TUTORIAL

Access to Routinely Collected Clinical Data for Research: A Process Implemented at an Academic Medical Center

Susan C. Guerrero1,2, Sujatha Sridhar2,3, Cynthia Edmonds3, Christina F. Solis4, Jiajie Zhang1,2, David D. McPherson2,5 and Elmer V. Bernstam1,2,6,*

Electronic health records are valuable for clinical and translational research. Institutions must protect patient privacy and comply with applicable regulations while allowing appropriate access to clinical data for research. The processes that investigators must follow to access clinical data can be substantially different at different institutions. In this paper, we describe the process developed at our institution that has been active for 5 years and was used to satisfy over 200 requests for access to identified clinical data, usually within 1 day for internal requests and 3 days for visiting researchers.

Routinely collected clinical data are increasingly collected in electronic form via electronic health records (EHRs) and aggregated within data warehouses, sometimes with additional data, including administrative (e.g., demographics, billing codes) and financial data (e.g., amount paid to the provider). Routinely collected clinical data are distinct from clinical trial data. Clinical trial data are generally collected from consented participants as part of a protocol that is specifically approved by an institutional review board (IRB). In contrast, routinely collected clinical data are essential to health care and, thus, do not require additional resources to create. As such, these data have advantages and disadvantages for clinical and translational research (CTR). Advantages of routinely collected data include the fact that these data represent the population, rather than the often very selected subpopulation that enrolls in clinical trials.1,2 On the other hand, drawing valid conclusions from these data can be challenging.3,4 Nevertheless, there is increasing interest to leverage routinely collected clinical data for CTR.5 Many institutions formed informatics groups that, with clinical and institutional information technology (IT) groups, created and manage data warehouses. Each institution must implement policies that govern research access to clinical data warehouses (CDWs). These policies must ensure compliance with applicable regulations (e.g., Health Insurance Portability and Accountability Act (HIPAA)) and protecting patient privacy while facilitating CTR.6

One common approach is to create deidentified data sets for research. This has multiple advantages, including decreasing, although not eliminating, risk to privacy (Table 1). On the other hand, deidentification can reduce the utility of data,7 necessitating project-specific deidentification strategies. Free text documents, such as dictated clinical notes, may be particularly difficult to deidentify.8 Thus, from the researcher’s perspective, access to fully identified data can be very valuable.

To a degree, applicable regulations (e.g., HIPAA) are subject to interpretation. Thus, a particular practice may be considered acceptable by one institution but not acceptable by another. Furthermore, institutions differ with respect to their assessment of and willingness to accept legal risk. Therefore, policies regarding access to clinical data containing protected health information vary among institutions. Motivated by the desire to decrease the barriers to clinical and translational research, we describe policies and procedures for data access at our institution.

PROCESS IMPLEMENTED AT THE UNIVERSITY OF TEXAS HEALTH SCIENCE CENTER AT HOUSTON

The University of Texas Health Science Center at Houston (UTHSC-H) has maintained a CDW since 2006. Clinical data are updated nightly, and the CDW currently contains information on over 4.3 million individuals derived from multiple sources, including administrative/billing systems and institutional EHRs dating back to 2004. In addition to structured data, the CDW contains over 30 million notes, updated on an as-needed basis.

In 2013, an “umbrella protocol” entitled “Clinical and administrative data reuse for research and quality improvement” was submitted to the Committee for the Protection of Human Subjects (CPHS), the IRB. This protocol (Data S1) was deemed exempt by the CPHS due to the use of existing data and received a waiver of authorization for HIPAA purposes. Although usage tracking and access logs must be maintained, they do not have to be reported on a regular

1School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, Texas, USA; 2Center for Clinical and Translational Sciences, The University of Texas Health Science Center at Houston, Houston, Texas, USA; 3Research Administration, The University of Texas Health Science Center at Houston, Houston, Texas, USA; 4Office of Legal Affairs, The University of Texas Health Science Center at Houston, Houston, Texas, USA; 5Division of Cardiology and Department of Internal Medicine, McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, Texas, USA; 6Division of General Internal Medicine, Department of Internal Medicine, McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, Texas, USA.

