MEDINFO 2013
C.U. Lehmann et al. (Eds.)

doi:10.3233/978-1-61499-289-9-632

Developing a Self-Service Query Interface for Re-Using De-Identified Electronic Health Record Data

James J. Cimino,[a] Elaine J. Ayres,[a] Andrea Beri,[b] Robert Freedman,[b] Ellen Oberholtzer,[b] Sachi Rath,[b]

[a] *Laboratory for Informatics Development, the Clinical Center of the US National Institutes of Health, Bethesda, Maryland, USA*
[b] *Computer Sciences Corporation, Bethesda, Maryland, USA*

Abstract

The US National Institutes of Health has developed a repository of clinical research data drawn in part from electronic health records. A new de-identified data query tool under development has been developed to support re-use of these data. We used a collection of 30 human-mediated user queries to determine whether features of the tool will be sufficient to allow users to carry out the queries themselves. The results show that the tool implemented in February 2013 will carry out a small percentage of user queries but the planned extensions will be sufficient for carrying out the majority of such queries. Future development of the tool will include extensions that correspond to the features found in human-mediated queries.

Keywords:

Clinical Research, Data Repositories, Controlled Terminologies and Ontologies, Data Reuse.

Introduction

Data captured in the course of routine patient care are widely seen as having great potential for re-use in clinical and translational research [1]. The Biomedical Translational Research Information System (BTRIS) at the US National Institutes of Health (NIH) contains such data and provides a user interface that supports direct queries by clinical researchers [2]. While BTRIS is popular with investigators who desire access to data collected in their own studies (in identified form), use of data in de-identified form for secondary purposes has been less widely used. This paper describes our efforts at understanding the kinds of queries that our investigators seek to carry out and efforts to design a user interface that supports such queries.

Background

Access to De-Identified Clinical Data Repositories

Clinical research data repositories vary in their approaches for user access to de-identified data sets. Many systems provide human-mediated queries (typically carried out by a systems analyst), while some systems offer a self-service user interface.

The i2b2 system at Partners HealthCare allows direct user access to de-identified data based on the role and training of the user, as well as the technical security of the client machine [3]. Self-initiated queries are permissible for approximate patient counts or for summary data. Those with additional training and credentials may access limited data sets with line item patient data. Other academic medical centers have implemented the open source i2b2 software "hive" for their clinical research data repositories. Cincinnati Children's Hospital [4] and Georgia Health Sciences University [5] allow users direct access to the repository only for de-identified data summaries.

The Stanford Translational Research Integrated Database Environment (STRIDE) allows direct user access to their Cohort Discovery Tool providing summary statistics about patient research cohorts. With appropriate permissions, users may conduct a line-item review of de-identified patient data to screen subjects for inclusion in a study cohort [6].

BioGRID Australia is a federated national research repository. Users may query approved research databases for de-identified data. However, these queries require a working knowledge of databases and data organization and the ability to build a logical query statement [7].

While each of the systems described above provides an interface for end users, when researchers find that they are unable to obtain the desired data directly, they need a human intermediary to obtain their data. The kinds of queries that require this type of intervention, and the rate at which they occur, does not generally appear in the published informatics literature.

The NIH Biomedical Translational Research Information System (BTRIS)

Development of the first full version of BTRIS began in October 2008 and became operational in July of 2009. The system is comprised of data from over 35 NIH sources containing 477,000 human subjects seen at the NIH Clinical Center from 1953 to the present.

Major data sets include patient demographics, research study enrollment information, laboratory results, medication orders and administration, and a wide variety of clinical notes and reports from the two electronic health record (EHR) systems that have operated at the Clinical Center from 1976 to 2004 and 2004 to present. BTRIS's database is a hybrid between a traditional entity-relation model and an entity-attribute-value (EAV) model. All data in BTRIS are coded with the NIH's Research Entities Dictionary (RED), which is a single, controlled terminology created from a merger of terminologies from BTRIS data sources [2].

