light on needs and barriers that are unique to clinical translational science.

STUDY DESIGN

We used a semi-structured interview technique for informant data collection, since it allowed for an open-ended discussion that could capture the situational aspects of data use and provide us an approach to finding consistencies among responses (Strauss & Corbin, 1998). We developed a list of structured interview questions, focusing on user and task analysis techniques (Hackos & Redish, 1999) to understand the work environments and tasks of participants. We used this instrument to initiate the interview and probed into interesting responses by asking unstructured follow-up questions. Each interview lasted approximately one hour. Where possible we interviewed participants in their work settings to establish context and facilitate recall.

Our participants were identified through word-of-mouth, email and snowball sampling (where our current participants helped identify

Table 1. Profile of interview participants

	Role	Research Area
P01	Clinical Data Support	Basic/bench science & retrospective clinical
P02	Research Scientist	Clinical trials, health quality
P03	Clinical Data Support	Various
P04	Clinical Data Support	Various
P05	Principal Investigator	Pre- and Post-tests, drug effects, healthcare guidelines
P06	Research Scientist/ Clinical Data Support	Health services, quality metrics
P07	Principal Investigator	Health services, clinical trials
P08	Principal Investigator	Clinical trials, cardiovascular health
P09	Principal Investigator	Epidemiology, observational studies
P010	Principal Investigator	Clinical Trials
P011	Clinical Data Support	Various
P012	Clinical Data Support	Various

other participants). There were 12 participants in total (see summary in Table 1).

The interviews began with a focus on understanding how scientists used clinical and EMR data for research purposes. Based on the critical incident technique, we asked participants to show us the steps they recently carried out in exploring or analyzing such data. Responses included explanations of the tools they used, the types of queries they formulated and descriptions of how they formatted and visualized the output, where applicable. In particular, we were interested in learning about the types of technical and/or non-technical challenges or barriers they faced in using clinical data for secondary research purposes. Lastly, we sought to understand what scientists expected or wished they could do with a federated clinical data querying system that would provide access to anonymized EMR data across a larger pool of institutions.

Two of the authors transcribed the interviews independently. We analyzed the transcripts inductively, by following an iterative process of open coding and axial coding to discover relationships among emerging concepts in our data (Strauss & Corbin, 1998). This was followed by selective coding where all the results of axial coding were integrated. We identified recurring themes in the interviews by following this inductive analysis approach.

RESULTS

Our key findings confirm the existence of different types of users of clinical systems, the varying range in their clinical data discovery and analysis needs, and the barriers scientists face in accessing clinical data for translational research purposes.

Different Needs for Different Users

Our interviewees represented two different user groups based on the tasks they described: (1) researchers who collected and analyzed clinical data on their own and (2) intermediaries or "research IT" (coined by Bernstam et al., 2009) staff members who worked closely with investigators to access and distill complex data requests.

Within the group of researchers, participants had a wide range of technical skills. For example, half of our researcher participants were accustomed to extracting and manipulating clinical data on their own and could write sophisticated SQL queries. One researcher explained that technical skills were necessary to have control over the data:

I would want to do my own analysis – totally depends on what the question is...I want more control over the data, I would want to do subsets...my experience with other reporting software is that they are very limited...it's just a simple report, you can't do back-end processing...

In contrast, one researcher confessed to having very limited experience using any kind of EMR system or other tools. Most other researchers fell somewhere in-between—they had basic knowledge about database design and how EMR systems function in general but could not extract or manipulate data on their own. They consulted with research IT personnel as needed.

Researchers also pointed out that sometimes they had no choice but to go through specialized research IT staff because they did not have permission to access certain patient databases. For example, one researcher described a situation where she needed patient data housed in the hospital billing database:

...they [billing dept] don't have people sitting around willing to help us...there [are] long delays to when we make a take request to when we actually get data back...have to be considerate of their work load... it would be good to get things that may be harder for them but valuable for us....

Research IT participants typically had expertise in database design and programming, but little or no formal training in biomedicine. They acquired the relevant domain knowledge through on the job experience. The type of help they provided to researchers ranged from basic data transformations to complex queries involving multiple tables and joins from disparate data sources that could sometimes take days to run. For instance, one research IT participant described the intricacies of providing the requested data to a group of clinical researchers as follows:

A lot of what I do is pulling raw data and then putting[it] into tables and then handing off tables to researchers. I don't have statistical background. For some researchers, sometimes they want us to do basic analytics (like counts) – but a lot of the researchers want raw data so they can do high-level analysis.

