Journal of the American Medical Informatics Association, 29(4), 2022, 707–712
https://doi.org/10.1093/jamia/ocab259
Advance Access Publication Date: 6 December 2021
Case Report

Case Report

Supporting research, protecting data: one institution's approach to clinical data warehouse governance

Kellie M. Walters[1], **Anna Jojic**[1], **Emily R. Pfaff**[2], **Marie Rape**[1], **Donald C. Spencer**[3], **Nicholas J. Shaheen**[4], **Brent Lamm**[3], **and Timothy S. Carey**[5]

[1]North Carolina Translational and Clinical Sciences Institute, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA, [2]Department of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA, [3]Information Services Division, UNC Health, Morrisville, North Carolina, USA, [4]Division of Gastroenterology and Hepatology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA, and [5]Division of General Medicine and Clinical Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

Corresponding Author: Kellie M. Walters, MPH, NC TraCS Institute, 160 N Medical Drive, Chapel Hill, NC 27599, USA; kellie_walters@med.unc.edu

Received 9 July 2021; Revised 21 September 2021; Editorial Decision 7 November 2021; Accepted 11 November 2021

ABSTRACT

Institutions must decide how to manage the use of clinical data to support research while ensuring appropriate protections are in place. Questions about data use and sharing often go beyond what the Health Insurance Portability and Accountability Act of 1996 (HIPAA) considers. In this article, we describe our institution's governance model and approach. Common questions we consider include (1) Is a request limited to the minimum data necessary to carry the research forward? (2) What plans are there for sharing data externally?, and (3) What impact will the proposed use of data have on patients and the institution? In 2020, 302 of the 319 requests reviewed were approved. The majority of requests were approved in less than 2 weeks, with few or no stipulations. For the remaining requests, the governance committee works with researchers to find solutions to meet their needs while also addressing our collective goal of protecting patients.

Key words: EHR data, data governance, clinical data warehouse, clinical informatics, data privacy

INTRODUCTION

Data collected as part of usual clinical care are a powerful resource for research.[1,2] Researchers can leverage these data to find patients potentially eligible for a trial,[2–4] conduct secondary data analyses,[5,6] or follow clinical outcomes of study participants.[5,7] The Health Insurance Portability and Accountability Act of 1996 (HIPAA) allows for such uses of these data, provided patients sign a HIPAA authorization form or an Institutional Review Board (IRB) grants a waiver of HIPAA authorization.[8] While this ensures the appropriate legal protections are in place, concerns about data release and use do not end there.

Data brokers and clinical leaders face the following questions:

- How much, and what type, of data is appropriate to share with an external party?
- How will patients react to a study recruitment letter related to their medical history, as identified from the electronic health record?[9]
- Does disclosure of these data present a risk to institutional reputation?

Different stakeholders have different reactions to these questions, and data brokers must carefully balance the benefits of data use with potential drawbacks.[10] A recent systematic review on data access and use in clinical data warehouses found a lack of in-depth information on data governance and criteria used for reviewing requests.[11] This article seeks to fill that void by describing the governance approach at the University of North Carolina at Chapel Hill for research uses of the Carolina Data Warehouse for Health (CDW-H), the central repository for electronic health record (EHR) data for UNC Health.

Institutional context

The University of North Carolina at Chapel Hill and UNC Health work in close partnership to carry out their combined research mission. UNC-Chapel Hill, a large Research 1 University, is home to schools of medicine, public health, pharmacy, nursing, and dentistry. UNC Health encompasses the academic medical center in Chapel Hill and community practices and hospitals throughout the state of North Carolina.

The CDW-H, UNC Health's institutional EHR data warehouse, was established in 2009. The CDW-H is used for operational, quality improvement, and research activities. This article addresses the research component, which is jointly governed by the UNC School of Medicine and UNC Health through the CDW-H Oversight and Operations Committees. The North Carolina Translational and Clinical Sciences Institute (NC TraCS), UNC's Clinical and Translational Science Award (CTSA) hub, housed in the School of Medicine, is charged with the stewardship of the data request and approval process.

