Understanding Enterprise Data Warehouses to Support Clinical and Translational Research: Initial Findings on Architecture, Processes, and Service Management

Boyd Knosp, MS¹, Catherine K. Craven, PhD, MA, MLS², David A. Dorr, MD, MS³,4, Thomas R. Campion, Jr., PhD⁵,6,7,8

¹Roy J. and Lucille A. Carver College of Medicine and the Institute for Clinical & Translational Science, University of Iowa, Iowa City, IA; ²Institute for Healthcare Delivery Science, Icahn School of Medicine at Mount Sinai, New York, NY; ³Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR; ⁴Department of Medicine, Oregon Health & Science University, Portland, OR; ⁵Department of Population Health Sciences, Weill Cornell Medicine, New York, NY; ⁶Department of Pediatrics, Weill Cornell Medicine, New York NY; ⁷Information Technologies & Services Department, Weill Cornell Medicine, New York, NY; ⁸Clinical & Translational Science Center, Weill Cornell Medicine, New York, NY

Introduction

Electronic patient data is critical to the conduct of clinical and translational science, and experts have noted that “an [enterprise] data warehouse (EDW) and related 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 EHR data for research has increased steadily over time (2). Although EDW for research (EDW4R) services are now nearly ubiquitous (2), little is known about how individual CTSA hubs have implemented EDW4R activities, hindering the optimization of informatics for biomedical research across the CTSA consortium. 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.

Methods

The Informatics Domain Task Force (iDTF) of the CTSA consortium commissioned this study through the Enterprise Data Warehouse Working Group (EDW WG) co-led by two co-authors (BK, TRC) with support from one co-author (DAD) as liaison from the NCATS National Center for Data to Health (CD2H).

We conducted semi-structured interviews with informatics leaders responsible for EDW4R activities at 22 CTSA hubs. Interviews occurred at the American Medical Informatics Association (AMIA) Informatics Summit as well as the iDTF/CD2H Face-to-Face meeting held March 2019 in San Francisco and teleconferences through May 2019. All semi-structured interviews followed a guide covering three areas—organizational and technical architecture, processes for access, and service management—developed by two authors (BK, TRC) based on their experience operating EDW4R activities in two CTSA hubs. Using transcripts of interview notes, the study team performed a thematic coding analysis using a modified grounded theory approach. The University of Iowa Institutional Review Board determined this study to be non-human subjects research.

Results

Architecture

The relationship between a CTSA hub and health systems (i.e., clinical enterprises) influenced EDW4R organization. Approaches spanned the spectrum from CTSA hubs having no EDW4R to a separate EDW4R. Specifically, where no EDW4R existed, a health system-only EDW was in operation and provided limited research support. Some hubs reported having a shared EDW between health system and research groups. Others described a separate EDW4R containing data from a single health system whereas some hubs described a separate EDW4R storing data from multiple systems and/or statewide health information exchange efforts. CTSA informatics leaders characterized their relationships with health system information technology (IT) organizations variably with some noting a close partnership and others noting a “significant separation” or being treated as a “client.”
All CTSA hubs hosted their EDW4R in an on-premises data center, although six described planning but not yet implementing cloud approaches. Some hubs described remote access to a health system EDW to perform specific data extracts.

Respondents indicated their hubs supported multiple common data models (CDMs) including PCORnet CDM (n=10), i2b2 (n=17), and OMOP (n=7) as well as a number of research networks such as PCORnet, NCATS ACT, SHRINE networks separate from NCATS ACT, the NIH All of Us Research program, TriNetX, and professional society registries. Hubs also described supporting a number of self-service tools for investigators to interact with electronic patient data including those designed for clinical data (e.g., i2b2, TriNetX, Epic SlicerDicer), commercial business intelligence tools (e.g., BusinessObjects, Tableau, QlikView), statistical software (e.g., Rshiny), and direct structured query language (SQL) access along with other custom tools. Consolidation trends appeared with hubs describing NCATS ACT subsuming separate SHRINE networks as well as a shift from i2b2 to TriNetX for self-service queries.

EDW4Rs contained data from health systems’ Epic and/or Cerner EHR implementations as well biorepository, clinical trials management, electronic data capture, and legacy clinical systems along with external data sources such as the Social Security Death Master File. Respondents also described availability of secure data analysis environments and reported challenges in managing user identity, especially across multiple institutions involved in a CTSA hub.

Processes for access

In processes for access, we sought to understand pathways for approving and managing how investigators obtain data. Governance and engagement emerged as major themes. For governance, not all hubs described having a formal governance process for determining researcher access to EDW4R resources. Some hubs described an IRB protocol covering the EDW4R with the EDW4R leader as principal investigators that required clinical researchers to be added as co-investigators; others reported separate IRB protocols from investigators needing to have EDW4R personnel added to them. For investigator engagement, approaches varied with respect to community awareness. Hubs described in-person training sessions as well as online announcements and training. Recognizing varying skillsets of investigators, some hubs provided levels for novice (those requiring concierge service), power users (adept at self-service tools), and expert users (SQL/SAS/Python programmers). Managing demand was an area where hubs struggled with some using governance driven prioritization processes and others fee-for-service mechanisms.

Service management

Service management, or methods of delivering activities, was the most varied of all domains. Almost all institutions lacked any formal service management framework. One institution reported having ITIL (Information Technology Infrastructure Library) training a number of years ago but not continuing to practice it. Many had a ticketing system for tracking requests. Some reported a connection with their enterprise IT service desk either as a front end to the EDW4R services or as a referral mechanism. There was a growing trend to improve service management with hubs stating “we are working on a new ticketing system” and “we are looking at ways to work closer with enterprise IT.” Hubs described capturing four broad areas of service metrics including usage (number of requests, number of interactions), impact (number of grants supported, number of publications), performance (time to respond to data requests), and financial (cost of supporting requests, staff time, return on investment).

Discussion

Our motivation for adopting a qualitative approach was the observation that the CTSA community lacks a common vocabulary for discussing operational aspects of EDW4R. Initial observations described here may yield more substantive relationships as we continue our analysis across domains. For example, is there a link between formality of service management and data governance as well as architecture and processes for accessing the data? Such higher order trends could provide key elements to inform best practices and models for maturity in EDW4R operations.

Acknowledgements

This study received support from UL1TR002537, UL1TR00457, UL1TR002369, and UL1TR001433.

References