Characterizing Secondary Use of Clinical Data

E. Sally Lee, PhD[1], R. Anthony Black, MS[1], Robert D. Harrington, MD[5], Peter Tarczy-Hornoch, MD, FACMI[1, 2, 3, 4]

Institute of Translational Health Sciences[1]; Department of Biomedical Informatics and Medical Education[2]; Department of Pediatrics[3]; Department of Computer Science and Engineering[4]; Department of Medicine, Harborview Medical Center[5]; University of Washington, Seattle, WA

Abstract

The increasing reliance on electronic health data has created new opportunities for the secondary use of clinical data to impact practice. We analyzed the secondary uses of clinical data at the University of Washington (UW) to better understand the types of users and uses as well as the benefits and limitations of these electronic data. At the UW, a diverse population is utilizing different elements of clinical data to conduct a wide variety of studies. Investigators are using clinical data to explore research questions, determine study feasibility and to reduce the burden of manual chart abstraction. Discovered limitations include difficult-to-*use data formatting, researchers'* lack of understanding about the data structure and organization resulting in mistrust, and difficulty generalizing data to fit needs of many specialized users.

Introduction

Adoption of electronic medical record (EMR) systems have increased significantly over last decade due to national initiatives such as Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009[1]. These initiatives provided financial incentives and investments to support the adoption and use of EMRs, and mandated "meaningful use" of such systems[2]. The resulting expansion of electronic health records (EHRs) created new opportunities for the secondary use of clinical data, leveraging existing data to rapidly impact practice through a learning healthcare system that facilitates research and quality improvement[3,4,5].

According to a survey of 35 Clinical and Translational Science Award (CTSA) organizations and the NIH Clinical Center in 2008 and 2010, many institutions have created or are developing research data repositories derived from EHRs[6]. Another survey of 17 institutions in 2012 demonstrated a marked increase in clinical data repositories used for research since 2007[7]. These repositories represent the foundational infrastructure needed to enable the secondary use of clinical data.

Since electronic clinical data has become widely available, the potential uses, applications and successes for secondary use have been described extensively[3,5,8] as have the challenges and barriers of using the data[9,10,11]. In a 2014 article, Danciu et al. described Vanderbilt's approach to secondary use of clinical data and presented a set of practical "lessons learned"[3]. We complement their contribution by describing users and usage characteristics of secondary use of clinical data at the University of Washington (UW). By doing so, we provide insight into the types of users and uses of the secondary clinical data as well as some of the benefits and limitations.

Methods

Overview of secondary data use services

The Biomedical informatics (BMI) team at the Institute of Translational Health Science (ITHS, a CTSA funded institution at the UW) offers two main services for researchers seeking to access existing clinical data: 1) customized data consultation and queries, and 2) cohort estimation through the i2b2 based De-identified Clinical Data Repository (DCDR). The BMI team provides low-cost, fee-based access to the University of Washington Clinical Data Repository (UWCDR), which includes data from various UW clinical systems (currently the UWCDR based

on the Caradigm Amalga platform contains data from ~130 interfaces and systems). Since 2009, the BMI team, beginning with one person and growing into a three person team by November 2012, has provided customized solutions for researchers wanting clinical data.

Since its pilot launch in September 2013, the DCDR has been available as a no-cost cohort estimation and feasibility determination tool that users can access without an individual IRB approval due to the data repository IRB that is in place. The DCDR contains a subset of de-identified data from the UW CDR that researchers can query through a web-based graphical query interface in the i2b2 platform[12]. The tool requires the identification of search criteria and returns an aggregate count and summary of the patients who meet the criteria. The current data elements available in DCDR are: 1) allergy, 2) demographic (age, gender, language, marital status, race, religion, vital status), 3) diagnoses (through billing and order summary), 4) immunization, 5) labs (subset), 6) medication orders, 7) microbiology labs (culture, specimen, subset of susceptibility), 8) procedures (through billing and order summary), 9) problem list (outpatient only), 10) visit details (age at visit, institution, discharge disposition, visit type), and 11) vitals (blood pressure, BMI, height, weight, heart rate, temperature).