MATERIALS AND METHODS

Data request process

In order to ensure appropriate data protections and compliance with study protocols, investigators are not permitted to directly access the CDW-H. Instead, the CDW-H research program works on a request model, with NC TraCS serving as the single point of entry for all research requests. All requests go through a standard intake and review process (see Figure 1).

The data request form (see Supplementary Appendix A) includes questions about inclusion/exclusion criteria, data elements requested, planned use of the data, and data sharing plans (if any). The structured nature of the data request form ensures that there is adequate information to define the scope of the project and conduct governance review.

Requests are fulfilled by CDW-H honest brokers, designated analysts within NC TraCS or approved departments who are trained in querying healthcare data and navigating regulatory issues. The honest broker is not a member of the study team and does not participate in the research or analyses.

CDW-H supports requests for fully identified, HIPAA limited, and deidentified EHR data. Requests fall into 3 categories:

- Study recruitment: With appropriate approvals, study teams can request datasets of patients that potentially meet their study criteria and may receive approval to contact these patients via mail, phone, patient portal, or in clinic.
- Longitudinal EHR data for patients enrolled in trials, cohort studies, or registries: After a patient is enrolled in a study, information about that patient's medical history and future clinical outcomes may be requested, supplementing the data collected by the study team.

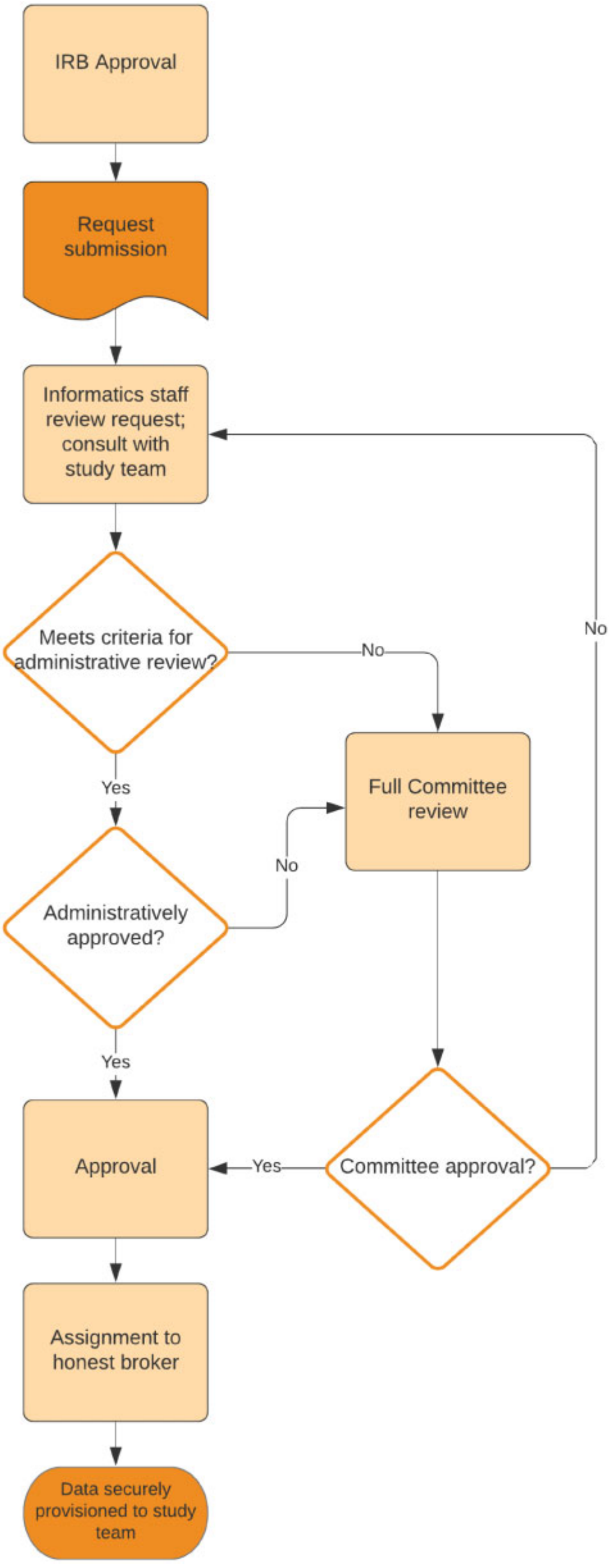
IRB Approval
Request submission
Informatics staff review request; consult with study team
Meets criteria for administrative review?
No
Yes
Full Committee review
No
Administratively approved?
Yes
Approval
Yes
Committee approval?
No
Assignment to honest broker
Data securely provisioned to study team