He further explained that researchers were sometimes not aware of the level of detail required in formulating queries and data transformations to seemingly "simple" research questions. As a result, often there were unrealistic expectations about timelines.

Another research IT participant pointed out that working with researchers required an ongoing partnership approach as the researchers did not always know what to do with raw data:

...they [researchers] had the notion of pulling the data in batch, but once they see the data in tabular form, they have no idea what to do with it...they don't want data, they want the question answered...it's an iterative process...the data has to be cleaned or managed.

Thus, we saw that not only researchers and research IT had different needs and experiences in accessing and making use of clinical data repositories, but that there existed individual differences even within these subgroups. As an example, principal investigators often possessed domain-specific statistical and study design skills – or could gain assistance with this – but were not necessarily those who were directly involved in resolving this experience with the practice of directly accessing the patient data. It is clear that developing high-utility access to clinical data for research purposes will require tools and expertise that cover both the technical skills as well as the ability to transform formal research questions into structured queries that meet necessary study design criteria.

Table 2. Examples of secondary use of patient data

- identifying target patient cohort(s) for clinical trials
- observational studies of drug benefits and adverse effects
- clinical quality improvement for patients (e.g., for diabetes)
- monitoring errors (e.g., in the emergency room)
- analyzing rates of use of medication to prevent certain conditions (e.g., blood clots)
- monitoring the type of lab test requests

Range of Clinical Data Discovery and Analysis Needs

Another theme that emerged was the variability and range in participants' responses on the use of clinical data for secondary research purposes.

There was general agreement among our participants about the utility of certain types of clinical data for pursuing translational research questions: patient demographics, laboratory values, diagnosis data and medication data. However, participants also described instances where they had more specific data needs for answering larger or more complex questions and there was wide variability within each research domain (i.e., clinical science, epidemiology, quality of care). For example, participants listed sources such as:

- Problem lists
- Nursing notes
- Therapeutic data
- More detailed orders
- Diagnostic test data
- Socio-economic data
- Procedure data (i.e., surgical procedures)

For some researchers, the use of data was more exploratory in that they wanted to find interesting patterns in the data to formulate relevant research questions. For other researchers, it was more important to have access to defined data points to answer various domain-specific research questions.

For instance, a common use of clinical data was for pre- and post-analysis of various groups of patients with a specific condition, such as blood clots. Another use was to determine if an intervention had the expected outcome for which it was designed. Some participants were also engaged in research involving quality of care and were interested in data, such as patient admission and re-admission rates, in a given timeframe. Table 2 summarizes the types of translational research questions researchers were pursuing by making secondary use of patient data.

Barriers to Accessing Clinical Data

Our findings also revealed a number of common barriers to accessing and using clinical data for translational research. These were from the perspective of both researchers and research IT staff. Some of the barriers were related to general medical practices and conventions, while others were more specific to experiences using commercial, open-source and/or locally developed software tools housing clinical data.

Accessing data in disparate sources: Participants agreed that one of the biggest challenges in doing translational research was obtaining access to patient data housed in different systems and formats suitable for secondary research.

For example, one researcher described that the challenge was in dealing with additional cost and time overhead for gathering clinical data from disconnected sources:

We mostly [had] registries of patients – type in their number and go look at their record... nursing notes were more detailed on what actually happened, but that's a separate database and had to be integrated - very costly and time consuming...about an hour per patient and we had a couple of thousand patients...

Another researcher explained that gathering data for certain retrospective studies could be challenging—sometimes requiring 10 year old records that were available only through paper-based records or microfilms. Having access to recent EMR data was crucial, but not always sufficient:

You may have 10 variables you want to satisfy [that] you don't have for whatever reason...you may not have all 10 variables in a database, but you may have in your chart..need photocopy of the medical record,...[through the] hardcopy, you may have [access to] some things you may not know about [in the EMR]...

Thus, the lack of data integration, different formats, and inefficiency in accessing non-EMR data was a major barrier in pursuing many types of translational research questions in a timely manner.

