Journal of the American Medical Informatics Association, 0(0), 2020, 1–8
doi: 10.1093/jamia/ocaa089
Research and Applications

Research and Applications

Understanding enterprise data warehouses to support clinical and translational research

Thomas R Campion Jr [1], Catherine K Craven,[2] David A Dorr,[3] and Boyd M Knosp[4]

[1]Department of Population Health Sciences, Weill Cornell Medicine, New York, New York, USA, [2]Institute for Health Care Delivery Science, Icahn School of Medicine at Mount Sinai, New York, New York, USA, [3]Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, Portland, Oregon, USA, and [4]Institute for Clinical and Translational Science, Roy J. and Lucille A. Carver College of Medicine, University of Iowa, Iowa City, Iowa, USA

Corresponding Author: Thomas R. Campion, Jr, PhD, Department of Population Health Sciences, Weill Cornell Medicine, Cornell University, 575 Lexington Avenue, Third Floor, New York, NY 10022, USA; thc2015@med.cornell.edu

Received 28 January 2020; Revised 24 April 2020; Editorial Decision 27 April 2020; Accepted 12 May 2020

ABSTRACT

Objective: Among National Institutes of Health Clinical and Translational Science Award (CTSA) hubs, adoption of electronic data warehouses for research (EDW4R) containing data from electronic health record systems is nearly ubiquitous. Although benefits of EDW4R include more effective, efficient support of scientists, little is known about how CTSA hubs have implemented EDW4R services. The goal of this qualitative study was to understand the ways in which CTSA hubs have operationalized EDW4R to support clinical and translational researchers.
Materials and Methods: After conducting semistructured interviews with informatics leaders from 20 CTSA hubs, we performed a directed content analysis of interview notes informed by naturalistic inquiry.
Results: We identified 12 themes: organization and data; oversight and governance; data access request process; data access modalities; data access for users with different skill sets; engagement, communication, and literacy; service management coordinated with enterprise information technology; service management coordinated within a CTSA hub; service management coordinated between informatics and biostatistics; funding approaches; performance metrics; and future trends and current technology challenges.
Discussion: This study is a step in developing an improved understanding and creating a common vocabulary about EDW4R operations across institutions. Findings indicate an opportunity for establishing best practices for EDW4R operations in academic medicine. Such guidance could reduce the costs associated with developing an EDW4R by establishing a clear roadmap and maturity path for institutions to follow.
Conclusions: CTSA hubs described varying approaches to EDW4R operations that may assist other institutions in better serving investigators with electronic patient data.

Key words: data warehouse, secondary use, CTSA, EHR

INTRODUCTION

Objective

Electronic patient data are critical to the conduct of clinical and translational science, and experts have noted that "an [enterprise] data warehouse (EDW) and related [informatics core] services for research purposes are no longer optional components for a robust translational research enterprise."[1] Among Clinical and Translational Science Award (CTSA) hubs funded by the National Institutes of Health (NIH) National Center for Advancing Translational Science (NCATS), adoption of EDWs containing electronic health

record (EHR) data for research has increased from 64% in 2008[2] to 94% in 2017.[3] Potential benefits of EDW for research (EDW4R) services include more effective and efficient clinical and translational research support through access to EHR data that have been curated by informatics experts for scientific purposes.[4] For example, EDW4R have enabled acceleration of drug development pipelines,[5] phenome-wide association studies when linked with genotyped biospecimens,[6] rapid pharmacovigilance,[7] characterization of treatment pathways for 250 million patients in 11 countries,[8] multisite clinical trial recruitment,[9] automated data collection,[10] and delivery of real-world evidence at the point of care to actualize the learning healthcare system.[11] Despite their ubiquity at CTSA hubs, little is known about implementation of EDW4R to optimize these benefits. The goal of this qualitative study was to understand the ways in which CTSA hubs have operationalized EDW4R infrastructure and services to support clinical and translational investigators.

BACKGROUND AND SIGNIFICANCE

To enable an EDW4R, a CTSA hub requires financial resources, technical skills, and intraorganizational collaborations. As reported in a recent survey of CTSA hubs, funding EDW4R activities requires different combinations of grants, institutional subsidy, and fee for service.[1] CTSA informatics leaders have consistently identified financial sustainability as a concern for EDW4R activities.[1,2]

With respect to technology, the literature contains several descriptions of EDW4R infrastructure at CTSA hubs.[12–21] From this literature comes a wide variety of functions, capabilities, and needs, including data integration, management, education, support, tooling, governance, optimization, and alignment across missions. A common factor is the commonality of EHR systems as a data source, which require sophisticated engineering approaches using structured query language (SQL) to extract, transform, and load data in accordance with particular tool specifications for analytical purposes.