BTRIS users access data in identified form with a set of reports developed based on a commercial "business intelligence" system (Cognos, IBM, Armonk, New York, USA). Reports are typically related to specific data domains (laboratory results, medication data, etc.) with some cross-domain and drill-through capabilities.

BTRIS De-Identified Queries, Version 1.0

Access to data in de-identified form is accomplished in BTRIS using a custom modification to the Cognos user interface with several features that were difficult to achieve in the native Cognos system, including workflow steps related to automatically documenting data use to allow appropriate oversight by the NIH Office of Human Subjects Research Protections. These steps allow users to perform queries prior to obtaining authorization, and then when desired data are found, to complete the required documentation for automatic certification of use of the data prior to downloading them for analysis. Current reports include summaries of demographic data, laboratory and vital sign data, and medication data. Detailed reports are available for laboratory and vital signs data. While summary reports provide access to the entire database, access to detailed data is limited to those related to protocols terminated over five years* ago, to avoid the need to notify original investigators of more recent studies, as per NIH policy.

While usage of identified data averages about 350 reports per week, usage of de-identified data averages fewer than five per week. Investigators interested in such data report that they are unable to carry out the queries they need using the current self-service user interface. When this occurs, we offer to mediate information retrieval by encouraging investigators to describe their research questions to us. We then perform the queries manually, using Structured Query Language (SQL) queries directly against the BTRIS tables.

Based on initial user queries, we established a set of requirements for the next version of the BTRIS de-identified query tool. These requirements included types of data, constraints on data, and data sets formed from inclusion from multiple data sources. In particular, users expressed a need for selecting data based on temporal relationships (e.g., "find me all the patients that had a complete blood count less than one month after the third dose of drug X"), cardinality (e.g., "find me all the patients who had at least three doses of drug X"), and merged data sets (e.g., "get me all the hematology labs and radiology reports on all these patients").

Based on these requirements, we designed the next version of the BTRIS de-identified query tool. Development began in early 2012, with a release in February 2013. In the interim, we continued to collect and execute requests for mediated queries.

Methods

Design of De-Identified Queries, Version 2.0

In creating our initial design, we were favorably influenced by the "drag-and-drop" user interface adopted by the i2b2 warehouse query tool [8]. We sought to extend the user interface to allow more than three panels, additional prompts for cardinality, temporal relationships between sets of query results, and cross-domain queries. We also developed a terminology look-up [9,10] that allows users to search the RED for specific classes of data within each data domain. System requirements were enumerated and then divided into those to be included in the first version of the new tool and later versions.

Collection of User Queries

User queries were collected by the BTRIS User Support team as part of their routine user interactions. Each query was developed manually after discussion between a BTRIS database analyst and user. Requests were classified based on the data domains requested (demographics, laboratory results, etc.), the types of data attributes specified (date range, value range, cardinality, etc.), and the relationships between the data (Boolean, temporal, etc.).

Comparison of User Requirements with Current Design

We assessed the theoretical capabilities for the de-identified query tool to address the previous user query requirements with respect to the various domains, attributes and relationships. Analysis considered the capabilities of the initial version of the tool as well as subsequent versions.

Results

Design of De-Identified Queries, Version 2.0

The tool is based on a set of domain-specific query modules that can be selected by the users and related to each other through "AND" and "OR" relationships. Multiple modules may be selected. Each module includes a collection of optional query prompts (e.g., date and value and patient age ranges).

The modules are defined using XML data structures that specify the prompts to be included and the SQL queries generated by the tool based on user input to the prompts. The prompts themselves are defined with reusable widgets. Figure 1 shows a screen shot of the current user interface, depicting the selection of several modules, completion of the prompt fields with various values (including terminology look-up), and the relationships between the prompts.