Naming and classification conventions: Participants also discussed the difficulties of working with EMR data because of inconsistent naming conventions and non-intuitive classification schemes.

For example, one research IT participant described problems he faced when merging data based on attributes that were defined inconsistently between systems:

A data dictionary is very important/necessary with this system, but also there's a need for deep semantics of the data itself. For instance, deceased is a data attribute in this system, but it only applies if the patient dies at [this hospital], and if the patient dies somewhere else it's not marked/included in the software accurately... also not easy to interpret.

Although proper supporting documentation could be used to prevent such misinterpretations, this documentation was often incomplete or unavailable:

..not enough documentation of the data...it's hard to find the person who knows what a particular data field is there or why it's used ... list of codes don't mean anything...not an easy way to just look it up...not enough consistent documentation or availability of expertise.

Ongoing work in biomedical ontologies and terminologies will resolve some of these issues with naming conventions in the near future, but our data suggests that these ontologies should be developed with a user-centered approach to minimize misinterpretations by researchers in the long run.

Inaccurate or missing information: All participants also cited examples of difficulty they faced in working with incomplete and inaccurate data in the clinical repositories they queried.

This problem often appeared, due to data entry errors influencing quality and requiring additional work:

There's always missing data, some of the data is wrong...[we know] after checking multiple sources – in one area it's different than another area...it's hard to determine..

The other problem was that due to the exploratory nature of research, scientists could not predict ahead of time or enforce what patient data should be collected or protected over time:

One of the issues with our data systems is that certain tables are at the patient level and certain ones are the visit level – if someone changes their address...then I don't get to see it changed... like was the patient homeless 3 years ago? As the data goes in, it would be helpful [to know]... how [data] looked on a particular day... what did it look like at that time, instead of keeping messy audit trails...

The researcher participants agreed that an improvement in the design and use of EMRs in clinical settings would be helpful to prevent errors during translational research, but such a change would require long-term institutional changes and could be difficult in practice.

Limitation of systems interfaces: When asked to show example queries in existing systems, many of our participants pointed out the other barriers they faced.

For example, one research IT participant explained that although he could create complex queries with the current system he was using, he could not handoff the system to researchers if they wanted more control over their data and analysis:

...functionally [the system] does pretty much what I ever wanted it to...users are limited... can't customize query [as] they need to know how backend of a database works...I don't know anyone [researcher] who would use it... a lot of things are hard-coded in, have to change all the way back to the underlying C++ code to create custom queries..

In addition to lacking intuitive querying and end-user programming facilities that are usually required to create customized queries, current systems altogether lacked the facility for answering certain research questions, as explained by one researcher:

A lot of the data would get better info if we were able to have natural language processing to extract some free text...some of the coded data in EMR is easier to work with and there are also algorithms that can only access coded data. However, for some of the diagnoses, important data resides as free text...our analytic algorithms can't look at it because we don't have any automatic way of looking at it...

In summary, the barriers discussed above show how researchers were sometimes forced to extend timelines or invest additional resources to obtain and transform the appropriate data. These results further stimulate enthusiasm for tools that would allow and simplify the process of using clinical data for secondary research and better streamline clinical research workflows.

DISCUSSION

As CTSAs and other research organizations seek to develop and deploy tools and processes to meet the needs of the clinical translation research environment, access to high quality clinical data remains a key issue. Although there are significant technical, semantic and policy-level issues underlying the development and implementation of clinical research systems, our study shows that for improving the long-term adoption and success of these emerging systems, there is much value to be gained by understanding the variation in needs of end-user scientists in the context of their translational research workflows.

A number of the themes that have emerged in our data have direct implications for supporting clinical data discovery and designing systems and processes that accommodate the varied needs of different users (i.e., researchers versus research IT staff). As suggested by our results, for some researchers who possess aptitude for development-stage end-user computing, completeness of interfaces is less of a barrier and they may be willing to experiment with tools and formulate complex queries on their own. But, for others, the role of the research IT and collaborative research design and data extraction is critical in getting assistance in effectively exploring data. Although our current focus has been on the *i2b2* system within the CICTR project, we believe the emergent themes apply to other similar clinical data querying tools.