Figure 1. Carolina Data Warehouse for Health request process. All research requests for data go through this intake and approval process. Of note is the dual review pathway. Most requests are approved through administrative (expedited) review, which includes a step to ensure the request aligns with the IRB protocol. Requests that are controversial or represent additional risk, such as large data sharing projects or requests that involve recruitment of children, are reviewed by the CDW-H governance committees.

- Secondary data analyses: With waivers of both consent and HIPAA authorization in place, researchers may request datasets without patient contact. Such studies allow investigators to analyze trends and outcomes in real-world data.

Governance structure

CDW-H governance is comprised of the CDW-H Oversight and Operations Committees. The CDW-H Operations Committee reviews all data requests through an administrative pathway or full committee meeting. The CDW-H Oversight Committee sets policy for the CDW-H and reviews precedent-setting requests and appeals.

Both Committees benefit from interdisciplinary membership rosters that include clinician scientists, Office of Human Research Ethics (IRB) leadership, informatics researchers, public health researchers, legal counsel, privacy office staff, and patient representatives. Members include representatives from both the University and Health System.

An administrative review process was developed in response to an increased volume in requests and recognition that many requests did not require extensive discussion. In this expedited review, a staff member and the CDW-H Operations Chair assess whether the request meets HIPAA and IRB requirements. Examples of requests eligible for administrative review include many datasets for secondary analyses if data will remain within UNC, recruitment lists comprised of adults, and data regarding patients who have consented to participate in the associated study.

The CDW-H review process is not intended to replace or circumvent the IRB process. Rather, the reviews complement one another and often evaluate different issues.

Governance approach

The primary goal of CDW-H governance is to determine how best to safely use clinical data for research to ultimately improve patient health. The challenge is that it can be difficult to define what uses are appropriate. While HIPAA provides us with the legal guardrails we must operate within, we have learned that questions about data disclosure and sharing often go beyond what HIPAA considers.[10] UNC's interdisciplinary governance committees exist to address this exact challenge.

All requests must receive appropriate IRB review. The CDW-H review process includes a step to ensure the request and protocol align; and, if not, the discrepancy must be addressed before the request can be approved. Beyond IRB and HIPAA requirements, governance often considers the following:

- Is the request limited to the minimum necessary data?

As machine learning and analyses of large cohorts become more common, researchers are asking for more and more data—more patients, more years of data, or more data elements for each patient.[12,13] However, HIPAA requires only the minimum necessary information to complete a task to be disclosed[8]; therefore, the Committee may require such requests be narrowed in scope, or the scope of the request be well justified (by, for example, a consultation with a statistician).

- Are data sharing plans justified and in compliance with legal requirements?

Projects requiring the sharing of data outside the institution are becoming more common.[14–17] This represents additional inadvertent disclosure risk, so the Committee must weigh the value of the data sharing with that risk. Key considerations include what data will be shared, what plans there are for data reuse, how shared data will be stored, and whether a data sharing agreement will be in place.

When sharing data, a study team may also request approval for data linkage. Linking CDW-H data with, for example, claims data has the benefit of filling in gaps in EHR data, reducing missing data bias.[18,19] The Committee considers how data will be linked, what identifiers (if any) need to be shared to support the linkage, and whether the linked data itself presents risk of reidentification.

- What impact could this request have on patients?