Collection of the characteristics of users and usage statistics for the secondary clinical data use

User profiles and clinical data usage statistics were collected for the two main services described above. Information was gathered regarding custom consult data requests from January 2012 to April 2014 and DCDR usage from September 2013 to August 2014. To characterize the custom consult data, we examined email correspondences, hours billed and the queries created during the services. To characterize the DCDR usage, we studied access request information as well as logged queries.

We also contacted twenty users who have used the DCDR more than once and conducted brief, thirty-minute semi-structured interviews with ten users who responded to our request to explore users' opinions about the DCDR and more broadly about secondary use of clinical data. The interview questions were open-ended and asked both specifically about users' opinion on the DCDR as well as the type of informatics service that would be helpful for the users conducting research (e.g. difficulties using the tool, enhancements would like to see, any type of informatics service that would help in research). The preliminary analysis of the interview content was performed using the standard qualitative methodologies guided by grounded theory[13]. As the tool becomes more widely used, we will conduct a more thorough qualitative evaluation study of the DCDR.

Table 1. Top 20 data elements that were requested

Data queried	Requests
Demographic (Age, Gender, Race, Ethnicity, Language, Vital Status, Address/Zip)	89
Visit Details (Institution, Clinic, Type, Date, Frequency, Insurance, Service, Interpreter)	85
Diagnoses (by ICD9)	67
Lab Values (Discrete)	41
Procedures by CPT or ICD9	28
Inpatient Clinical Events (Meds, Echo, Notes, Infusion, Ventilation, Gas stats)	26
Medication Orders	22
Vitals (Weight, Height, BMI, BP)	21
Radiology Information	14
Clinical Notes	14
Provider Information	13
Problem List	11
Pathology Report	8
Social History (e.g. Smoking)	7
Appointments	7
Lab Values (Textual; e.g. Microbiology)	9
Allergy Information	3
Surgery information in Anesthesia System	2
Immunization	2
Emergency Tracking Information	2

Results

BMI consult requests

The BMI consult team served 106 distinct users between January 2012 and April 2014 (excluding brief telephone consults), resulting in 129 distinct consult requests. Sixty seven (64%) of 106 users received clinical data and 15 users used consult services more than once. Users belonged to more than 25 distinct academic departments and held various ranks ranging from fellows and residents to full professors.

A unique consult request was defined as one that had a single funding source or a single human subjects application even though multiple interactions might have occurred over time for that request. Ninety-seven (75%) of 129 requests were completely fulfilled and 32 (24%) were incompletely fulfilled. The majority of consult requests fell into three categories: 1) data for a retrospective study (39), 2) research recruitment screening (36), and 3) grant or project feasibility (36). Consult request size was based on the number of hours required to fulfil the request and was categorized as: a) small $\leq$ 5 hours (52), b) medium 6-20 hours (32), and c) large $\geq$ 20 hours (24). Most of the data requests involved demographics, visit information and coded diagnoses. Table 1 summarizes top 20 data elements that were requested. If a single data request contained more than one data element, it was counted more than once.

Incompletely fulfilled requests were often due to client-side issues such as lack of funding or the project being placed on hold; however, some were due to difficulties with the data request. The three major difficulties with data requests were: 1) data were only available in textual notes and required natural language processing to extract, 2) requests included financial data which required extra permission, effort and time, and 3) data could only be obtained through sophisticated calculations and required time and resources.

DCDR usage statistics

The DCDR is currently being used by 100 distinct users (Figure 1), 71 of whom participated in in-person tool training. Users were able to run simple cohort estimation queries with minimal training (e.g. self-directed online learning modules); however, more difficult queries such as tying encounter level information to other data required more training. Sixty-six users queried the system more than once (excluding training queries), and 5 requested data after the cohort estimation. DCDR users belonged to more than 15 distinct academic departments and ranged in rank from residents and fellows to full professors.