To support the spectrum of clinical and translational science activities, researchers can access data from an EDW4R using a number of tools commonly but not uniformly implemented at CTSA hubs. For activities preparatory to research, i2b2 enables investigators to obtain de-identified counts of patients meeting study eligibility criteria documented in EHR data; no SQL programming is required by researchers.[22] To support multisite investigator-initiated clinical trials, NCATS launched the Accrual to Clinical Trials (ACT) network,[23] which enables researchers to obtain patient counts using i2b2 data from nearly all CTSA hubs connected via the SHRINE (Shared Health Research Information Network).[24] To support sponsor-initiated clinical trials, a private company, TriNetX (Cambridge, MA), enables a CTSA hub to make de-identified patient counts from EDW4R data available to biopharmaceutical companies.[25] For retrospective observational studies, the Observational Medical Outcomes Partnership common data model (CDM) has gained traction among data scientists who use standardized queries and advanced statistical techniques such as machine learning to analyze EHR and claims data.[26] The Observational Medical Outcomes Partnership also serves as the CDM for the NIH *All of Us* Research Program,[27] which aims to gather clinical and genomic data on more than 1 million patients. Additionally, the Patient-Centered Outcomes Institute developed the PCORnet CDM to support prospective clinical trial enrollment and retrospective observational studies across a nationwide network.[28,29] Although scientists can readily interact with EDW4R assets using these tools and CDMs, challenges include completeness, quality, and bias of the underlying EHR data.[30]

With respect to intraorganizational collaborations, a robust EDW4R can involve cooperation of enterprise information technology (IT) organizations and informatics faculty and staff from a CTSA hub.[31] Enterprise IT may consist of clinical IT, which oversees the EHR used in hospitals and practices, as well as operational IT, which provides baseline infrastructure (eg, network, email) across a medical center or university. Common features of enterprise IT organizations include service desk support and an underlying IT service management approach, such as the Information Technology Infrastructure Library,[32] often operationalized with IT workflow software such as ServiceNow (Santa Clara, CA). Whereas the focus of enterprise IT is technology, the purview of biomedical informatics is the structuring, acquisition, and use of patient data to support clinical and translational research using IT as a tool.[31] Although an enterprise IT organization often provides an EDW to address administrative, clinical, financial, and research activities in an organization, scholars distinguish it from an EDW4R that addresses only research activities.[33]

Given the financial, technical, and organizational complexity of EDW4R operations, optimizing an institution's approach is challenging. The purpose of this article is to provide insight as to current implementations, facilitators, and challenges in EDW4R services at a variety of CTSA hubs.

MATERIALS AND METHODS

The Informatics Domain Task Force (iDTF) of the CTSA consortium commissioned this study through the EDW Working Group co-led by 2 co-authors (B.M.K., T.R.C.) with support from 1 co-author (D.A.D.) as liaison from the NCATS National Center for Data to Health and 1 co-author (C.K.C.) as liaison from a separate iDTF-supported quantitative survey of EDW4R practices and research reproducibility. The University of Iowa Institutional Review Board (IRB) determined this study to be non–human subjects research.

Data collection

Two authors (B.M.K., T.R.C.) conducted semistructured interviews with informatics leaders responsible for EDW4R activities at CTSA hubs. Interviews occurred at the American Medical Informatics Association Informatics Summit as well as the iDTF/National Center for Data to Health Face-to-Face meeting held March 2019 in San Francisco along with teleconferences held through May 2019. All semistructured interviews followed a guide covering 3 areas—organizational and technical architecture, processes for access, and service management—developed by 2 authors (B.M.K., T.R.C.) based on their experience operating EDW4R activities in 2 CTSA hubs (Supplementary Appendix 1). While conducting interviews, interviewers recorded handwritten notes that they later transcribed into electronic format. Interviewers then shared notes with interviewees by email for correction and elaboration.