Ever present on the Committee's mind is the impact a request may have on patients. For example, CDW-H data is frequently used to sup-

Table 1. Common request scenarios and examples of the Committee's past responses

Scenario	Example responses
Sharing fully identified dataset with outside institution	• Ask research team to justify the requested data (eg, "What analytical purpose does exact street address serve?") • Suggest alternative variables to achieve a similar goal (eg, providing census tract instead of full address) • Evaluate options to avoid release of identifiers unless necessary (eg, date shifting where exact dates are not required)
Cohort size or control group size appears excessively large (eg, 100 controls for each case)	• Request justification for cohort or control group size • Recommend (or require) a consult with CTSA biostatistics service • Add inclusion criteria when appropriate (eg, only include patients with at least 3 encounters in the study period)
Cohort definition for a recruitment dataset is very broad, while recruitment goal is small (eg, a list of 500 000 patients in order to recruit 25 participants)	• Educate researcher that CDW-H is more appropriately used to recruit more narrowly defined populations • Recommend consult with CTSA recruitment service
Cohort definition targets sensitive recruitment population (eg, teenagers with suicidal ideation)	• Recruitment criteria may be narrowed by, for example, requiring a specified diagnosis code to appear multiple times on a patient's record, rather than once, or requiring chart review after receipt of dataset but prior to patient contact • Amendments to recruitment materials may be required to ensure language is benign and unlikely to cause distress
Request to link data with an external dataset, such as claims data or EHR data from another institution	• Recommend a linkage methodology that does not require sharing identifiers (ie, privacy preserving record linkage)[22,23]

port recruitment, and the Committee recognizes that patients may have concerns if they receive a recruitment letter based on information in their medical records.[9] UNC created template recruitment language to help address this concern.[20] The Committee also advises on methods to prevent negative impact on patients—for example, by requesting that recruitment lists of pregnant people exclude patients with diagnosis or procedure codes signifying miscarriage.

- What impact could this request have on our institution?

The Committee considers how a request may reflect on UNC. A common concern when sharing EHR data is the possibility that data could be misused for competitive purposes.[16,21] Consider a project that allows researchers to access data from multiple health systems; the pooled data could enable comparing rates of post-surgical complications among competing institutions. The Committee may require that UNC's name not be disclosed in the combined dataset.

The above is not an exhaustive list but does represent many of the most common issues the Committee considers. The Committee judges each case individually with the goal of finding common ground with the study team in an attempt to "get to yes." If immediate approval is not possible, the Committee may respond by coaching the study team to make modifications in order to meet the research needs of their project, while also meeting the goal of protecting patients. Table 1 outlines common scenarios the Committee has faced and examples of their responses.

RESULTS

We aim for a governance process that is compliant, efficient, and supportive of research. In practice, this means that data requests: (1) are reviewed in a timely manner, (2) align with the corresponding IRB protocol, and (3) comply with institutional policy and federal law. To measure success, we reviewed the outcomes of data requests received in 2020, as shown in Table 2.

Of the 319 requests reviewed, 302 (94.7%) were approved. Most requests (265, 83.1%) were reviewed via administrative re-

Table 2. CDWH governance outcomes

Summary of requests	Requests (#)	Requests (%)
Data requests reviewed	319	100%
Data requests approved	302	94.7%
Requests reviewed via administrative pathway	265	83.1%
Requests reviewed via full committee pathway	54	16.9%
Administrative pathway time from submission to approval[a]	Admin. reviewed requests (#)	Admin. reviewed requests (%)
0–2 days	98	37.0%
3–14 days	94	35.5%
15–28 days	34	12.8%
29 days and over	29	10.9%
Not approved[b]	10	3.8%
Administrative pathway action	Admin. reviewed requests (#)	Admin. reviewed requests (%)
Requests may receive stipulations in multiple categories		
Approve without stipulations	202	76.2%
Regulatory-related changes required (eg, IRB protocol modification)	57	21.5%
Other (eg, clarification needed to understand if data will be shared)	8	3.0%
Committee pathway time from submission to approval[a]	Comm. reviewed requests (#)	Comm. reviewed requests (%)
0–28 days	14	25.9%
29–60 days	14	25.9%
61 days and over	19	35.2%
Not approved[b]	7	13.0%
Committee pathway action	Comm. reviewed requests (#)	Comm. reviewed requests (%)
Requests may receive stipulations in multiple categories		
Approve without stipulations	14	25.9%
Regulatory-related changes required (eg, IRB protocol modification)	14	25.9%
Data sharing agreement required and/or modification to data sharing plans required	26	48.1%
Scope modification or justification required	4	7.4%
Other (eg, add inclusion criteria to increase specificity; add criteria to ensure appropriate guardian for child is contacted)	2	3.7%

[a] Request approval may be dependent on factors outside of CDW-H governance's control. For example, approval may depend on the study team making an IRB protocol modification.

