Secondary Use of Patients' Electronic Records (SUPER): An Approach for Meeting Specific Data Needs of Clinical and Translational Researchers

Evan T. Sholle, MS[1], Joseph Kabariti, BS[1], Stephen B. Johnson, PhD[2,3], John P. Leonard, MD[4], Jyotishman Pathak, PhD[2], Vinay I. Varughese, BS[1], Curtis L. Cole, MD[1,2,4], Thomas R. Campion, Jr., PhD[1,2,3,5]

[1]Information Technologies and Services Department, Weill Cornell Medicine, New York, NY; [2]Department of Healthcare Policy and Research, Weill Cornell Medicine, New York, NY; [3]Clinical and Translational Science Center, Weill Cornell Medicine, New York, NY; [4]Department of Medicine, Weill Cornell Medicine, New York, NY; [5]Department of Pediatrics, Weill Cornell Medicine, New York, NY

Abstract

Academic medical centers commonly approach secondary use of electronic health record (EHR) data by implementing centralized clinical data warehouses (CDWs). However, CDWs require extensive resources to model data dimensions and harmonize clinical terminology, which can hinder effective support of the specific and varied data needs of investigators. We hypothesized that an approach that aggregates raw data from source systems, ignores initial modeling typical of CDWs, and transforms raw data for specific research purposes would meet investigator needs. The approach has successfully enabled multiple tools that provide utility to the institutional research enterprise. To our knowledge, this is the first complete description of a methodology for electronic patient data acquisition and provisioning that ignores data harmonization at the time of initial storage in favor of downstream transformation to address specific research questions and applications.

Introduction

Secondary use of electronic health record (EHR) data to support biomedical research is challenging. A common approach to secondary use of EHR data is a centralized clinical data warehouse (CDW)[1-5] involving dimensional modeling and terminological harmonization of source system data as a precondition for storage and use. To provide CDWs, institutions require extensive resources and may experience slow delivery of solutions for investigators due to maintenance of centralized data curation methods[6]. Additionally, the centralized modeling of data may limit the ability to meet researchers' specific needs[1,6]. Optimal approaches to secondary use of EHR data for meeting investigator needs are unknown.

Lacking a centralized clinical data warehouse, our institution needed to meet investigator needs for secondary use of EHR data. We hypothesized that an approach that aggregates raw data from source systems, ignores initial modeling typical of clinical data warehouses, and transforms raw data for specific research purposes would meet investigator needs. To our knowledge, literature describing such an approach to secondary use of EHR data is limited. The objective of this case report is to describe our approach to inform efforts at other institutions.

Methods

Setting

Weill Cornell Medicine (WCM), located on the Upper East Side of Manhattan in New York City, is the clinical research facility and medical college of Cornell University. WCM supports an academic staff of 1,049 physicians and educators and trains over 950 resident physicians and 450 medical students yearly. Faculty clinicians see patients through the Weill Cornell Physician Organization, a multispecialty group practice with more than 900 physicians and more than 20 sites across New York City. WCM faculty physicians have admitting privileges to NewYork-Presbyterian Hospital (NYP), a 2,478-bed hospital with multiple facilities. As long-time clinical affiliates yet separate legal entities, WCM and NYPH frequently collaborate in regards to clinical and translational research. In order to strengthen this collaboration, WCM and NYPH established the Joint Clinical Trials Office (JCTO) in 2013 to grow research activities between the two institutions. The JCTO regularly collaborates with the Clinical and Translational

Science Center (CTSC) and the information technology (IT) departments at WCM and NYPH for research integrity and resource access purposes.

WCM and NYPH make use of multiple EHR systems from different vendors. In the outpatient setting, WCM Physician Organization Information Services (POIS) provides Epic Ambulatory. In the inpatient and emergency setting, NYPH uses Allscripts Sunrise Clinical Manager (SCM). Additional ancillary EHR systems cover various specialty areas, including perioperative documentation and imaging. Across WCM and NYPH, patients have a shared medical record number (MRN). Multiple interfaces between WCM and NYPH enable sharing of data.

The Information Technology and Services (ITS) department at WCM provides IT infrastructure, management, and service for the WCM community. Included within ITS is the Research Informatics division (RI), which, with financial support from the JCTO and CTSC, administers the Architecture for Research Computing in Healthcare (ARCH) program, a suite of tools and services for investigators to obtain electronic patient data.