Data analysis

The study team performed a directed content analysis[34] of interview notes informed by naturalistic inquiry.[35] Two authors (B.M.K., T.R.C.) independently coded interview notes by identifying individual concepts and relationships between concepts. Additionally, the 2 authors independently generated memos that further elucidated

Table 1. Characteristics of interview respondents

Role	
Director	14 (70)
Chief research informatics officer	4 (20)
Technical lead	2 (10)
Time in role	
1-5 y	10 (50)
6-10 y	6 (30)
10+ y	4 (20)
Education (highest obtained)	
Nonterminal degree	10 (50)
PhD	5 (25)
MD	4 (20)
MD, PhD	1 (5)
Sex	
Male	17 (85)
Female	3 (15)

Values are n (%).

concepts and associations. The 2 authors then compared codes and memos, iterating analysis further into a single memo for review by 2 authors not involved in the interview process (D.A.D., C.K.C.), a process known as a peer debriefing.[35] After incorporating peer debriefing feedback into analysis, the study team engaged in member checking[35] by presenting preliminary findings to interviewees and others in EDW4R leadership roles as part of monthly EDW4R Working Group teleconferences. All members of the research team maintained reflexivity,[35] or awareness of how their biases could impact the research process.

RESULTS

Interviews with 20 respondents (Table 1) required approximately 25 hours to complete and yielded 35 pages of transcribed text.

The analysis of notes identified 12 themes: organization and data; oversight and governance; data access request process; data access modalities; data access for users with different skill sets; engagement, communication, and literacy; service management coordinated with enterprise IT; service management coordinated within a CTSA hub; service management coordinated between informatics and biostatistics; funding approaches; performance metrics; and future trends and current technology challenges.

Organization and data

The relationship between a CTSA hub and its health systems (ie, clinical enterprises) influenced EDW4R organization. CTSA informatics leaders characterized their relationships with clinical IT organizations variably with some noting a close partnership and others noting a "significant separation" or being treated as a "client." Approaches spanned the spectrum from CTSA hubs having no EDW4R to a dedicated EDW4R. Specifically, at one institution where no EDW4R existed, a health system–only EDW was in operation and provided limited research support. More frequently, hubs (n = 8) reported having a shared EDW between health system and research groups. Most commonly, hubs (n = 12) described a dedicated EDW4R containing data from a single health system that was "downstream and separate" from a health system's EDW. Additionally, 2 hubs described a dedicated EDW4R storing data from multiple health systems and/or statewide health information exchange efforts.

Oversight and governance

Oversight procedures varied from being nonexistent to a CTSA hub's informatics group creating and enacting policy to multistakeholder bodies defining activities for CTSA hub informatics teams to implement. To release data to investigators, the majority of institutions described requiring approval from both health system and university representatives including IRB, security, privacy, legal, and informatics. Several hubs had data governance committees of varying forms with some reviewing and setting policies for internal data requests and others vetting external data sharing agreements. At 2 hubs, only the EDW4R team controlled data access and governance.

Several institutions described the IRB as the EDW4R's controlling entity. IRB review and approval of study protocols was required for requests for identified data extraction, and the EDW4R team checked each IRB-approved protocol to ensure requested data matched terms for data release described in the protocol. Two institutions described the need for only IRB approval to release data due to IRB membership including informaticians. IRB participation was also critical in establishing guidelines for approving or denying requests involving massive data extractions (eg, "all EHR data") from an EDW4R. Specifically, IRBs helped determined the threshold (eg, number of patients) at which an investigator would require additional approval from health system and university overseers to obtain data.

Data access request process

To provide access to EDW4R resources, CTSA hubs described needing to verify a requester's IRB approval (n = 11) and Collaborative Institutional Training Initiative training status (n = 4), along with institutional affiliation via Active Directory or other identity management application (n = 3). Two institutions reported automating IRB protocol and Collaborative Institutional Training Initiative training verification, while 1 institution described a need to automate comparison of EHR elements described in a data request against permitted data elements defined in an IRB protocol.

Each CTSA informatics group generally had an IRB protocol governing its EDW4R. Some institutions reported requiring the addition of clinical researchers to the EDW4R IRB protocol to obtain access to data. Others described requiring clinical researchers to obtain separate IRB approval for investigating research questions using EDW4R data. In addition to verifying regulatory training and ethics approval, several institutions required investigators to sign a data sharing agreement acknowledging the sensitivity of EHR data and attesting to proper handling procedures. For delivering data, one institution described trusting investigators to store sensitive information on institutionally managed and encrypted hardware but lacking the ability to enforce policies. To manage Health Insurance Portability and Accountability Act–defined patient identifiers, most institutions had some form of honest broker activities performed by the EDW4R, although only 7 used the term *honest broker* in their descriptions.[36]

Data access modalities

Hubs described supporting a number of self-service tools for investigators to interact with electronic patient data including those designed for clinical data (eg, i2b2, TriNetX, Epic SlicerDicer), commercial business intelligence tools (eg, BusinessObjects, Tableau), and statistical software (eg, Rshiny), as well as direct database access for SQL queries and other custom tools. EDW4Rs contained data from health systems' Epic or Cerner EHR implementations as

well from biorepository, clinical trials management, electronic data capture, and legacy clinical systems, along with external data sources such as the Social Security Death Master File.