Providing Control Over the Expressiveness of Data Queries

In our results we found the existence of two end-user groups (researchers and research IT) and varying levels of technical expertise and clinical domain knowledge. This finding suggests that different levels of control are needed to tailor queries to research questions. Thus, systems should (1) offer a facility to create custom queries and in this, (2) address the balance between providing simple query interfaces for data characterization (e.g., counts and summary data) and advanced interfaces for formulating complex questions (e.g., trends, limited data sets, visualizations) . This is consistent with other findings on EMR-based query construction (Murphy et al., 2003).

Making Sense of Query Results

Our results showed that clinical scientists had different preferences and expectations for processing the results of the data access queries that they carried out themselves or through research IT. Thus, a data discovery system should provide multiple options for export and data delivery (e.g., raw tabular view and visual overview), and support evolving national data representation standards. There may be value in further exploring options for integrating statistical tools and sophisticated visualizations, such as recent work on temporal patterns in numerical and categorical EMR data (Plaisant et al., 2008;Wang et al., 2008).

Improving Usability of Interfaces

A consistent theme in our results was the need to make input and output behavior consistent to encourage reuse and customization, both in view of the researchers and the research IT participants who used different clinical data querying tools. Thus, users should be able to leverage experience from using other data discovery systems where appropriate and not face an arduous learning curve in customizing new interfaces since that takes away time and resources from the main goal of scientific data discovery. Ongoing usability testing and iterative design (Hackos & Redish, 1999) and training approaches could be valuable for creating user-centered clinical data querying systems.