Rather than wait until all modules and all prompts were completed, we chose to deploy an initial version of the system with what we consider to be a "critical mass" of the most popular data domain modules and the most important query parameter prompts. The creation of additional domains will be fairly straightforward through the reuse of existing prompt widgets, while the introduction of new widgets can proceed in a non-disruptive manner by simply extending the XML for the relevant module(s). The domains and features of the initial and later versions of the system are shown in Table 1.

Collection of User Queries

A total of 30 user queries were collected over the past year. The domains of interest, the attributes of those domains and any relationships are shown in Table 2.

Comparison of User Requirements with Current Design

The ability of the initial and later versions of the query tool to represent features of each of the 30 queries is represented as bold font and italic font (respectively) in Table 2. The user queries touched on all the domains implemented in the current version of the query tool, as well as most of the planned domains. As of this writing, a few domains (Alerts, Allergies, ECG, and Pulmonary Function) have yet to be requested. Four queries involved domains that we had not previously planned to provide: Research Study (three queries) and Admission/Discharge/Transfer events (ADT; one query).

The user queries also touched on all of the query attributes implemented in the current version of the query tool. No queries required the use of the one attribute planned for a future version of the tool (Cardinality). One query would require an unplanned attribute (location).

* Since implementation of the new system, the NIH policy has reduced this time limit to two years.