System Description

To support the research enterprise with patient electronic data, RI has implemented a methodology and technical infrastructure termed Secondary Use of Patient Electronic Records (SUPER). SUPER undergirds all of RI's efforts to provide investigators with patient data processes, and comprises multiple components, from data acquisition and manipulation, to feeding other clinical and research tools, to the underlying technical infrastructure. Figure 1 illustrates the workflow that powers SUPER activities.

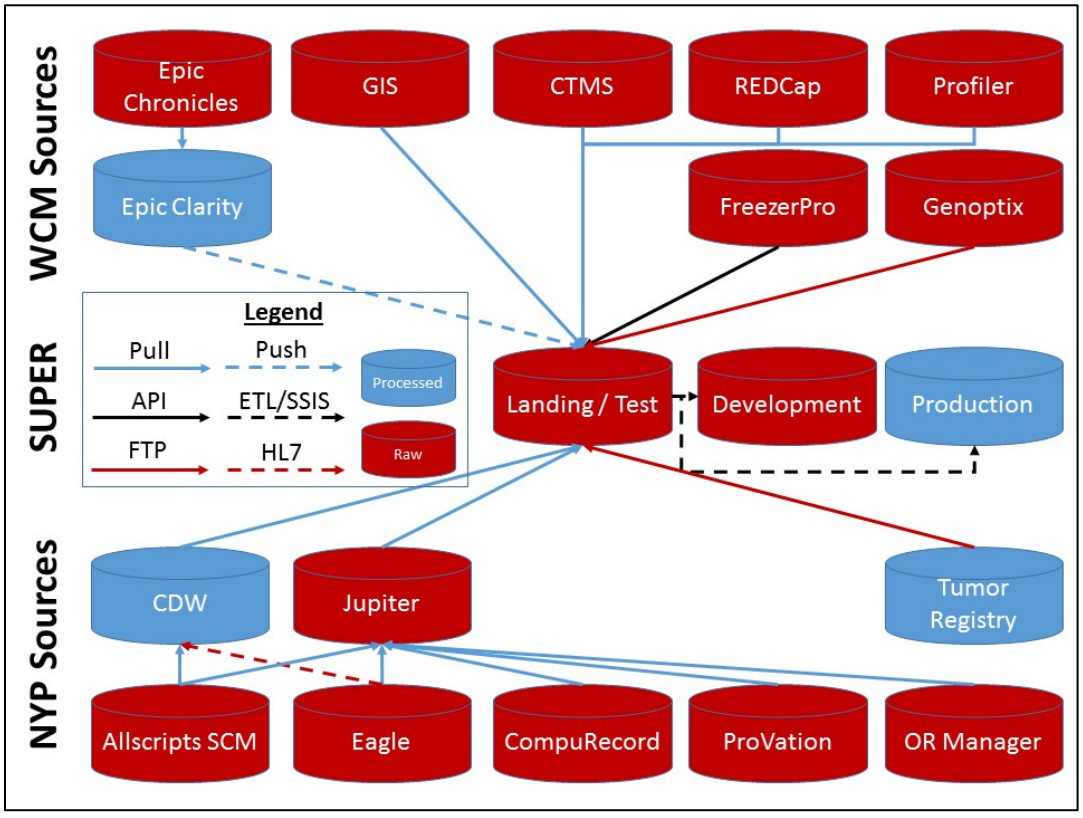

Figure 1. Model of flow from source systems to SUPER infrastructure

Data sources

SUPER stores data from multiple clinical and research information systems, comprising the entirety of data gathered to support clinical, billing, and research activities across WCM and NYP. Data enter SUPER and are maintained in the same format in which they exist in their source systems.

For outpatient clinical and billing data, SUPER receives copies of Epic Clarity database tables which is a relational subset of the original transactional data from the EpicCare EHR. These relational data enter the SUPER infrastructure directly without any additional transformation. From the inpatient EHR, SUPER obtains a regular data feed from

NYP's Clinical Data Warehouse (CDW)[7], which similarly aggregates inpatient billing data in the form of HL7 messages from NYP's Eagle patient accounting system, and the Jupiter analytics database. Jupiter contains tables copied directly from the Allscripts SCM inpatient EHR, as well as data from ancillary systems, including CompuRecord, the inpatient perioperative EHR for anesthesia administration, OR Manager provides operative data, and ProVation provides documentation from endoscopic procedures. SUPER imports all data from NYP's CDW and Jupiter database without any form of transformation or harmonization; tables are replicated exactly as they exist in the source systems.