*Correspondence: Elmer V. Bernstam (Elmer.V.Bernstam@uth.tmc.edu)
Received: October 15, 2018; accepted: December 16, 2018. doi:10.1111/cts.12614
Clinical Data Access Process

Guerrero et al.

basis to the CPHS. Change requests, such as inclusion of additional data in the CDW, still require CPHS approval.

Briefly, the protocol stipulates that no project-specific CPHS (IRB) approval is required as long as the following criteria are satisfied:

1. Data do not leave the servers maintained by the Biomedical Informatics Group (i.e., the group responsible for maintaining the CDW)
2. No identifiable data are shared with individual researchers or published, and
3. No contact is made with patients (e.g., to collect additional data)

Table 1 Identified vs. deidentified data

|                  | Deidentified data                                                                 | Identified data                                                                 |
|------------------|-----------------------------------------------------------------------------------|---------------------------------------------------------------------------------|
| Volume           | Very large data sets aggregated from multiple institutions (> 100 million patients) | Smaller data sets, usually limited to a particular study or institution (usually < 10 million patients) |
| Analysis         | Deidentification can affect analysis results                                       | Not affected by deidentification                                                |
| Risk (required protections) | Generally less sensitive                                                            | Highly sensitive with large penalties for inappropriate release                |
| Patient follow-up | Impossible unless data are reidentifiable                                          | Possible                                                                       |
| Text (unstructured data) | Difficult to deidentify and thus relatively difficult to share                     | Can be included                                                                |
| Rare data (e.g., rare diseases, unusual laboratory results) | By definition highly identifying and thus usually censored                         | Can be included                                                                |
| Genetic/genomic data | Nucleic acid sequence is difficult or impossible to deidentify                    | Can be included                                                                |

The protocol is managed by an administrator according to the flowchart shown in Figure 1. Key stakeholders and activities are shown in Figure 2. Requests for access to clinical data are handled using a local instance of the Research Electronic Data Capture (REDCap) system. The administrator is responsible for maintaining records, including documentation that the requesting investigator has completed the IT, human subjects’ research, and HIPAA training required of all investigators at the institution who work with clinical data. In cases that are not covered by the umbrella protocol, the administrator must verify that the investigator has been granted (or not granted) permission to access data and what specific data can be accessed by that investigator for the specific CPHS-approved project.

Figure 1 Data access request process. REDCap, Research Electronic Data Capture; SBMI, School of Biomedical Informatics.
Figure 2: Key stakeholders. REDCap, Research Electronic Data Capture.

In addition to granting access, the administrator is also responsible for revoking access to clinical data. User access is automatically revoked when a user leaves the institution (i.e., their IT account is terminated and their CDW credentials are invalidated). Additionally, user access is reviewed quarterly to determine if CDW access should be revoked due to expired training credentials, research protocol expiration (in cases in which access to clinical data was not gained via the umbrella protocol), or employee transfers (e.g., leaving the institution).

DATA ACCESS PROCESS UTILIZATION AND AUDIT

Since 2013, 233 investigators have obtained access to clinical data. At the present time (September 2018), 83 investigators have active data access privileges under the umbrella protocol. Usual time to from investigator request to obtaining access to clinical data has been 1 business day for internal requests and up to 3 business days for visiting researchers.

Many journals require a statement confirming IRB approval to be included in publications; this is enabled by the umbrella protocol. Authors write: “This study has been approved by the Committee for the Protection of Human Subjects (the UTHSC-H IRB) under protocol HSC-SBMI-13-0549.” Investigators whose projects do not conform to the umbrella protocol (e.g., if they wish to move data from institutional servers) may submit their projects to the CPHS for project-specific review.

Research productivity and attribution have been difficult to track. In general, projects have included:

1. Study feasibility (e.g., in preparation for a grant submission)
2. Quality improvement efforts (e.g., how many patients with X have had a prescription for drug Y?, clinical quality metrics)
3. Support for clinical operations (e.g., inform resource allocation decisions)
4. Basic and applied informatics studies (e.g., implement and evaluate an algorithm)
5. Participation in data-driven research networks (e.g., PCORNet10)
6. With separate CPHS approval, identifying patients who are potentially eligible for specific clinical trials, and
7. Population health management (e.g., identify patients with poorly controlled diabetes and implement intervention).