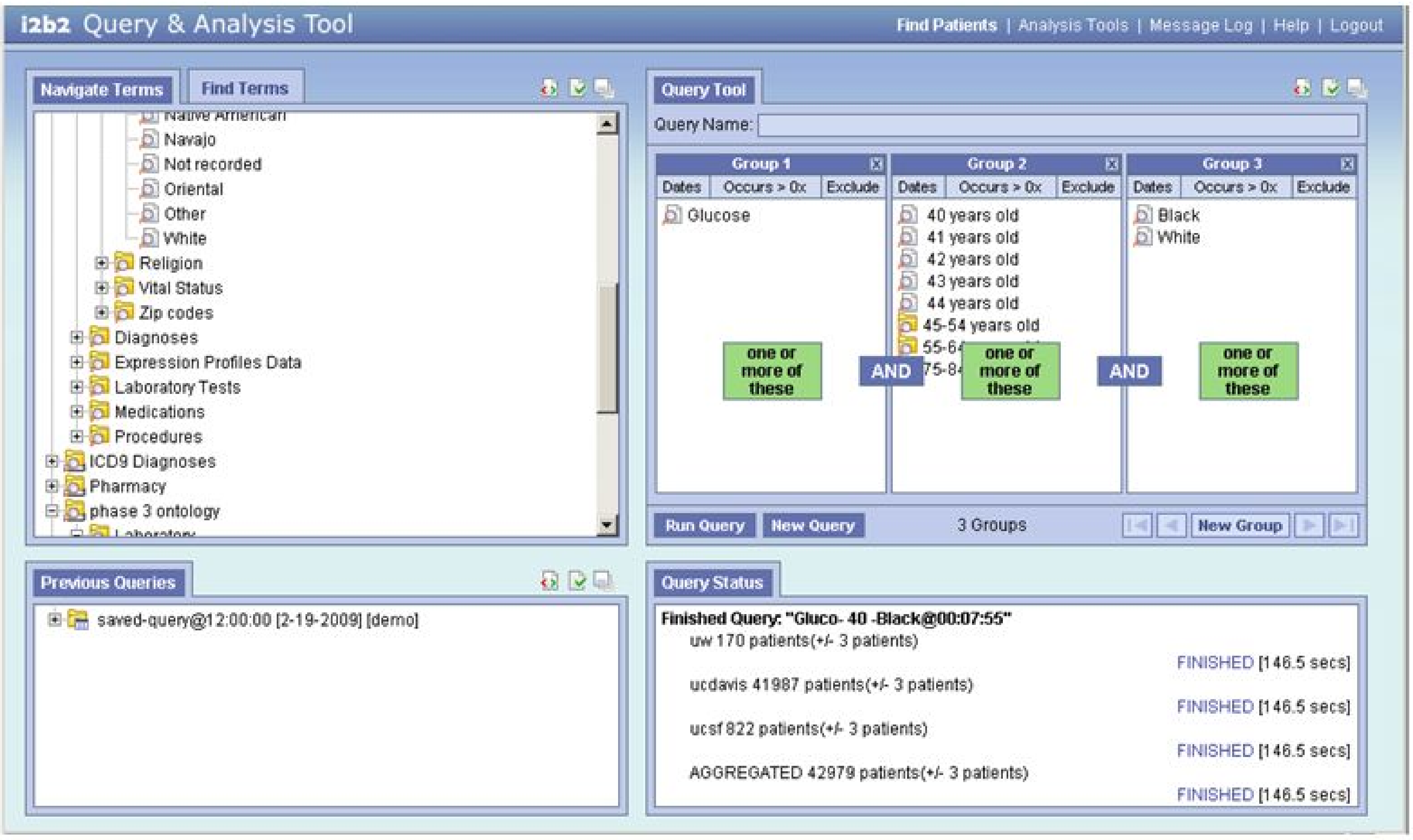
Data Transformation in *i2b2*

To better support end-user needs, the CICTR project has been building rapid iteration data transformation workflows to allow for each of the three university-based i2b2 nodes to develop common end-user data environments. At the heart of these are tools which allow for raw data to be imported easily into i2b2, including a "Universal" Extraction, Transformation and Load (UETL) tool a ontology tool that allows for customizable "mapping" scripts to leverage web-based terminology servers to generate rich end-user query taxonomies in the existing i2b2 interface. These tools have allowed the researchers and evaluators to separate the mechanics of Extraction, Transformation and Load (ETL) and terminology alignment from the end-user domain-focused experience. As a result, the current i2b2 workbench (Figure 1) remains identical for all end-users within CICTR, though the abilities and data representations contained within the interface are becoming more flexible as the project engages with different domain experts.

Managing Cultural and Social-Technical Issues

Apart from the implications for design, our interviews shed new light on the diversity of data discovery needs affecting translational research and the barriers that still need to be resolved beyond the system level. We believe the key to successful development and integration of translational research systems rests on how well we address the higher level barriers. Our results show that cultural and social-technical issues pose major challenges to clinical translational scientists in obtaining the appropriate data. For example, the barrier of accessing data in disparate sources involves knowledge of who has control and who can grant access or securely "hand off" data between clinical and research environments. Provenance issues, due to data-entry mistakes or data corruption errors, can limit the types of research questions that can be credibly pursued as well as the utility

Figure 1. Screen shot from i2b2 web client within CICTR



Finished Query: "Gluco- 40 -Black@00:07:55"

uw 170 patients(+/- 3 patients)
FINISHED [146.5 secs]

ucdavis 41987 patients(+/- 3 patients)
FINISHED [146.5 secs]

ucsf 822 patients(+/- 3 patients)
FINISHED [146.5 secs]

AGGREGATED 42979 patients(+/- 3 patients)
FINISHED [146.5 secs]

of the data. There are clear comparable risks in translational research to clinical practice – particularly when the potential impact of translational research can be on entire populations, but the present emphasis of this research is on the secondary use of clinical data, in forms that are highly regulated and de-identified as to mediate risk to individual patient groups, researchers or institutions. Based in part on this project, though also in the context of the larger CTSA informatics community, researchers are increasingly interested in applying their unique domain-focused requirements to informing how upstream clinical source systems capture and semantically store data, which shifts the expected role of a researcher from an end-user to one of a collaborator or participant in the development process.

Ongoing Evaluation

We see the need for evaluation of the utility of this end-user clinical-translational query interface to be on-going. Given the diversity of research needs and the rapid advancement in the quantity of electronic clinical and biological data available, the ability to measure the relative utility of this form of service interface within the context of researchers design and implementation workflows will provide valuable requirements and insights to advancing data-driven collaborations.

While our current study focused on discovery of generalized end-user computing needs for clinical translational scientists, our future evaluation plans are more specific to the i2b2 workbench environment as a method to address our users' needs. Multi-institutional survey tools will be used to determine user response to either a video demonstration of the tool or after a 'hands-on' training session to see how the reality of the i2b2 workbench meshes with their anticipated needs. Next, in depth usability testing, via think aloud sessions with pre-defined use cases, will be done to examine how researchers build queries and interact with the tool. In each phase of the evaluation, a report will be returned to the development team highlighting findings with recommendations for improvement.