For biospecimen data, SUPER obtains data from Profiler, a locally developed system used by one of the larger tissue banks, and RURO FreezerPro, a commercial specimen management system used to track inventory for other research programs. For tumor registry data, SUPER receives registry data from a third-party vendor responsible for abstracting charts according to North American Association of Central Cancer Registries (NAACCR) standards. WCM's clinical trials management system[8] feeds directly to SUPER, providing records of patient enrollment in research protocols. For next generation sequence (NGS) data, SUPER receives data from the Standard Molecular genomic information system (GIS) for molecular tests, including the EXaCT1 whole exome test performed at WCM, and from Genoptix, a third-party laboratory. For case report form data, SUPER obtains data from WCM's instance of REDCap[9].

Data acquisition process

As described above, the SUPER infrastructure and methodology depends on multiple partner organizations and teams, all of whom maintain their own clinical and research data systems. Obtaining data from these sources systems requires careful coordination with the teams responsible for their maintenance and care – the data systems feeding SUPER also send data to other systems and, in some cases, support user-facing activities. Given the significant demands imposed by regular data transfers, flexibility is crucial in order to avoid interference with clinical operations and other activities. To facilitate a flexible and modular approach, a series of extract, transform, load (ETL) database operations, application program interface (API) calls, and flat file imports provide data to SUPER. While SUPER pulls data from most source systems, some external systems push data to SUPER. Most source-to-SUPER data transmissions occur monthly while some occur weekly and daily to support specific projects.

The primary intake of data from the outpatient EHR occurs monthly over a one-week period, with each night of the seven-day period handling one section of a large ETL. Generally, each night's job covers some of the significantly larger tables, along with related smaller lookup and metadata tables. The modular nature of this process allows for one night's worth of data to be available immediately for use: in the event that a subsequent transfer fails, not all data is useless. Due to the size of the data set and underlying infrastructure limitations, the refresh is limited to 5 to 7 nights a week per source system. To ensure each night's job proceeds as quickly as possible and ensure the capture of any retroactively imposed changes, WCM POIS will wipe any relevant data from the servers before initiating the transfer of the data set. A delta load approach, while appealing, would render the process overly time-consuming.

For data intake from the inpatient EHR via NYP's CDW and Jupiter, the acquisition process differs. To allow for fine-grained control, we automate the process through the combination of Microsoft SQL Server Integration Services (SSIS) and SQL Server Agent. Using the proper driver – i.e., a DB2 driver to power the transfer from a DB2 server – we configure an encrypted SSIS package to make a connection to the relevant source tables, run a SQL query pulling out that data, and transfer the data to the SUPER landing server. The SSIS package is configured to ensure the preservation of the source schema. Overall these ETLs span a seven-day period, with each day encompassing several table transfers.

Other sources require different processes. To obtain NGS data from Genoptix, we rely on a cron job run from a virtual machine to pull data from a CSV file located on an FTP server into the SUPER infrastructure. To integrate data from FreezerPro, a Python script polls FreezerPro's API and pulls data into SUPER. Intake of data from NAACCR's tumor registry requires a SAS dictionary invoked from the command line, in conjunction with a Python script to transform the data and load it into SUPER. SUPER obtains data from small-scale systems, including the clinical trials management system and REDCap, more frequently, as the load is significantly smaller than other incoming data sources.

ETL code management

All ETL scripts are crucial to SUPER workflows. To ensure their security, currency, and homogeneity, SUPER relies on Subversion (SVN), a version control system that allows developers to save and compare their ETL scripts across

multiple domains and develop the versions of the process over time. With SVN, one change to an ETL script will propagate across multiple systems at once, ensuring that all iterations of a data model are identical.

To ensure that no individual ETL usurps all available resources, thereby affecting other jobs, the RI schedule places refreshes of the larger data sources at the last and first week of the month, and the smaller sources over a weekend. Individual data mart requests and *ad hoc* reporting take place during the two to three weeks in between each larger refresh, with bigger jobs occurring over a weekend. To avoid interference with day-to-day operations, weeknights and weekends are dedicated for resource-intensive ETL jobs.