An unanticipated side effect of streamlining data access and removing the need for project-specific IRB review has been the lack of a centralized protocol repository in the electronic IRB system. Thus, the actual number of publications that resulted from access to clinical data is even more difficult to track.

Since initial review by the CPHS, the protocol has been amended several times to include new data sources, including data from an affiliated psychiatric hospital (i.e., a distinct institution affiliated with the UTHSC-H) and several de-identified locally managed commercial data sets. Access to some data sets requires additional approvals (e.g., the psychiatric hospital must approve access to their data). Access to data by trainees requires approval by a supervising faculty member.

The data access process has enabled cross-institutional collaboration. When a collaborator needs to access clinical data, they do so by obtaining a guest account from institutional IT and demonstrating completion of required HIPAA and human subjects’ research training. If collaborators need to move data to their home institution, they are no longer covered by the umbrella protocol and must obtain project-specific approval from the CPHS.

Although quantitative comparison across institutions is challenging, discussions with collaborators suggests that the process described above is significantly faster and easier than at other institutions. Anecdotally, the data access process has been very well-accepted by the local investigator community and collaborators.

In 2018, the data access process was subjected to a formal audit. Although no fundamental issues were found, deficiencies in record keeping were identified based on random chart review. To address these deficiencies, we transitioned from a paper-based system to using REDCap for request processing. In addition, we transitioned process administration to a single administrative professional. These findings emphasize the need for a formal process.

DISCUSSION, CHALLENGES, AND LESSONS LEARNED

We developed and implemented a data access process for investigators who wish to reuse routinely collected, identified clinical data for research. The process has been deemed exempt by our IRB, removing the need for project-specific IRB review as well as annual reporting and renewal requirements for projects that are limited to retrospective
review of existing clinical data. The project also received a waiver of HIPAA authorization from the CPHS. Over 200 individual researchers have used this process. Average time to access data has been reduced to 1 business day for users from our institution and 3 business days for visiting researchers.

Traditionally, IRBs require a project-specific protocol. In the protocol document, the investigator must specify the data to be abstracted, the analyses to be run, and the results to be reported. However, these requirements are not always compatible with the “big data science” approach. Often, projects are exploratory (e.g., what correlates with what?) and use a variety of algorithms that are difficult to specify in advance.

For example, consider the problem of designing a “high-throughput phenotyping” algorithm to identify patients with an undiagnosed rare disease, such as systemic lupus erythematosus (SLE) within a large clinical data set. It is important to diagnose SLE, but diagnosis is often delayed by years due in part to the wide variety of symptoms that may be associated with the disease and the lack of signs or symptoms that would reliably rule in (or rule out) SLE. To identify patients with undiagnosed SLE, one might attempt to apply machine learning. Specifically, one might design a classification algorithm to identify likely cases based on all data in the EHR. After running an initial version of the algorithm, one might conduct an error analysis by manually reviewing selected cases that were misclassified, modify the algorithm, and repeat.

Thus, it may not be possible for the investigator to pre-specify the data fields that the algorithm will use to classify cases, how many cases will be manually reviewed, or which fields will be accessed. It is difficult to determine how many investigators, particularly junior investigators, are deterred by the apparent contradiction between institutional data access policies and the exploratory nature of their data analysis plan. Adoption of umbrella protocols for data access, along with appropriate investigator support and guidance by the IRB and privacy office, can decrease barriers to the reuse of clinical data for research.

Furthermore, IRBs and institutions increasingly recognize the potential liability posed by large-scale clinical data collection required for popular “big data” approaches. Indeed, there is now an insurance industry that provides financial coverage to institutions in case of data breach. Approaches that limit the distribution of identified clinical data (i.e., protected health information) are desirable and may even be cost-saving if they can reduce insurance premiums. The protocol described in this paper specifically excludes projects that wish to transfer data from the institutional servers (e.g., to laptops or noninstitutional computers). Thus, investigators have an additional incentive to maintain clinical data in a safe computational environment provided by the institution.

The requirement to maintain the data on institutional servers does have implications for investigators. For example, analyses must be conducted using software on the servers rather than being downloaded to personal computers where users may have administrative privileges that allow them to install software or modify system configuration. To address this issue, we have dedicated research server capacity (i.e., physical servers and/or virtual machines) within the protected zone where investigators can install software that may be needed for specialized analyses.