Finally, although this study is limited by using data from interviews at a single academic site, it is unique because it establishes an understanding of the needs of scientists in data-rich translational research workflow environments. In-depth studies illuminating other aspects of collaboration, impact, and work practices in translational research will serve as a useful supplement to statistical accounts of user needs and system use.

CONCLUSION

Our study is one of the first of its kind to characterize the data discovery and end-user computing needs of clinical translational scientists. As suggested by our results, there remain major challenges for clinical translational scientists to gain access to and subsequently integrate clinical and laboratory data that is located within and across multiple institutions. Some of these challenges can be met with sustaining and information-needs focused partnerships with end-user scientists and their research IT staff throughout the development and deployment process of clinical translational data discovery tools.

The CICTR project has used the themes of this study to inform the refinement of the data discovery environment as to begin to meet the different end user needs. With the increase in data available in the network (in excess of 4 million patients as of early 2010), and richness of data sources (now including data on medications and laboratory tests) there have been corresponding on-going challenges to introduce these capabilities and limitations of such an interventional architecture. We are now focusing on establishing knowledge development workflows that allow us to capture and then map the different data sources such that we can leverage the input of end-user scientists' perspectives on utility, richness and quality against the data-level transformation requirements that is necessary for our clinical terminologists and software developers to implement. These complementary and overlapping workflows are allowing us to capture and operationalize information representation needs more efficiently, and are in turn enhancing scientists' ability to ask challenging research questions.

Our next steps are to iteratively address further issues raised by users in the current prototype and expand to a multi-site evaluation method that employs other forms of summative and formative evaluation as described above. We hope that this user-centered approach to designing our collaborative information environment will help us better understand and serve the needs of scientists as they address novel clinical translational research challenges.

ACKNOWLEDGMENT

This work has been supported by NCRR grants UL1 RR025014 and UL1 RR024146, and NCRR, NIH and DHHS under Contract No. HHSN268200700031C.

REFERENCES

Anderson, N. R., Lee, E. S., Brockenbrough, J. S., Minie, M. E., Fuller, S., Brinkley, J., & Tarczy-Hornoch, P. (2007). Issues in biomedical research data management and analysis: Needs and barriers. *Journal of the American Medical Informatics Association*, *14*(4), 478–488. doi:10.1197/jamia.M2114

Ash, J. S., Anderson, N., & Tarczy-Hornoch, P. (2008). People and organizational issues in research systems implementation. *Journal of the American Medical Informatics Association*, *15*(3), 283–289. doi:10.1197/jamia.M2582

Ash, J. S., Berg, M., & Coiera, E. (2004). Some unintended consequences of information technology in health care: The nature of patient care information system-related errors. *Journal of the American Medical Informatics Association*, *11*(2), 104–112. doi:10.1197/jamia.M1471

Bartlett, J. C., & Toms, E. G. (2005). Developing a protocol for bioinformatics analysis: An integrated information behavior and task analysis approach. *Journal of the American Society for Information Science and Technology*, *56*(5), 469–482. doi:10.1002/asi.20136

Bernstam, E. V., Hersh, W. R., Johnson, S. B., Chute, C. G., Nguyen, H., & Nagarajan, R. (2009). Synergies and distinctions between computational disciplines in biomedical research: Perspective from the Clinical and Translational Science Award programs. *Academic Medicine*, *84*(7), 964–970. doi:10.1097/ACM.0b013e3181a8144d

Carver, J. C., Kendall, R. P., Squires, S. E., & Post, D. E. (2007). Software development environments for scientific and engineering software: A series of case studies. In *Proceedings of the International Conference on Software Engineering* (pp. 550-559).

Chilana, P. K., Palmer, C. L., & Ko, A. J. (2009). Comparing bioinformatics software development by computer scientists and biologists: An exploratory study. In *Proceedings of the ICSE Workshop on Software Engineering for Computational Science and Engineering* (pp. 72-79).

Hackos, J., & Redish, J. (1998). *User and task analysis for interface design*. New York, NY: John Wiley & Sons.