ETLs are inherently ambiguous: while they inform the *what* and *where* for the transfer of data, they do not inform the *how*. Paver, a task manager developed in Python, manages priority levels and the order of task execution[10]. With the use of SVN and a collaborative development process, ETL scripts are automatable and reusable. Paver renders them modular, affording the ability to configure the order of a set of scripts and have them run piecewise, thereby automating those reusable scripts in different contexts.

Indexing

SUPER stores several terabytes of source data, with some tables holding billions of rows of data. Without an indexing process, it is impossible to expect any ETL, or even query, to finish within a practical timeframe. However, indexing is time-consuming and must take place after every refresh cycle. To optimize the indexing process and ensure it takes place as quickly as possible, a cron job powered by Paver syncs index generation code with the latest revision and runs the index refresh script every six hours unless an ETL is currently running, signaled by a system flag created as part of the ETL script. The indexing script uses dynamic SQL to drop and recreate stored procedures. These stored procedures reduce active indexing time through the use of a procedural log, rather than a transactional log, and through previewing and optimization of the index processing time. Indexing scripts first check for presence of the index, with its presence signaling the job to move on to the next index. Since indexes are recreated with every refresh, there is no possibility of long term index usage and bug appearances - therefore, the method of ignoring an already created index should be sufficient.

Terminology management

SUPER does not conduct terminology management or data harmonization on sources accessioned to the landing server, instead relying on efforts already put in place by POIS and NYP IS to harmonize data. While not detailed here, a great deal of effort goes into standardizing data as it enters the EHR using an internally developed terminology server, TruData©, that normalizes data against reference and interface terminologies.

RI pulls data from a variety of sources, each of which is organized according to its own internal schema. Some data sources have already been subject to a data harmonization process, leveraging proprietary data models or other schemas and consistently using reference terminologies. Examples include Epic Clarity, which transforms Epic Chronicles data, and the NYP CDW, which has rules for data representation[7]. Others consist of relatively raw transactional data. While internal data dictionaries for individual source systems afford developers and analysts the ability to understand the relations between tables within source systems, exploration of data is necessary to understand table relationships within and across source systems. Therefore, rather than maintain a common data dictionary for the various systems[5], SUPER relies on source system use of reference terminologies and engineering expertise to query data within and across applications.

Documentation and workflow management

To ensure a streamlined process for maintaining SUPER, analysts and engineers use multiple off-the-shelf work management tools, including ServiceNow, Jira, Subversion, Sharepoint, Box, Outlook, and Slack. RI has implemented a system of workflow management that leverages these tools to ensure all team members know what to do, when to do it, and how to ask for help. ServiceNow, WCM ITS's internal ticketing system, allows for the triage of data requests and inquiries from investigators and research personnel, as well as infrastructure change requests and access control issues. To track work, including acquisition and integration of new data sources, data quality efforts, and the development of user-facing informatics tools, Jira allows for ticket-level tracking of tasks, priorities, and day-to-day work logging. As previously discussed, SVN allows developers to maintain and share code.

The use of SharePoint and Box allows team members to collaborate on external-facing reports and presentations to the research community, as well to aggregate and track documents gathered from and produced for investigators. A shared Outlook calendar makes all team members aware of ongoing ETL jobs or other pertinent events. Confluence stores wiki documentation written for the reusable code available to all team members, as well as ongoing issues

relating to individual tables, data sources, or user-facing tools. After a researcher submits a request via ServiceNow and engineers use Jira to work toward a solution, team members use Confluence to represent the knowledge distilled through the development process.

A weekly code review session allows software engineers to circulate and present code to their colleagues, exposing problems or issues, requesting help, or working together to develop and implement novel solutions. WCM ITS's departmental Slack instance allows all team members to stay in constant contact, notifying each other of pertinent issues, as well as facilitating contact with other ITS business units to address issues relating to infrastructure, identity management, or project management.

Hardware and software infrastructure

SUPER consists of four Microsoft SQL Server 2014 database servers and five Linux virtual machines. WCM ITS provides clusters of virtual servers, which allow a more robust and flexible system of maintenance and deployment. The primary server is the landing server[11], which uses 40 gigabytes of random access memory (RAM), 21 terabytes of hard disk drive storage, and 20 virtual CPU cores. The landing server requires extensive system resources to ensure that data flows unhindered from source systems. In addition to enabling queries for the most recent data, the landing server provides data to development and production servers.