[b] In some cases, a request may be reviewed and receive stipulations, but the study team will choose not to proceed with the required changes. These requests are marked as not approved.

view. Seventy-two percent (192) of these requests received approval in less than 2 weeks. Fifty-seven requests (21.5%) required modification to the data request, IRB protocol, or both in order to ensure alignment across the documents and be compliant.

Fifty-four requests (16.9%) were reviewed by the CDW-H Operations Committee. About half of these requests received approval within 2 months. The committee process is lengthier, because requests tend to be complex and require multiple consultations. Most requests receive stipulations, which must be addressed before approval. Common stipulations include modifications to ensure request and IRB protocol alignment or execution of a data sharing agreement. In rare instances, the CDW-H Operations Committee escalates requests to the CDW-H Oversight Committee. Three requests from 2020 were escalated and ultimately approved.

DISCUSSION

Data governance practices will necessarily vary by institution. Our governance approach has commonalities with the spectrum of criteria and procedures described in Pavlenko et al[11] including requirements for human subjects protection training, IRB approval, and data sharing agreements, and attention to patients' perspectives and institutional reputation when reviewing requests. Notably, however, the spectrum in Pavlenko does not include an explicit focus on whether a request is limited to the minimum necessary data. As interest and capacity for working with larger datasets grow, we expect more institutions will need to address this issue.

The success and longevity of CDW-H governance is attributable to a combination of factors. The bedrock of CDW-H governance is strong institutional support from UNC Health and UNC-Chapel Hill and close relationships among stakeholders within both organizations. Importantly, CDW-H governance's purview is narrow, limited to the review of requests for CDW-H data. The Committee intentionally avoids commenting on the protocol or scientific merit of a project. Having a single point of entry helps to reduce confusion and ensures requests are reviewed consistently.

The CDW-H governance process and committees have proven capable of adapting, and this will be critical to continued success. Trends in clinical informatics, including increased interest in data sharing, a growing appetite for larger analytical datasets, and heightened interest in analyzing clinical notes, will present new challenges for our governance system as we seek to meet the needs of researchers while also protecting patients and data.

CONCLUSION

Our governance process has proven effective and efficient for UNC over the past decade. The Committees are a valuable resource for the University and Health System. They help ensure clinical data are provisioned appropriately and researchers are educated about the benefits and sensitivities of working with clinical data. Though many data governance challenges lie ahead, our past experience demonstrates that this system is a robust one that is able to address a dynamic clinical research environment.

FUNDING

This work was supported by the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health, through Grant Award Number UL1TR002489. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

AUTHOR CONTRIBUTIONS

Manuscript drafting: KMW, AJ, ERP, and TSC. Initial implementation of governance processes: TSC, BL, DCS, and MR. Ongoing leadership of governance processes: AJ, BL, ERP, NJS, DCS, MR, and KMW. Manuscript revisions and final approval: TSC, AJ, BL, ERP, NJS, DCS, MR, and KMW.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

We would like to express our gratitude to the current and former members for the CDW-H Oversight and Operations Committees.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABLITY

No new data were generated or analyzed in support of this research.