Letondal, C. (2005). Participatory programming: Developing programmable bioinformatics tools for end users. In Lieberman, H., Paterno, F., & Wulf, V. (Eds.), *End-user development* (pp. 207–242). Dordrecht, The Netherlands: Springer-Verlag.

MacMullen, W. J., & Denn, S. O. (2005). Information problems in molecular biology and bioinformatics. *Journal of the American Society for Information Science and Technology*, *56*(5), 447–456. doi:10.1002/asi.20134

Massar, J., Travers, M., Elhai, J., & Shrager, J. (2005). BioLingua: A programmable knowledge environment for biologists. *Bioinformatics (Oxford, England)*, *21*(2), 199–207. doi:10.1093/bioinformatics/bth465

Mathe, J., Duncavage, S., Werner, J., Malin, B., Ledeczi, A., & Sztipanovits, J. (2007). Implementing a model-based design environment for clinical information systems. In *Proceedings of the ACM/IEEE International Workshop on Model-Based Trustworthy Health Information Systems* (pp. 399-408).

Miller, R. H., & Sim, I. (2004). Physicians' use of electronic medical records: barriers and solutions. *Health Affairs*, *23*(2), 116–126. doi:10.1377/hlthaff.23.2.116

Murphy, S., Gainer, V., & Chueh, H. (2003). A visual interface designed for novice users to find research patient cohorts in a large biomedical database. In *Proceedings of the AMIA Annual Symposium* (pp. 489-493).

Plaisant, C., Lam, S., Shneiderman, B., Smith, M., Roseman, D., Marchand, G., et al. (2008). Searching electronic health records for temporal patterns in patient histories: A case study with microsoft Amalga. In *Proceedings of the AMIA Annual Symposium* (pp. 601-605).

Rose, A. F., Schnipper, J. L., Park, E. R., Poon, E. G., Li, Q., & Middleton, B. (2005). Using qualitative studies to improve the usability of an EMR. *Journal of Biomedical Informatics*, *38*(1), 51–60. doi:10.1016/j.jbi.2004.11.006

Segal, J. (2005). When software engineers met research scientists: A case study. *Empirical Software Engineering*, *10*(4), 517–536. doi:10.1007/s10664-005-3865-y

Segal, J. (2009). Software development cultures and cooperation problems: A field study of the early stages of development of software for a scientific community. *Computer Supported Cooperative Work*, *18*(5-6), 581–606. doi:10.1007/s10606-009-9096-9

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Newbury Park, CA: Sage.

Tran, D., Dubay, C., Gorman, P., & Hersh, W. (2004). Applying task analysis to describe and facilitate bioinformatics tasks. *Medinfo*, *11*(2), 818–822.

Umarji, M., & Seaman, C. (2008). Informing design of a search tool for bioinformatics. In *Proceedings of the ICSE Workshop on Software Engineering for Computational Science and Engineering.*

Wang, T. D., Plaisant, C., Quinn, A. J., Stanchak, R., Murphy, S., & Shneiderman, B. (2008). Aligning temporal data by sentinel events: discovering patterns in electronic health records. In *Proceeding of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 457-466).

Zerhouni, E. (2007). Translational research: Moving discovery to practice. *Clinical Pharmacology and Therapeutics*, *81*, 126–128. doi:10.1038/sj.clpt.6100029

Parmit K. Chilana is a PhD student at the Information School of the University of Washington (UW), specializing in human-computer interaction (HCI). A research interest of Parmit has been the application of usability and user-centered design principles for improving software design in biomedical and health informatics domains and she has collaborated on various projects within UW Medicine. Based on this experience, she has done further investigation into the challenges of usability practices in other highly complex domains and has synthesized pedagogical implications for HCI. Parmit's current research focus is on better understanding post-deployment software usability and redesigning help interfaces. Her recent projects have looked at how users express unwanted software behaviors and how support professionals and developers respond to user-reported software issues in open source and commercial software development contexts. She is currently exploring the design of contextual-help tools for web applications which leverage crowdsourced solutions from users. Parmit received her MS in Library and Information Science from the University of Illinois at Urbana-Champaign and BSc in Computing Science from Simon Fraser University, Canada.