The SUPER development server is populated with data a week behind the regular refresh schedule of the landing server so as not to perturb data pulls; it also provides a backup should the landing server fail. Configured with 16 gigabytes random access memory, approximately 12 terabytes of hard disk drive storage, and 13 virtual CPU cores, the development server is not as powerful as the landing server, but allows for the development and trials of small ETLs, data mining queries, and relational data exploring. RI endeavors to keep the server as closely in line as possible with the landing server to ensure backup capabilities as well as utility for query transferring.

For production purposes, SUPER has two separate servers. For identified data available to researchers who access RI services, the production servers use 16 gigabytes of RAM, approximately 5 terabytes of solid state drive space, and 9 virtual CPU cores. Production servers are intended to handle multiple concurrent users.

Results

As shown in Figure 2, SUPER has enabled ARCH to support WCM researchers with multiple tools for using electronic patient data. For cohort discovery, i2b2 enables WCM investigators to query de-identified data from EHR and research systems[12]. For EHR analytics, the Observational Medical Outcomes Partnership common data model (OMOP CDM)[13] allows researchers to perform robust queries using standardized vocabularies. Alternately, for investigators with specific needs, developers and analysts work together to generate custom data marts from SUPER tables. Using REDCap in conjunction with the REDCap dynamic data pull (DDP) plugin and generalizable middleware to support REDCap DDP, investigators can populate REDCap fields with EHR data stored in SUPER[14]. For multi-institutional data sharing, the New York City Clinical Data Research Network (NYC-CDRN), which receives funding from the Patient-Centered Outcomes Research Initiative (PCORI), enables queries of citywide EHR data; the SUPER infrastructure enables WCM to contribute data to the effort. Use of SUPER to support additional multi-institutional data sharing initiatives for clinical trial recruitment, including TriNetX[15] and the National Center for the Advancement

of Translational Science's Accrual for Clinical Trials program (NCATS ACT)[16], is underway. SUPER also previously enabled use of RexDB from Prometheus Research[17].

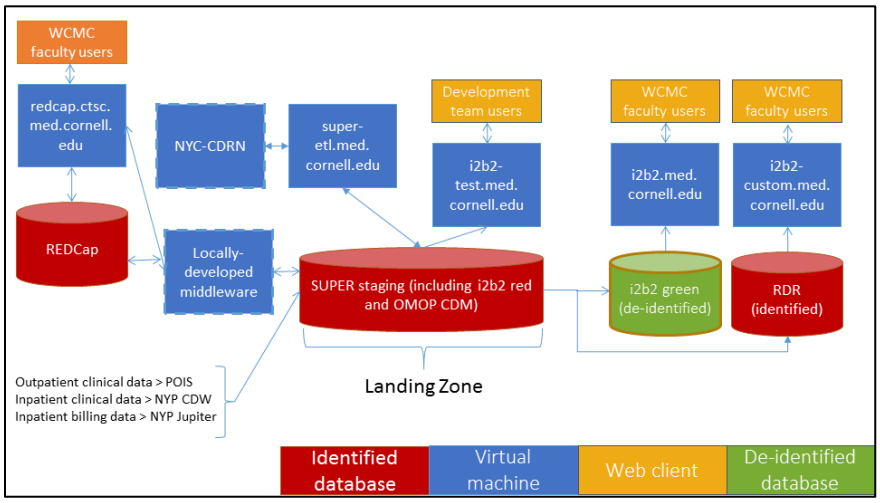

Figure 2. Model of data flow from sources to WCM SUPER staging environment.

Discussion

To our knowledge, this is the first complete description of a methodology for electronic patient data acquisition and provisioning that ignores data harmonization at the time of initial storage in favor of downstream transformation to address specific research questions and applications. While other institutions have outlined approaches that integrate data from different sources in their raw format and transform them on an ad hoc basis[18], the SUPER approach expands previous domain-specific work into the complex realm of health informatics. SUPER has successfully enabled multiple tools that provide utility to the WCM research enterprise, including cohort discovery, electronic data capture, and robust querying. In contrast to centralized data warehouses, the SUPER approach can enable institutions to tailor solutions that meet investigator needs rather than focus on dimensional modeling and other technology-oriented activities.