We have adopted the perspective that vetted investigators can be trusted, rather than attempting to restrict user capabilities to prevent breach. One reason for this choice is that it is very difficult to prevent a malicious user from circumventing technical safeguards. For example, by taking a photograph of a screen using a personal smart phone or simply writing down information on paper. Thus, we have chosen to rely on careful user vetting rather than technical restrictions.

CONCLUSION

The purpose of this paper is to describe the clinical data access process that is in place at our institution. This process has been operational for over 5 years (i.e., has passed the test of time) and has withstood formal audit. Compared with traditional project-by-project IRB review, this process decreases the burden on investigators wishing to reuse clinical data for research and quality improvement while appropriately protecting patient privacy and complying with relevant regulations. We hope that this example can help other institutions lower barriers to research by developing appropriately compliant but research-friendly policies and procedures.

Supporting Information. Supplementary information accompanies this paper on the Clinical and Translational Science website (www.cts-journal.com).

Data S1. Clinical and administrative data reuse for research protocol.

Funding. This study was supported in part by the National Center for Advanced Translational Sciences (NCATS) grants UL1 TR000371 (Center for Clinical and Translational Sciences), UL1 TR001105 (UT-Southwestern Clinical and Translational Alliance for Research), UL1 TR001857 and U01 TR002393, the National Library of Medicine grant R01 LM011829, PCORI CDRN-1306-04608, the Reynolds and Reynolds Professorship in Clinical Informatics, and the Cancer Prevention Research Institute of Texas (CPRIT) Data Science and Informatics Core for Cancer Research (RP170668).

Conflict of Interest. The authors declared no competing interests for this work.

1. Murthy, V.H., Krumholz, H.M. & Gross, C.P. Participation in cancer clinical trials: race-, sex-, and age-based disparities. JAMA 291, 2720–2726 (2004).
2. Oh, S.S. et al. Diversity in clinical and biomedical research: a promise yet to be fulfilled. PLoS Med. 12, e1001918 (2015).
3. Diaz-Garelli, J.F., Bernstam, E.V., Mse Rahbar, M.H. & Johnson, T. Rediscovering drug side effects: the impact of analytical assumptions on the detection of associations in EHR data. AMIA Jt. Summits Transl. Sci. Proc. 2015, 51–55 (2015).
4. Hersh, W.R. et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med. Care 51 (8 suppl. 3), S30–S37 (2013).
5. Safran, C. Reuse of clinical data. Yearbook Med. Inform. 9, 52–54 (2014).
6. Ness, R.B., Joint Policy Committee, Societies of Epidemiology. Influence of the HIPAA privacy rule on health research. JAMA 298, 2164–2170 (2007).
7. Gal, T.S., Tucker, T.C., Gangopadhyay, A. & Chen, Z. A data recipient centered de-identification method to retain statistical attributes. J. Biomed. Inform. 50, 32–45 (2014).
8. Stubbs, A., Kofilia, C. & Uzuner, O. Automated systems for the de-identification of longitudinal clinical narratives: overview of 2014 i2b2/UTHealth shared task track 1. J. Biomed. Inform. 58 (suppl.), S11–S19 (2015).
9. Harris, P.A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N. & Conde, J.G. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J. Biomed. Inform.* **42**, 377–381 (2009).

10. Collins, F.S., Hudson, K.L., Briggs, J.P. & Lauer, M.S. PCORnet: turning a dream into reality. *J. Am. Med. Inform. Assoc.* **21**, 576–577 (2014).

11. Doria, A. *et al.* SLE diagnosis and treatment: when early is early. *Autoimmun. Rev.* **10**, 55–60 (2010).

12. Brown, S. Ten things you need to know about cybersecurity insurance. *Data Center J.* (2016); [http://www.datacenterjournal.com/ten-things-need-know-cybersecurity-insurance/](http://www.datacenterjournal.com/ten-things-need-know-cybersecurity-insurance/). Accessed October 11, 2018.

© 2019 The Authors. *Clinical and Translational Science* published by Wiley Periodicals, Inc. on behalf of the American Society for Clinical Pharmacology and Therapeutics. This is an open access article under the terms of the Creative Commons Attribution-Non-Commercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.