Elishema Fishman is a recent graduate of the MS in Information Management program at the University of Washington's Information School. She holds a BA in English from UCLA. Elishema's research interests lie in user-centered design and content management, with a goal of effectively looking at and understanding user needs. Throughout her studies at the University of Washington, Elishema worked as a Research Assistant at the Institute of Translational Health Sciences, part of UW Medicine. Elishema spent several years working as a public relations specialist, prior to obtaining her Masters degree where she gained experience in writing and editing. Elishema's long-term career interests lie in information architecture with a focus on User Experience design.

Estella M. Geraghty is an Assistant Professor of Clinical Internal Medicine at the University of California, Davis. She earned her MD from UC Davis in 2002 and also holds both Masters of Medical Informatics and Masters in Public Health degrees from UC Davis. Dr. Geraghty is board certified in Internal Medicine and is also among the charter class of Certified in Public Health professionals. Her research interests revolve around spatial epidemiology and geographic information systems (GIS) as methodologies for understanding the interplay between health and the environment. One current project investigates the relationship between aerial pesticide spraying for West Nile virus and health effects. She is also working on a multi-disciplinary, multi-scalar, mixed-method approach to understanding youth outcomes in a 9-county Sacramento region. Outcomes, including health, are analyzed by their geographies to understand disparities and vulnerabilities in the population. For the last two years she has been involved in the evaluation component of the Cohort Discovery Tool, powered by i2b2 (informatics for integrating the bench and the bedside). This is a multi-institution project leveraging EMR data to improve cohort discovery among NIH designated CTSAs.

Peter Tarczy-Hornoch is an elected Fellow of the American College of Medical Informatics and an elected member of the Society for Pediatric Research. He serves as the Head of the Division of Biomedical and Health Informatics. He also serves in a variety of leadership roles throughout the School of Medicine including leading the research and service activities of the Biomedical Informatics Core of the Institute of Translational Health Sciences (the regional CTSA award) and serving as the Director of Research and Data Integration for UW Medicine Information Technology Services (the operational clinical computing group). His current research focuses on data integration of biomedical and health data including looking at ways of handling semi structured data, representing uncertainty at various levels in the system, and doing computerized reasoning over integrated data. His research builds on collaborations with biologists and clinical and translational researchers looking at: a) large scale functional gene annotation, b) SNPs for elucidation of disease mechanisms, and c) as part of the Institute of Translational Health Sciences and the Northwest Institute of Genetic Medicine research in the area of collaborative integrated analysis of a combination of clinical data, experimental biological data, and clinical/translational research study data.

Fredric M. (Fred) Wolf is Professor and Chair of the Department of Medical Education and Biomedical and Health Informatics in the School of Medicine and Adjunct Professor of Health Services and Epidemiology in the School of Public Health at the University of Washington. He was formerly Professor of Medical Education and Director of the Learning Resource Center and the Laboratory for Computing and Cognition at the University of Michigan Medical School. He has many years of experience in educational psychology/evaluation and measurement, medical education, and health services research. He is a member of the international Cochrane Collaboration and former Visiting Scholar at UK Cochrane Centre and Green College, University of Oxford. His research has focused on a) dissemination and evaluation of new technology, including decision support systems, b) clinical decision making and judgment under uncertainty, c) evidence based medicine, systematic reviews and meta-analysis of educational and healthcare interventions, and d) evaluation of clinical and translational research interventions and training.

Nick R. Anderson's academic research areas include user needs analysis of information management issues faced by small research laboratories, clinical decision support, knowledge transformation and delivery, and the study of how information tools support policy development for biospecimen data sharing. He is the Principal Investigator of the Cross-Institutional Clinical Translational Research project (CICTR), which is building a collaboration to provide query discovery across de-identified clinical data from 3 academic research hospitals for cohort recruitment, and Co-Investigator on the Cancer Biospecimen Portal project, a collaboration between the University of Washington, Fred Hutchinson Cancer Research Center and Seattle Children's Research Institute. He also leads a range of collaborative large-scale clinical data projects that are exploring the boundaries of sharing of sensitive clinical information under evolving regulatory policies. He has faculty appointments as Assistant Professor in the Division of Biomedical Health Informatics and Adjunct Assistant Professor in the Department of Bioethics and Humanities at the University of Washington and is the Associate Director of the Biomedical Informatics Core for the Institute of Translational Health Sciences (ITHS).