In characterizing SUPER's methodology and infrastructure, the question naturally emerged: what is SUPER? Popular terms include data warehouse, data mart, and data lake. A data warehouse aims to render organizational data easily accessible in a credible and consistent fashion, adapt to shifting methodologies and use cases, secure sensitive data elements, drive enhancements in decision making, and serve as a usable and accessible resource to end-users[19]. The Mayo Clinic's Enterprise Data Trust (EDT), a "top-down, subject-oriented, integrated, time-variant, and non-volatile collection of data in support of Mayo Clinic's analytic and decision-making processes," serves as an example of the application of the data warehouse model within the healthcare setting. The EDT seeks to normalize data from disparate sources, implementing "consistent information models" and shared vocabulary to enable queries that touch data elements enterprise-wide[1]. Other institutions have also pursued centralized data warehouses with controlled terminology and modeling rules[3,4,6,7]. While the clinical data warehouse is defined by efforts to conduct semantic harmonization and data structuring during the ETL process from source systems, SUPER does not conduct any data harmonization, instead relying on efforts already put in place by teams maintaining source data.

An alternate approach to the data warehouse is the data mart – a specifically tailored data set specific to one research question, scientific workflow, or operational/quality improvement requirement[20]. Often extracted from enterprise-wide data repositories, data marts offer an agile approach that is tailored directly to a specific use case. Purpose-built, they offer greater potential utility at the cost of flexibility and effort – each data mart often requires iterative definitional work with a business analyst to determine the specific requirements and often cannot be reused to support research questions outside of its bailiwick. Data marts also come with a particular suite of risks, including redundancy of effort of aggregation, inconsistency between marts, and difficulty connecting results with larger data sets[20]. Although SUPER enables the creation of data marts, it also constitutes a more holistic effort to aggregate data across the entire institution and is not confined to individual research questions.

A third and relatively novel dimension for considering approaches to enterprise data aggregation and integration is that of the data lake. Coined by James Dixon, the data lake refers to an approach by which institutions aggregate

enterprise data in their native formats and store them unaltered[21]. Analysts and developers then configure analytical components and export data to respond to individual queries and use cases as they emerge, working with users on an *ad hoc* basis. The data lake is not mutually exclusive with the data mart – indeed, they are intrinsically related, as the data lake relies on the data mart for the analysis and dissemination of the data that feeds its "aquifer." SUPER seems to most closely fit the definition of a data lake, as it is characterized by its avoidance of data harmonization efforts during the intake process. However, its reliance on terminology management and semantic restructuring by groups maintaining source data feeds distinguishes it from a data lake in the strictest sense of the word. SUPER is, perhaps, comparable to a "data kitchen," which regularly receives shipments of ingredients from vendors – some processed and some in their raw state – and uses these ingredients to "cook" an array of meals, some simple and designed to satisfy large numbers of people, and some complex, prepared for a specialized audience with particular tastes.

Our implementation of SUPER has yielded a number of pertinent lessons both technical and organizational. From a technical standpoint, SUPER's ETL jobs are notable for their exhaustive resource use and lengthy duration. Our experience shows the value of developing large ETL jobs in a modular fashion and running them piecewise. While a large transaction log can have substantial utility, this approach offers greater benefit to the overall workflows the system facilitates and enables rapid recovery in the event of an individual component's failure. Maximizing preprocessing is also crucial, with indexes configured to help as much of the process as possible. Additionally, retaining previous iterations of data sets on a development server creates a constant backup for operational use, obviating the need to ensure that the landing server is constantly available. While teams responsible for maintaining clinical and billing source systems may be initially reluctant to establish a workflow for transferring raw clinical tables, the SUPER approach offloads much of the iterative workflow in determining data flow and conducting tailored exports from source system teams onto RI. Investigators desirous of using clinical data for research may also be unprepared for the ramifications of its nature – underlining that data are accessioned from multiple sources and may not always be in agreement is crucial in setting expectations for the utility and quality of data delivered, especially for research use cases where data cleanliness is paramount.

Some limitations apply to this analysis. The study was conducted at a single site and may not generalize to other institutions. However, it may be potentially useful to sites with existing clinical data warehouses or sites considering their adaptation. Furthermore, the lack of terminology management may limit our approach: however, in a post-meaningful use healthcare setting, this may not be as applicable.