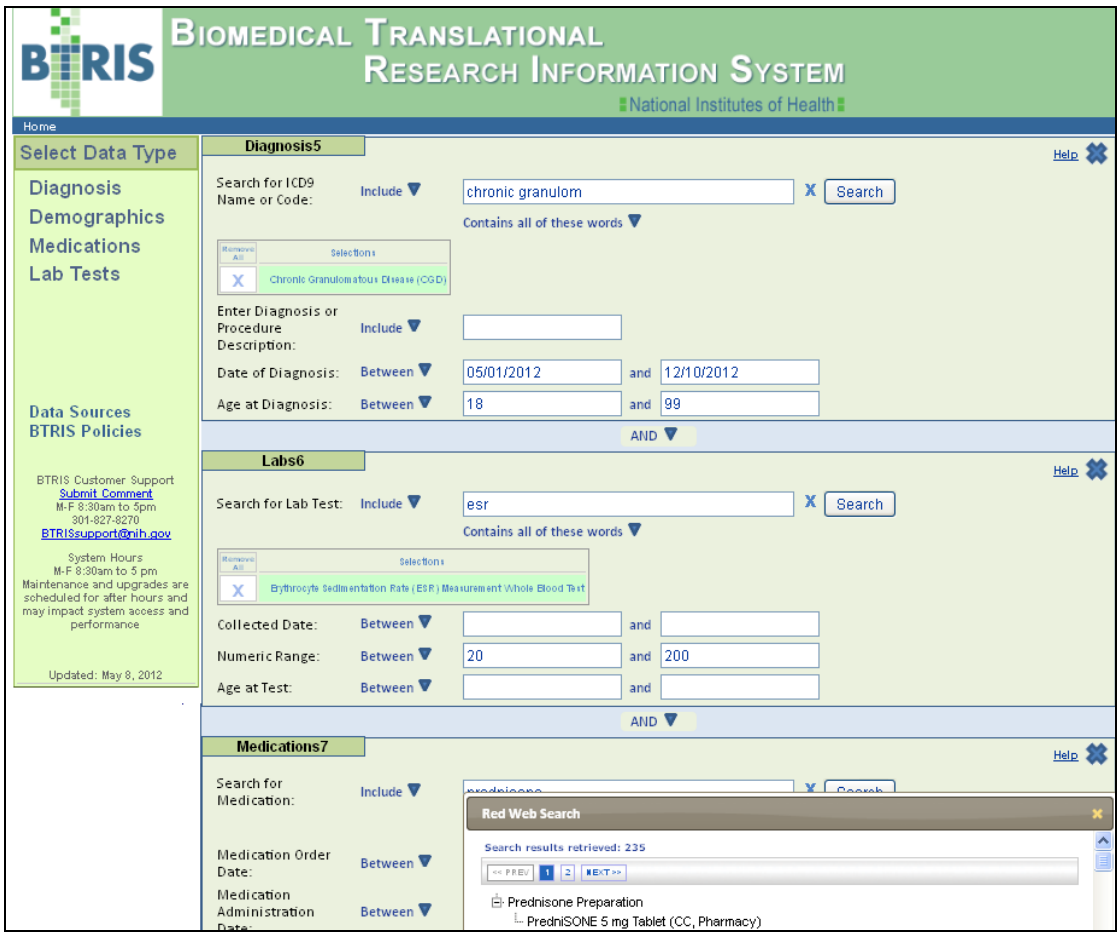

Figure 1- The first version of the de-identified query tool. Three query modules have been dragged from the left-hand menu into the main frame and "ANDed" to each other. Note that the Diagnosis module has a controlled term (Chronic Granulomatous Disease) specified, as well as date and age ranges. The Laboratory Test module also has a controlled term (Erythrocyte Sedimentation Rate) selected and a range specified for the test value. The Medication Module shows the term look up ("prednisone").

Finally, the user queries made use of both the "AND" and "OR" relationships. In addition, four queries will require the planned "DATE" relationship and one query will require the planned "NOT" relationship. No query would require any unplanned relationships.

Discussion

The results of this preliminary analysis of our new query tool are encouraging. Although only four of the 30 queries (queries 10, 19, 26 and 30, in Table 2) can be handled by the initial version of the tool, 25 of the 30 queries (all but queries 7, 15, 20, 27, and 28 in Table 2) will be handled by the later planned versions of the tool. Based on our experience with the imple-

Table 1- Features Planned for De-Identified Query Tool

Query Function	Version 1	Later Versions
Domains	Demographics, Diagnoses, Procedures, Medications, Laboratory Results	Admission/Discharge/Transfer, Alerts, Allergies, Blood Bank, Clinical Documents (notes written by healthcare personnel caring for the patients), ECG, Echo, Pathology, Pulmonary Functions, Radiology, Vital Signs
Query Features	Controlled Terminology, Age Range, Date Ranges, Value Range	Cardinality
Relationships	AND, OR	NOT, Before, After

mentation of the tool thus far, we believe that the addition of new domains and attributes will be straightforward; their limitations are development time, not algorithmic complexity.

We hold no illusions that the simple inclusion of query features in our tool will guarantee success for our users. However, the tool underwent user acceptance testing that demonstrated the need for some minor adjustments to the user interface that has now been released to NIH users. We expect that, no matter how easy-to-use we think the system will be, user training will be needed to teach the overall paradigm of Boolean combinations in attribute-limited domain data sets. However, we are encouraged that the features users seem to need are present, planned, or not beyond our ability to provide.

The three types of data needed to satisfy the queries is already available in the BTRIS database (the patients' research study enrollments) or planned for inclusion (ADT and location data). Expanding the query tool to access these data should not present major challenges.

A fourth data type, Adverse Events, will be more problematic. Neither of the EHRs that BTRIS derives data from includes event data that is specifically identified as "adverse". Instead, "adverse" may be implied by an abnormal test result, the initiation or cessation of a medication, or the addition of a new diagnosis to the patient's problem list. In fact, detection of such events in EHR data is a widely regarded informatics research are [11]. However, NIH investigators are now beginning to add adverse event data directly into BTRIS to facilitate automated reporting of study trial results to clinical trials registries, [12] such as ClinicalTrials.gov [13]. Thus, at least for a subset of patients in the BTRIS database, adverse event data *will* be available.

While the tool may support the creation of an effective query to the BTRIS database, this is only part of the task of data reuse. In mediating queries for users, our analysts have also carried out extensive data pruning and mining. For example, finding relevant clinical documents requires not only searching for specific text phrases, but also the expansion of those searches to include additional synonymous expressions and then manually filtering the results to remove the false positives. Natural language processing may be brought to bear on this task, but for the near term, users will likely need some assistance with this process.