CTSA hubs reported adopting a number of CDMs to participate in multiple research networks, each associated with its own CDM, including ACT (n = 17), TriNetX (n = 11), PCORnet (n = 10), and Observational Health Data Sciences and Informatics (n = 7). The majority of institutions participated in 2 networks, while 2 indicated that they participate in all 4. Strategic drivers for CDM adoption and research network participation included research leadership (n = 2), CTSA hub informatics leadership (n = 3), and both research and informatics leadership (n = 2). Consolidation trends appeared with 3 institutions reporting a shift from i2b2 to TriNetX for local queries and 2 hubs describing a move from separate local SHRINE instances to the national ACT network.

Data access for users with different skill sets

EDW4R leaders described a range of investigators' technical abilities, from needing point-and-click self-service tools like i2b2, to analyst-mediated reports extracted from EDW4R, to expert SQL access for "power users" seeking to use Python, R, or other command line interfaces. CTSA informatics leaders indicated that most users sought self-service tools and analyst-mediated reports; few described researchers seeking direct SQL access, with the exception of some informaticians and computational biologists. However, one hub reported requiring research groups requesting large datasets to have a data-savvy team member to interact directly with a SQL database. One CTSA hub predicted a decline in power users in favor of analyst-mediated queries due to the effort required for noninformatics staff to learn the complexity of underlying data stored in an EDW4R.

Of institutions that supported power users, governance committees typically reviewed and approved investigator access to EDW4R SQL resources. Some hubs addressed the needs of SQL-writing power users by provisioning a separate database to ensure that errant queries executed by researchers did not consumer computational resources and interfere with operational work performed by EDW4R staff.

In addition to varying technical abilities of users, EDW4R leaders described a range of powers users' clinical data expertise. At one end were data scientists with computational skills but little-to-no understanding of the impact of biological and clinical processes on EHR data. These users often had expertise in particular methods, such as machine learning, and sought to apply the approaches to clinical problems. At the other end were clinicians and those with biomedical informatics training or experience who understood the nuance and vagary of EHR data.

Engagement, communication, and literacy

As one respondent indicated, "constant outreach" was critical to ensuring awareness of EDW4R activities among investigators and required engagement through multiple channels. To raise awareness of EDW4R services, institutions employed a number of different strategies across online, in-person, funding opportunity-related, and informal modalities (Table 2). About half of respondents reported using websites, newsletters, and listservs websites to communicate.

For in-person meetings, respondents described a variety of workshop formats, which included presentations depicting examples of existing questions from researchers and drop-in sessions with customized support. One institution described a user-training program whereby investigators watched online training videos prior to attending an in-person class in which investigators interacted with informatics experts to analyze data. Presentations during faculty meetings were often challenging, with one respondent stating, "Department meetings don't work [because] research gets squeezed into a few minutes."

Table 2. Awareness activities for EDW4R

Activity	n
Online	
Website	7
Video training materials	4
Newsletter	3
Documentation	2
Self-training environment	2
Listserv	1
In-person	
Faculty meeting presentations	5
Half to multiday workshops	5
Training for faculty/staff	4
One-on-one training/consulting session	3
Drop-in office hours	2
Orientation	2
Training studio	2
Didactic lectures (eg, master's program)	1
Information table at events	1
Lunch-and-learn sessions	1
Recruitment core engagement	1
Funding opportunity-related	
Pilot grant program	2
Abstract contest	1
Informal	
Physician champions	2
Word of mouth	2

EDW4R: electronic data warehouses for research.

Two institutions described funding opportunity-related outreach, including a contest with 2 cash prizes that required investigators to present an abstract proposing a novel study based on i2b2 queries. The contest increased investigator engagement with clinical data for research through increased i2b2 usage. Another institution described a local pilot grant program that provided unfunded researchers with 12 hours of informatics consultation service based on merits of a scientific proposal.

Hubs indicated that informal engagement through word of mouth and champions contributed to cultural acceptance of informatics for research. As one respondent indicated, informatics leaders at a CTSA hub were adept at describing how EDW4R resources supported specific studies performed by clinical investigators. However, the respondent aspired for clinical investigators, without assistance from CTSA informatics hub leaders, to tell success stories about how they used EDW4R services to support their research.