To our knowledge, this is the first comprehensive description of a methodology and infrastructure for secondary use of patient electronic data for clinical research that stores data from source systems in their raw format to serve specific investigator needs. We feel that this methodology possesses unique advantages and that it may be of use to comparable organizations considering potential approaches for secondary use of patient electronic data for research.

Acknowledgements

This study received support from NewYork-Presbyterian Hospital (NYPH) and Weill Cornell Medical College (WCMC), including the Clinical and Translational Science Center (CTSC) (UL1 TR000457) and Joint Clinical Trials Office (JCTO). We thank Rajesh Bollapragada for help in conceptualizing the institutional setting, and Monika Ahuja for developing and describing some of the methodologies that support SUPER. We also thank the members of the Research Informatics team—Brant Lai, Julian Schwartz, Sean Pompea, David Kraemer, Marcos Davila, Jacob Weiser, Prakash Adekkanattu, and Steven Flores—as well as Project Management Office colleagues Cindy Chen and Anthony DiFazio.

References

1. 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 doi: 10.1136/jamia.2009.002691
2. Danciu I, Cowan JD, Basford M, et al. Secondary use of clinical data: the Vanderbilt approach. J Biomed Inform 2014;**52**:28-35 doi: 10.1016/j.jbi.2014.02.003
3. Kamal J, Liu J, Ostrander M, et al. Information Warehouse – A Comprehensive Informatics Platform for Business, Clinical, and Research Applications. AMIA Annu Symp Proc, 2010:452-6.
4. Lowe HJ, Ferris TA, Hernandez PM, Weber SC. STRIDE – An Integrated Standards-Based Translational Research Informatics Platform. AMIA Annu Symp Proc, 2009:391-5.
5. Cimino JJ, Johnson SB, Hripcsak G, Hill CL, Clayton PD. Managing vocabulary for a centralized clinical system. Medinfo 1995;**8 Pt 1**:117-20

6. 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-i102 doi: 10.1136/amiajnl-2011-000339
7. Wilcox AB, Vawdrey DK, Chen YH, Forman B, Hripcsak G. The evolving use of a clinical data repository: facilitating data access within an electronic medical record. AMIA Annu Symp Proc 2009;**2009**:701-5
8. Campion TR Jr, Blau VL, Brown SW, Iscovich D, Cole CL. Implementing a Clinical Research Management System: One Institution's Successful Approach Following Previous Failures. AMIA Jt Summits Transl Sci Proc. 2014 Apr 7;2014:12-7. eCollection 2014.
9. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009;**42**(2):377-81 doi: 10.1016/j.jbi.2008.08.010
10. 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. Epub 2011 Oct 22.
11. Abend A, Housman D, Johnson B. Integrating Clinical Data into the i2b2 Repository. Summit on Translat Bioinforma, 2009:1-5.
12. 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 doi: 10.1136/jamia.2009.000893
13. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc 2012;**19**(1):54-60 doi: 10.1136/amiajnl-2011-000376
14. Campion TR Jr, 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. 2017:76-81. eCollection 2017.
15. TriNetX. TriNetX Health Data Network [Internet]. Cambridge, MA: TriNetX, Inc; [cited 2017 Jul 3]. Available from: http://www.trinetx.com
16. National Center for Advancing Translational Sciences (US). CTSA Consortium Tackling Clinical Trial Recruitment Roadbloacks [Internet]. Bethesda (MD): US Department of Health and Human Services, National Institutes of Health; [cited 2017 Jul 3]. Available from https://ncats.nih.gov/pubs/features/ctsa-act.
17. Prometheus Research. Research Data Platform | Clinical Data Management System [Internet]. New Haven, CT: Prometheus Research, LLC; [cited 2017 Jul 3]. Available from: https://www.prometheusresearch.com/.
18. Kozhenkov S, Sedova M, Dubinina Y, et al. BiologicalNetworks--tools enabling the integration of multi-scale data for the host-pathogen studies. BMC Syst Biol 2011;**5**:7 doi: 10.1186/1752-0509-5-7
19. Kimball R, Ross M. *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*: Wiley Publishing, 2013.
20. Inmon WH. Data mart does not equal data warehouse. Data Manag Rev (1998)
21. Woods D. Big Data Requires a Big, New Architecture. Forbes [Internet]. 2011 [cited 2017 Jul 3]. Available from: https://www.forbes.com/sites/ciocentral/2011/07/21/big-data-requires-a-big-new-architecture/