The user interface for our tool is tightly linked to the BTRIS database. However, a number of aspects of our approach may have applicability to other institutions within the United States and internationally. First, the kinds of data we include in BTRIS are similar to those found in electronic health record, making our experience with user queries generally applicable. Second, the types of technical characteristics of user queries (controlled terminologies within various domains and Boolean combinations of data across domains) are, we believe, very typical of the kinds of queries that any researcher might pose to a clinical repository, making our user interface design relevant to developers of similar systems. Third, because the user interface generates SQL queries, our software could be adapted to work with other clinical data repositories, or even multiple repositories.

Some of the policy issues around use of our tool have been addressed by adapting existing policies related to patient protections and reuse of data from paper and electronic health records. Because we are at the NIH, where all patients are research subjects, additional policies have been needed to strike a balance between recognition of the original investigators' rights to first publication and the need to maximize the value of research data through promotion of their reuse. Expansion of the use of the tool to include data and users outside NIH, whether they be in the US or internationally, remains a challenge.

Table 2- Characteristics of User Queries. ***Bold items*** *are included in Version 1; italic items with underline are planned for future versions. Key to Domains: Admission/Discharge/Transfer, AE=Adverse events, B=Blood Bank, C=Clinical Documents(notes written by members of the healthcare team), D=Demographics, Dx=Diagnosis, E=Echo, L=Laboratory Tests, M=Medications, Mi=Microbiology, P=Pathology, R=Radiology, S=(Research) Study, V=Vital Signs. Key to Attributes: a=age, c=controlled terms, d=date range, e=expiration date, g=gender, l=location, n=normal ranges, t=text search, u=units of measure, v=discrete values. For example, Query #2 involved the domains Demographics (implemented) and Clinical Documents (planned), and one planned attribute (date range) and two unplanned attributes (normal ranges and units of measure). Of note, while searching is currently not allowed on these two attributes, they are currently included along with any reported laboratory results.*

Query #	Domains	Relationship	Attributes
1	*B* ,*C* ,**D,L,M**	**AND**	**c,d,t,v**
2	*C* ,**L**	**AND**	**d**,*n* ,*u*
3	**D**,*P*	**AND**	**d,t,v**
4	*C* ,**L**	**AND**	**d,t,v**
5	*A* ,*C* ,**D,Dx**,*E* ,**L,M,** *Mi* ,*V*	**AND**	**d,e,t,v**
6	**D**,*MI*	**AND**	**t,v**
7	AE**,D,L**	**AND**	**v**
8	**L**,*V*	**AND,** *Date*	**d,v**
9	**D,L,P**	**AND,** *Date*	**v,t**
10	**D,Di,L,M**	**AND**	**c,d,t,v**
11	**D,L**,*P*	**AND**	**a,v**
12	**D,L,M**,*P* ,*R*	**AND,** *Date*	**v,t**
13	*C* ,**D,Dx,L,M**,*Mi*	**AND,** *Date*	**t,v**
14	*C* ,**D,Dx,L,M**,UI,*V*	**AND**	**d,t,v**
15	**D,L,M**,S	**AND**	**v**
16	*C* ,**D,Dx,M**,*P* ,*V*	**AND**	**t,v**
17	*C* ,**D**	**AND**	**t,v**
18	*C* ,**D,L,M**,*V*	**AND**	**d,t,v**
19	**D,Dx,L**	**AND**	**t,v**
20	**Dx**	**AND**	**c**,**d**,*l*
21	**Dx**,*P* ,*R* ,*V*	**AND**	**t,v**
22	**Dx,L,M**,*P*	**AND**	**d,t,v**
23	*C* ,**Dx,L**,*P*	**AND, OR**	**t,v**
24	**L,M**,*P*	**AND, OR**	**v**
25	*C* ,**Dx,L,M**	**AND**	**t,v**
26	**D,L**	**OR**	**a,g,v**
27	*C* ,**Dx**,**L**,S	**AND**	**t,v**
28	*C* **,Dx,M**,S	**AND**, *Not*	**t,v**
29	*C* ,**D,L,M**	**AND**	**t,v**
30	**Dx,L**	**AND**	**v**

The results of this study are likely to lead to the development of additional requirements for the tool. We will continue with all the current planned extensions, even those that have not yet been required of user queries, because we believe it is only a matter of time before a user makes a request that will need them.