Several institutions sought to improve investigators' data literacy with respect to knowledge about clinical data, use of EHR data in research, and protection of patient privacy. One institution explicitly described a goal of moving from an approach of delivering data without substantial explanation—"here's your data, figure it out"—to establishing data literacy among researchers. In addition to providing training for all researchers, ensuring collaboration with biostatistics and encouraging researchers to consult with informatics services early in project planning were meant to help elevate literacy and reduce common problems in working with EDW4R data.

Service management coordinated with enterprise IT

About one-third of respondents described coordination between CTSA informatics and enterprise IT organizations, whereby enterprise IT organizations had a standard operating procedure to redirect research requests to CTSA informatics for fulfillment. Four hubs described use of the ServiceNow request management platform by enterprise IT organizations separate from CTSA informatics. At one institution, investigators submitted data requests for research, quality improvement, and operations using a central data request form managed by enterprise IT, which triaged research requests to CTSA informatics. One respondent indicated that CTSA informatics supported "quality improvement (QI) activities to try to sell research" to the clinical community and as a way "to demonstrate value to the health system." Additionally, the respondent indicated that CTSA informatics provided "better QI support than [enterprise IT]."

Service management coordinated within a CTSA hub

No CTSA hubs reported using a formal service management approach, such as the Information Technology Infrastructure Library, and most described some homegrown service management processes. At least 2 sites described ongoing efforts to formalize service management processes that were "undisciplined" or "organic." In contrast, another hub characterized its approach as "smooth."

Some institutions (n = 4) reported routing service requests through a CTSA hub managed processes for all services, not just informatics, while others (n = 9) had researchers directly contact the EDW4R team. To manage service requests from investigators, hubs used a variety of software tools focused on CTSA institutions (n = 7), including REDCap (Research Electronic Data Capture) and SPARC (Scholarly Publishing and Academic Resources Coalition), as well as commercial offerings (n = 8), such as ServiceNow, Jira, TRAC, and Salesforce.

To prioritize requests, most hubs followed a first-in-first-out approach with exceptions for grant deadlines and leadership requests. Some institutions reported not having a formal prioritization method, while others stated that grant-funded projects and clinical trial recruitment activities took precedence. At some institutions, smaller jobs (eg, taking <4 hours) would be given higher priority and more complex jobs would require funding.

Service management coordinated between informatics and biostatistics

Multiple institutions described informatics and biostatistics consultations as critical to study approval and data request processes. Of note, 3 CTSA hubs held "one-stop shop" in-person sessions including representatives from informatics, biostatistics, IRB, and clinical trials to provide comprehensive support to investigators. One institution described the research project intake process as requiring biostatistics and informatics review, while other institutions described encouraging investigators to seek informatics consultations "early and often" prior to submitting IRB applications. Compared with junior investigators, senior investigators at one institution were more likely to include biostatistics personnel when requesting informatics support.

Funding approaches

To manage demand, institutions described use of fee-for-service and full-time equivalent funding approaches (Table 3). Most CTSA hubs reported having more demand than capacity to fulfill data requests.

Table 3. Examples of fee for service and full-time equivalent funding of EDW4R resources

Type	Example

Fee for service	• $75 hourly fee (subsidized by CTSA)
	• $120 hourly fee (subsidized by CTSA)
	• Variable fee schedule (subsidized by CTSA for academic partners)
	  • Internal principal investigator: $105 hourly fee
	  • External principal investigator (academic): $150 hourly fee
	  • External commercial sponsor: $175 hourly fee
	• Undisclosed fee if in excess of threshold of effort
	  • 120 h
	  • 8 h
	  • 1 h

FTE funding	• If effort is >10% FTE • 20%-30% of FTE funded by grants

CTSA: Center for Advancing Translational Sciences; EDW4R: electronic data warehouses for research; FTE: full-time equivalent.

At one institution, expanding from 1 to 10 database analysts failed to scale due to demand outpacing supply.

For fee for service, some institutions employed fee schedules with different rates for internal investigator–initiated studies, external investigator–initiated studies, and industry sponsor–initiated studies; rates were subsidized by CTSA hubs. Others implemented an hourly chargeback rate for database analyst time. Although many institutions described providing certain services for free (eg, i2b2), respondents described a threshold of hours or effort whereby institutions required funding from investigators to proceed with data extraction. While one institution reported fee for service as supporting growth of informatics staff, another reported little revenue generated. One institution described investigator preference for funding full-time equivalent portions from grants to support study activities as compared with fee for service. Fee-for-service strategies were often described as being in a state of evolution; some institutions changed their recharge methods to address pushback from the community.