REFERENCES

1. Kim E, Rubinstein SM, Nead KT, *et al.* The evolving use of electronic health records (EHR) for research. *Semin Radiat Oncol* 2019; 29 (4): 354–61.
2. Cowie MR, Blomster JI, Curtis LH, *et al.* Electronic health records to facilitate clinical research. *Clin Res Cardiol* 2017; 106 (1): 1–9.
3. Jones WS, Mulder H, Wruck LM, *et al.*; ADAPTABLE Team. Comparative effectiveness of aspirin dosing in cardiovascular disease. *N Engl J Med* 2021; 384 (21): 1981–90.
4. Obeid JS, Beskow LM, Rape M, *et al.* A survey of practices for the use of electronic health records to support research recruitment. *J Clin Transl Sci* 2017; 1 (4): 246–52.
5. Arterburn D, Wellman R, Emiliano A, *et al.*; PCORnet Bariatric Study Collaborative. Comparative effectiveness and safety of bariatric procedures for weight loss: a pcornet cohort study. *Ann Intern Med* 2018; 169 (11): 741–50.
6. Safran C, Bloomrosen M, Hammond WE, *et al.* Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. *J Am Med Inform Assoc* 2007; 14: 1–9.
7. Richesson RL, Hammond WE, Nahm M, *et al.* Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. *J Am Med Inform Assoc* 2013; 20 (e2): e226–31–e231.
8. National Institutes of Health. HIPAA Privacy Rule and Its Impacts on Research. https://privacyruleandresearch.nih.gov/pr_08.asp Accessed May 25, 2021.
9. Beskow LM, Brelsford KM, Hammack CM. Patient perspectives on use of electronic health records for research recruitment. *BMC Med Res Methodol* 2019; 19 (1): 42.
10. Lee LM. Ethics and subsequent use of electronic health record data. *J Biomed Inform* 2017; 71: 143–6.
11. Pavlenko E, Strech D, Langhof H. Implementation of data access and use procedures in clinical data warehouses. A systematic review of literature and publicly available policies. *BMC Med Inform Decis Mak* 2020; 20 (1): 157.
12. Jiang F, Jiang Y, Zhi H, *et al.* Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol* 2017; 2 (4): 230–43.
13. Rajkomar A, Oren E, Chen K, *et al.* Scalable and accurate deep learning with electronic health records. *npj Digital Med* 2018;1:18.

14. Haendel MA, Chute CG, Bennett TD, *et al.*; N3C Consortium. The National COVID Cohort Collaborative (N3C): rationale, design, infrastructure, and deployment. *J Am Med Inform Assoc* 2021; 28 (3): 427–43.
15. Forrest CB, McTigue KM, Hernandez AF, *et al.* PCORnet® 2020: current state, accomplishments, and future directions. *J Clin Epidemiol* 2021; 129: 60–7.
16. Turley CB, Obeid J, Larsen R, *et al.* Leveraging a statewide clinical data warehouse to expand boundaries of the learning health system. *EGEMS (Wash DC)* 2016; 4 (1): 1245.
17. Hornik CP, Atz AM, Bendel C, *et al.*; on behalf of the Best Pharmaceuticals for Children Act–Pediatric Trials Network. Creation of a multicenter pediatric inpatient data repository derived from electronic health records. *Appl Clin Inform* 2019; 10 (02): 307–15.
18. McDermott CL, Engelberg RA, Woo C, *et al.* Novel data linkages to characterize palliative and end-of-life care: challenges and considerations. *J Pain Symptom Manage* 2019; 58 (5): 851–6.
19. Canterberry M, Kaul AF, Goel S, *et al.* The patient-centered outcomes research network antibiotics and childhood growth study: implementing patient data linkage. *Popul Health Manag* 2020; 23 (6): 438–44.
20. NC TraCS Institute. Carolina Data Warehouse for Health (CDW-H) Suggested Language to use in Recruitment Letters, Emails and Telephone Scripts. https://tracs.unc.edu/index.php/services/informatics-and-data-science/cdw-h/cdw-h-faq Accessed May 25, 2021.
21. Kuperman GJ, McGowan JJ. Potential unintended consequences of health information exchange. *J Gen Intern Med* 2013; 28 (12): 1663–6.
22. Bian J, Loiacono A, Sura A, *et al.* Implementing a hash-based privacy-preserving record linkage tool in the OneFlorida clinical research network. *JAMIA Open* 2019; 2 (4): 562–9.
23. Kho AN, Cashy JP, Jackson KL, *et al.* Design and implementation of a privacy preserving electronic health record linkage tool in Chicago. *J Am Med Inform Assoc* 2015; 22 (5): 1072–80.