We expect to encounter many additional challenges during the creation of a successful tool that meets our users' needs and expectations, including optimization of query performance, expanding the synonyms in our controlled terminology look-up tool, rendering query results into formats that meet user requirements, and adding post-processing data analysis and visualization tool, to name a few. However, no amount of work on these areas would help the users unless the query can be formulated, allowing the data to be obtained in the first place. This study supports the contention that our tool at least will have this part of the task done right.

Conclusions

User queries for de-identified data are diverse in content and structure. Our analysis of human-mediated user queries shows that there is a good match between the requirements of those queries and the capabilities of our tool.

Acknowledgment

This research was supported by intramural research funds from the NIH Clinical Center and the US National Library of Medicine.

References

[1] Friedman C, Rigby M. Conceptualising and creating a global learning health system. Int J Med Inform. 2013: 82(4):e63-71.

[2] Cimino JJ, Ayres EJ. The clinical research data repository of the US National Institutes of Health. Stud Health Technol Inform. 2010;160(Pt 2):1299-303.

[3] Murphy, S, Gainer, V, Mendis, M, Churchill, S, Kohane, I. Strategies for maintaining patient privacy in i2b2. J Am Med Inform Assoc 2011;18:i103-i108.

[4] Jonnalagadda SR, Li D, Sohn S, Wu ST, Wagholikar K, Torii M, Liu H. Coreference analysis in clinical notes: a multi-pass sieve with alternate anaphora resolution modules. J Am Med Inform Assoc. 2012 Sep-Oct;19(5):867-74.

[5] Georgia Health Sciences University i2b2 http://www.georg-iahealth.edu/OCIS/i2b2.html Accessed December 10, 2012

[6] Lowe H, Ferris T, Hernandez P, Weber S. STRIDE – An integrated standards-based translational research informatics platform. AMIA Annu Symp Proc 2009:391-95.

[7] Hibbert M, Lohrey J, Melnikoff S. Integration of data for research. Stud Health Technol Inform. 2010;151:461-75.

[8] Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010 Mar-Apr;17(2):124-30.

[9] Cimino JJ, Ayres EJ, Remennik L. Resolving Terminology Issues for a Clinical Research Data Warehouse. Payne PR, ed. 2009 AMIA Clinical Research Informatics Summit, San Francisco, CA (poster).

[10] Cimino JJ, Ayres EJ, Rath S. Bridging the Data-Terminology-Researcher Gap: User Interface Design for Terminology Browsing in the NIH's Biomedical Translational Research Information System. Kahn MG, ed. 2012 AMIA Clinical Research Informatics Summit, San Francisco, CA (poster).

[11] Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. J Am Med Inform Assoc. 2005 Jul-Aug;12(4):448-57.

[12] Cimino JJ, Ayres EJ, Rath S, Freedman R. Automated Submission of Clinical Trials Results to ClinicalTrials.gov. 2013 *AMIA* Summit on Clinical Research Informatics (Poster), San Francisco, CA.

[13] Zarin DA, Tse T, Williams RJ, Califf RM, Ide NC. The ClinicalTrials.gov results database--update and key issues. N Engl J Med. 2011 Mar 3;364(9):852-60.

Address for correspondence:

James J. Cimino, M.D.
Laboratory for Informatics Development
US National Institutes Clinical Center
10 Center Drive, Room 6-2551
Bethesda, Maryland, 20892 USA
James.Cimino@nih.gov