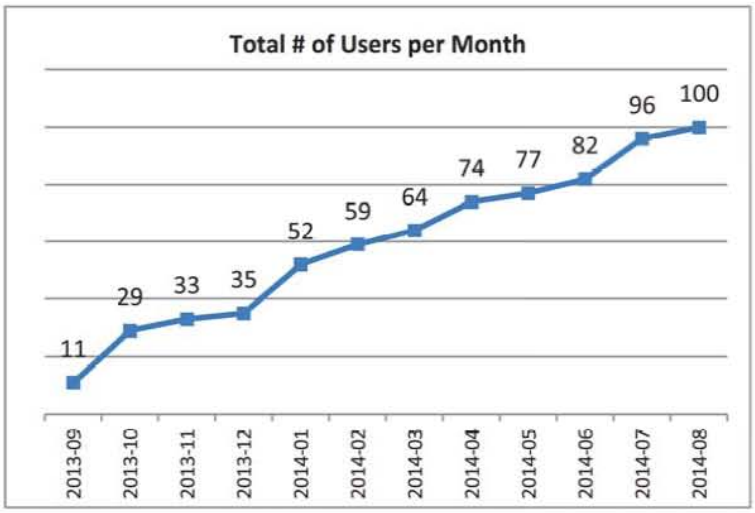

Figure 1. Total # of DCDR users per month.

Since its release, DCDR has been queried 1456 times (excluding training queries). The largest number of queries were run against the patient visit information such as date of service, location, and patient class (e.g. inpatient, outpatient), followed by medication orders and diagnoses. Table 2 summarizes the number of queries that looked at each of the data elements in DCDR. If a single query contained more than one element, that query was counted more than once.

Table 2. Number of queries that looked at various data elements.

Category	# Queries
Visit Details (date, location, patient class, disposition)	1292
Medication orders	620
Diagnoses	469
Procedures	263
Demographics (age, gender, language, race)	211
Microbiology lab results	181
Lab results	110
Problem List	98
Allergy	68
ID Note	33
Vitals (height, weight, temperature)	23
Immunization	6
Susceptibility	6

Preliminary analysis of interviews

Preliminary analysis of the interviews identified the values and limitation of the secondary use of clinical data. Users stated that simply having free access to aggregate data was critical for exploring research questions. Being able to readily

determine research feasibility was also noted as a great value. Outside the specific question of secondary use, users also noted that having access to support and training from those with a full understanding of the clinical data and the DCDR tool was significant.

The most frequently mentioned limitation of secondary use of clinical data was researcher's lack of knowledge about the data itself. When users were not operating with a common set of definitions and were not aware of full provenance of the data they received from the system (and thus unaware of the associated caveats and limitations) they had less confidence that the data would meet their needs. When operating on their own, they were often unsure that their queries were correct and were also unable to verify data validity. Furthermore, as researchers' questions are often specific to their specialties, data needs infrequently overlapped with those from other specialties and more generalized data was not seen as useful within each specialty.

Discussion

Benefits of secondary use of clinical data

The wide benefits of secondary use of clinical data is evident in the diversity of the user population and data requests, and the number of users who came back to use both services. A key benefit was the unhindered ability of investigators to explore data to generate research questions. Clinicians and researchers continually generate questions throughout their everyday activities, but without ready access to data collection and analytic capabilities they have no means to answer these questions. Providing immediate access to a clinical data repository enables the exploration of ideas and is critical to generating novel research questions. The determination of study feasibility is also extremely important since neither retrospective nor prospective studies can move forward without it. Not knowing whether there are sufficient study subjects or data to complete an investigation can lead to wasted time and effort and invalidate the results of a study that does errantly go forward. The low conversion rate from DCDR query to actual data requests may be evidence of this; clinicians and researchers may be probing the data and rejecting investigations that cannot be completed while focusing their energies on the minority of questions that can be explored by the available data. Another benefit of electronic data repositories is the significant reduction in time and effort required to conduct retrospective and prospective studies. Although manual chart abstraction is often still required, the ability to electronically screen a population to identify subjects and then collect a defined dataset on those subjects saves researchers enormous amount of time and effort.