Performance metrics

As shown in Table 4, respondents described capturing a number of EDW4R metrics with considerable variation. For tracking data request turnaround time, some hubs distinguished between time expended by informatics staff vs noninformatics personnel, such as investigators and regulators. One respondent said, "focus on [measuring] what [informatics] can control." Turnaround time varied based on complexity of data extraction performed by informatics staff as well as informatics staff waiting on institutional approvals (eg, IRB decisions) and input from investigators. Accordingly, 2 sites described measuring time from investigator request to informatics response, and one site described measuring estimated and actual turnaround time for data requests. Hubs noted challenges in measuring time to request fulfillment.

In addition to turnaround time, hubs measured the volume of services provided with respect to requests fulfilled and system users. One site emphasized the need to track consultations between investigators and informatics analysts to understand value added prior to or in lieu of data requests. After completing informatics service requests, 2 hubs formally surveyed investigators regarding data provided and experience with the request process, and one hub reported

Table 4. Metrics of EDW4R operations

Type	Measures	
Usage	Users, data requests, projects, and tickets	12
Duration	Turnaround time for requests and other processes	10
Outcomes	Grants and publications	10
Cost	Time or total dollars spent	6
Feedback	Surveys and evaluations from end users	4
User characteristics	Types of trainees, faculty members, and power users	3

EDW4R: electronic data warehouses for research.

conducting an annual survey of all investigators supported regarding satisfaction.

Measurement of impact of CTSA informatics focused mostly on grants and publications, with one respondent describing more than 50 articles supported in 8 years. However, 4 institutions described difficulty in tracking impact due to the need for investigators to respond to surveys about CTSA resources supporting research. One site described automated emails sent to all investigators who requested CTSA informatics services inquiring about whether the services supported grants and publications. Two sites described collaborations with CTSA evaluation cores to measure impact. One respondent described similarity between CTSA hub and National Cancer Institute Cancer Center impact tracking. Notably, 3 institutions reported not capturing metrics for EDW4R services.

Future trends and current technology challenges

All institutions hosted EDW4R on premises, but 6 described plans for future cloud migration with one institution calling the cloud "the next generation environment" to support the NIH Strategic Plan for Data Science.[37] Although EDW4Rs generally obtained copies of clinical data from EDWs managed by health systems, some CTSA hubs described use of remote access via virtual private network to extract data from particular legacy applications within health system firewalls. Three institutions described use or planned use of secure data analysis environments for researchers to interact with data obtained from EDW4Rs.[38,39] Managing user identity of researchers for provisioning and de-provisioning access to sensitive datasets was noted as a challenge by 2 respondents.

DISCUSSION

In this study, informatics leaders from 20 CTSA hubs described EDW4R operations with respect to architecture, processes for access, and service management. We identified 12 themes describing varying operational approaches and maturity levels. Findings demonstrate diversity of current structures, challenges in service management, and rapid change facing the current management and future sustainability of EDW4R activities. Additionally, results indicate growing maturity of practices in architecture and engagement. This study is a step in developing an improved understanding of EDW4R activities across CTSA hubs. An opportunity may exist to define best practices for EDW4R operations. With a clear roadmap and maturity path to follow, institutions could potentially improve efficiency, reduce costs, and better serve investigators. In particular, guidance in 3 areas—migration of EDW4R to the cloud, data governance, and relationships between CTSA informatics and enterprise IT—may warrant further investigation.

Although respondents did not explicitly describe successes and failures of EDW4R approaches, success appeared related to meeting local cultural needs. For example, some institutions wanted their EDW4R to support a broad research portfolio with a range of users, while others used the EDW4R to support a small set of use cases (eg, enable clinical trial cohort identification). Culture appeared to also determine engagement and communication approaches, as respondents demonstrated considerable variability. Success factors that might apply to any institution's EDW4R operations include executive sponsorship (ie, leadership awareness of and support for EDW4R), EDW4R representation (ie, EDW4R leadership representation in institutional data governance), defined and communication processes (ie, well-known methods for guiding researchers to access data), service management collaborations (ie, engagement between EDW4R, enterprise IT, and other CTSA cores), and service measurement (ie, tracking EDW4R activities).

This study has limitations. Our work was sponsored by the CTSA iDTF as a working group deliverable and was performed within a limited time frame, which reduced the scope of what we could cover as well as the how many institutions we could interview. As a result, we have not explored some key aspects of EDW4R operations, such as workforce expertise of informatics staff necessary to deliver services and which use cases yield greatest institutional value. Despite the limited scope, we initiated a dialog across the CTSA community on a topic of high interest.

At a time when academic medical centers are struggling to control costs, it can be difficult to prioritize investing in research infrastructures such as an EDW4R and informatics specialists. Understanding successful approaches from CTSA hubs may help establish best practice guidance that can reduce costs of EDW4R operations and better support scientists with electronic patient data.

CONCLUSION

Through qualitative analysis of interviews with 20 CTSA informatics leaders, we identified 12 themes describing EDW4R operations. The 12 themes may suggest best practices that can provide guidance in EDW4R development and establish groundwork for future investigations and development of maturity models.

FUNDING

This study received support from the National Institutes of Health National Center for Advancing Translational Sciences through grant numbers UL1TR002537 (BMK), UL1TR000457 (TRC), UL1TR002369 (DAD), and UL1TR001433 (CKC).

AUTHOR CONTRIBUTIONS

BMK conceptualized the study and interview guide, and BMK and TRC conducted interviews, transcribed notes, and performed analysis. CKC and DAD provided analytical feedback. TRC wrote the manuscript with contributions from BMK. CKC and DAD edited the manuscript. TRC and BMK revised the manuscript.