Limitations of secondary use of clinical data

Many of the limitations of using of secondary use of clinical data that we identified have been previously reported including the major problem of clinical data not being available in a format that investigators can easily access[3,5,9]. For instance, in the care-based setting, much of the data is recorded in textual notes and reports that require sophisticated natural language processing to interpret programmatically. Furthermore, even when data are available in a more structured and searchable formats they might be under additional protections by the institution or may require complex calculations on a large number of variables to make the data useful. The high expense in terms of time and resources required to use such data may prevent researchers from incorporating these methods of electronic data abstraction into their investigations. Researchers also often lack sufficient knowledge about data definitions and provenance that leads them to question the data's authenticity and expend time and effort on validation. The feature space of clinical data is large and complex with multiple systems generating enormous amounts of diverse data, each with its own history, caveats and intricacies that might be known only to data specialists. As such, researchers are often unaware of how to best navigate and effectively utilize the data. While providing researchers ready access to data has helped to promote study exploration and feasibility, the use of such tools is hampered by limitation in how easily complex data can be intuitively extracted by the researchers.

Conclusion

The growth in popularity of our services indicates a strong and continuing demand for the ability to use existing clinical data to support research efforts. Even though projects involving secondary use of data are in their infancy researchers are effectively utilizing services to explore research questions and determine study feasibility with minimal expenditures in time and resources. These BMI services effectively multiply the efforts of researchers and expand the capability to ask research questions compared to historically under-resourced researchers.

Most researchers ask complex research questions that cannot be queried in a simple format. Often these complex questions require customized queries that researches are typically incapable of performing on their own. Tools like the DCDR are most effective when coupled with a knowledgeable BMI team of clinical data specialists that can facilitate this critical need. Creating an infrastructure of "self-service" IT tools and easy access to IT consultants facilitates trust in the data sources and systems, allowing investigators to effectively navigate and extract their needed data.

Acknowledgements

This publication was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR000423.

References

1. Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 2009 [cited 2014 August 28]. Available from: http://www.healthit.gov/policy-researchers-implementers/select-portions-hitech-act-and-relationship-onc-work.
2. Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. N Engl J Med. 2010;363(6):501-4.
3. Danciu I, Cowan JD, Basford M, Wang X, Saip A, Osgood S, et al. Secondary use of clinical data: The Vanderbilt approach. J Biomed Inform. 2014.
4. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. Sci Transl Med. 2010;2(57):57cm29.
5. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PR, Bernstam EV, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care. 2013;51(8 Suppl 3):S30-7.
6. MacKenzie SL, Wyatt MC, Schuff R, Tenenbaum JD, Anderson N. Practices and perspectives on building integrated data repositories: results from a 2010 CTSA survey. J Am Med Inform Assoc. 2012;19(e1):e119-24.
7. Murphy SN, Dubey A, Embi PJ, Harris PA, Richter BG, Turisco F, et al. Current state of information technologies for the clinical research enterprise across academic medical centers. Clin Transl Sci. 2012;5(3):281-4.
8. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc. 2013;20(1):117-21.
9. Ancker JS, Shih S, Singh MP, Snyder A, Edwards A, Kaushal R, et al. Root causes underlying challenges to secondary use of data. AMIA Annu Symp Proc. 2011;2011:57-62.
10 Capurro MD, Yetisgen PhD M, van Eaton Md E, Black R, Tarczy-Hornoch MD. Availability of Structured and Unstructured Clinical Data for Comparative Effectiveness Research and Quality Improvement: A Multi-Site Assessment. eGEMs (Generating Evidence & Methods to improve patient outcomes). 2014;2(1):11.
11 Van Eaton EG, Devlin AB, Devine EB, Flum DR, Tarczy-Hornoch P. Achieving and Sustaining Automated Health Data Linkages for Learning Systems: Barriers and Solutions. eGEMs (Generating Evidence & Methods to improve patient outcomes). 2014;2(2):3
12. Murphy SN, Mendis M, Hackett K, Kuttan R, Pan W, Phillips LC, et al. Architecture of the open-source clinical research chart from Informatics for Integrating Biology and the Bedside. AMIA Annu Symp Proc. 2007:548-52.
13. Patton MQ. Qualitative research and evaluation methods. 3 ed. Thousand Oaks, Calif.: Sage Publications; 2002.