SUPPLEMENTARY MATERIAL

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

ACKNOWLEDGMENTS

The authors thank Jeanne Holden-Wiltse and Bradley Taylor for data collection assistance.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Obeid JS, Tarczy-Hornoch P, Harris PA, *et al*. Sustainability considerations for clinical and translational research informatics infrastructure. *J Clin Trans Sci* 2018; 2 (5): 267–75.
2. 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.
3. 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 Trans Sci* 2017; 1 (4): 246–52.
4. Payne PRO, Johnson SB, Starren JB, Tilson HH, Dowdy D. Breaking the translational barriers: the value of integrating biomedical informatics and translational research. *J Investig Med* 2005; 53 (4): 192–200.
5. Pulley JM, Shirey-Rice JK, Lavieri RR, *et al*. Accelerating precision drug development and drug repurposing by leveraging human genetics. *Assay Drug Dev Technol* 2017; 15 (3): 113–9.
6. Denny JC, Bastarache L, Roden DM. Phenome-Wide Association Studies as a Tool to Advance Precision Medicine. *Annu Rev Genom Hum Genet* 2016; 17 (1): 353–73.
7. Kohane IS, Churchill SE, Murphy SN. A translational engine at the national scale: informatics for integrating biology and the bedside. *J Am Med Inform Assoc* 2012; 19 (2): 181–5.
8. Hripcsak G, Ryan PB, Duke JD, *et al*. Characterizing treatment pathways at scale using the OHDSI network. *Proc Natl Acad Sci U S A* 2016; 113 (27): 7329–36.
9. Claerhout B, Kalra D, Mueller C, *et al*. Federated electronic health records research technology to support clinical trial protocol optimization: evidence from EHR4CR and the InSite platform. *J Biomed Inform* 2019; 90: 103090.
10. Campion TR, Sholle ET, Davila MA. Generalizable middleware to support use of redcap dynamic data pull for integrating clinical and research data. *AMIA Jt Summits Transl Sci Proc* 2017; 2017: 76–81.
11. Longhurst CA, Harrington RA, Shah NH. A "green button" for using aggregate patient data at the point of care. *Health Aff (Millwood)* 2014; 33 (7): 1229–35.
12. Baghal A, Zozus M, Baghal A, Al-Shukri S, Prior F. Factors associated with increased adoption of a research data warehouse. *Stud Health Technol Inform* 2019; 257: 31–5.
13. Danciu I, Cowan JD, Basford M, *et al*. Secondary use of clinical data: the Vanderbilt approach. *J Biomed Inform* 2014; 52: 28–35.
14. Chute CG, Beck SA, Fisk TB, Mohr DN. The enterprise data trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. *J Am Med Inform Assoc* 2010; 17 (2): 131–5.
15. Sholle ET, Kabariti J, Johnson SB, *et al*. Secondary use of patients' electronic records (SUPER): an approach for meeting specific data needs of clinical and translational researchers. *AMIA Annu Symp Proc* 2017; 2017: 1581–8.
16. Kamal J, Liu J, Ostrander M, *et al*. Information warehouse-a comprehensive informatics platform for business, clinical, and research applications. *AMIA Annu Symp Proc* 2010; 2010: 452–6.
17. Lowe HJ, Ferris TA, Hernandez PM, Weber SC. STRIDE–an integrated standards-based translational research informatics platform. *AMIA Annu Symp Proc* 2009; 2009: 391–5.
18. Wade TD, Hum RC, Murphy JR. A Dimensional Bus model for integrating clinical and research data. *J Am Med Inform Assoc* 2011; 18 (Suppl 1): i96–102.
19. Starren JB, Winter AQ, Lloyd-Jones DM. Enabling a learning health system through a unified enterprise data warehouse: the experience of the Northwestern University Clinical and Translational Sciences (NUCATS) Institute. *Clin Transl Sci* 2015; 8 (4): 269–71.
20. Mosa ASM, Yoo I, Apathy NC, Ko KJ, Parker JC. Secondary use of clinical data to enable data-driven translational science with trustworthy access management. *Mo Med* 2015; 112 (6): 443–8.
21. Waitman LR, Warren JJ, Manos EL, Connolly DW. Expressing observations from electronic medical record flowsheets in an i2b2 based clinical data repository to support research and quality improvement. *AMIA Annu Symp Proc* 2011; 2011: 1454–63.
22. Murphy SN, Weber G, Mendis M, *et al*. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc* 2010; 17 (2): 124–30.
23. Visweswaran S, Becich MJ, D'Itri VS, *et al*. Accrual to clinical trials (ACT): A clinical and translational science award consortium network. *JAMIA Open* 2018; 1 (2): 147–52.
24. Weber GM, Murphy SN, McMurry AJ, *et al*. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. *J Am Med Inform Assoc* 2009; 16 (5): 624–30.
25. Topaloglu U, Palchuk MB. Using a federated network of real-world data to optimize clinical trials operations. *JCO Clin Cancer Inform* 2018; 2 (2): 1–10.
26. Hripcsak G, Duke JD, Shah NH, *et al*. Observational health data sciences and informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform* 2015; 216: 574–8.
27. Denny JC, Rutter JL, Goldstein DB, *et al*. The "All of Us" Research Program. *N Engl J Med* 2019; 381 (7): 668–76.
28. Collins FS, Hudson KL, Briggs JP, Lauer MS. PCORnet: turning a dream into reality. *J Am Med Inform Assoc* 2014; 21 (4): 576–7.
29. Hernandez AF, Fleurence RL, Rothman RL. The ADAPTABLE trial and pcornet: shining light on a new research paradigm. *Ann Intern Med* 2015; 163 (8): 635–6.
30. Hersh WR, Cimino J, Payne PRO, *et al*. Recommendations for the use of operational electronic health record data in comparative effectiveness research. *EGEMS (Wash DC)* 2013; 1 (1): 1018.
31. Bernstam EV, Hersh WR, Johnson SB, *et al*. Synergies and distinctions between computational disciplines in biomedical research: perspective from the Clinical and Translational Science Award programs. *Acad Med* 2009; 84 (7): 964–70.
32. Axelos. What is IT service management? https://www.axelos.com/best-practice-solutions/itil/what-is-it-service-management Accessed December 2, 2019.
33. Shin S-Y, Kim WS, Lee J-H. Characteristics desired in clinical data warehouse for biomedical research. *Healthc Inform Res* 2014; 20 (2): 109–16.
34. Hsieh H-F, Shannon SE. Three approaches to qualitative content analysis. *Qual Health Res* 2005; 15 (9): 1277–88.
35. Lincoln YS. Naturalistic inquiry. In: Ritzer G, ed. *The Blackwell Encyclopedia of Sociology*. Oxford, UK: Wiley; 1985: 221–357.
36. Boyd AD, Saxman PR, Hunscher DA, *et al*. The University of Michigan Honest Broker: a Web-based service for clinical and translational research and practice. *J Am Med Inform Assoc* 2009; 16 (6): 784–91.
37. National Institutes of Health. STRIDES Initiative. https://datascience.nih.gov/strides Accessed December 1, 2019.
38. Bradford W, Hurdle JF, LaSalle B, Facelli JC. Development of a HIPAA-compliant environment for translational research data and analytics. *J Am Med Inform Assoc* 2014; 21 (1): 185–9.
39. Oxley PR, Ruffing J, Campion TR, Wheeler TR, Cole CL. Design and implementation of a secure computing environment for analysis of sensitive data at an academic medical center. *AMIA Annu Symp Proc* 2018; 2018: 857–66.