Journal of Biomedical Informatics xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Secondary use of clinical data: The Vanderbilt approach

Ioana Danciu [a,b,*], James D. Cowan [a], Melissa Basford [a], Xiaoming Wang [a], Alexander Saip [a], Susan Osgood [a], Jana Shirey-Rice [a], Jacqueline Kirby [a], Paul A. Harris [a,b]

[a] Vanderbilt Institute for Clinical and Translational Research, Vanderbilt University Medical Center, 2525 West End Ave., 6th Floor, Nashville, TN 37203, United States
[b] Department of Biomedical Informatics, Vanderbilt University Medical Center, 2209 Garland Ave., Nashville, TN 37232, United States

ARTICLE INFO

Article history:
Received 21 July 2013
Accepted 4 February 2014
Available online xxxx

Keywords:
Biomedical informatics
Secondary use of clinical data
Research data warehouse
Research enterprise

ABSTRACT

The last decade has seen an exponential growth in the quantity of clinical data collected nationwide, triggering an increase in opportunities to reuse the data for biomedical research. The Vanderbilt research data warehouse framework consists of identified and de-identified clinical data repositories, fee-for-service custom services, and tools built atop the data layer to assist researchers across the enterprise. Providing resources dedicated to research initiatives benefits not only the research community, but also clinicians, patients and institutional leadership. This work provides a summary of our approach in the secondary use of clinical data for research domain, including a description of key components and a list of lessons learned, designed to assist others assembling similar services and infrastructure.

1. Introduction

Over the last decade, the transition from paper medical records to electronic clinical systems has been accelerated by a national emphasis on modernizing our health care infrastructure. Legislative initiatives, such as the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 [1], which includes monetary incentives (and ultimately penalties), requires providers to show "meaningful use" of certified electronic health records (EHRs). This resulted in a significant growth in the amount of clinical data being collected. The transition from paper to electronic clinical systems has also created new opportunities for secondary use of clinical data in biomedical research. Rapid cohort identification, quality of care assessment, comparative effectiveness research, data privacy and de-/re-identification research, phenotyping methodology and predictive modeling represent a handful of areas where ready access to clinical data for research endeavors is beginning to make a real impact at academic medical centers across the country.

In 2006, responding to national trends, the American Medical Informatics Association (AMIA) compiled a set of recommendations [2] that defined challenges and stressed benefits of research-driven secondary use of clinical data. MacKenzie et al. [3] surveyed 35 Clinical and Translational Science Award (CTSA) organizations and the NIH Clinical Center in 2008 and 2010, reporting a positive trend for institutional development, management and utilization of integrated data repositories to support the research enterprise. Primary obstacles reported in the 2010 survey included data quality and standards issues related to assembling a common repository from multiple systems, sustainable funding to support infrastructure and operations, and meaningful data access services provided to research teams. In 2012, Murphy et al. [4] surveyed 17 institutions and observed a significant increase in the clinical repositories used for research since 2007. In 2013, Embi et al. [5] surveyed clinical research informatics (CRI) papers published in scientific journal and conference proceedings from 2009 to 2013 and observed six common themes: (1) clinical data reuse for research; (2) data and knowledge management, discovery and standards; (3) researcher support and resources; (4) participant recruitment; (5) patients/consumers and CRI; and (6) policy, regulatory and fiscal matters. Large-scale, integrated data repositories are foundational for work in many of these areas, resulting in a growing number of academic medical centers assembling big data programs to support the local research enterprise. Examples include Intermountain Healthcare [6], Massachusetts General Hospital [7], the Mayo Clinic [8], Columbia University Medical Center [9] and Stanford Medical Center [10]. Data exploration tools such as i2b2 [11] and Harvest [12] have been designed to directly support researcher data inquiry needs, though a combination of tools and human expert support are typically needed for optimal enterprise-wide researcher support.

* Corresponding author at: Vanderbilt Institute for Clinical and Translational Research, Vanderbilt University Medical Center, 2525 West End Ave., 6th Floor, Nashville, TN 37203, United States.
E-mail address: Ioana.danciu@vanderbilt.edu (I. Danciu).

http://dx.doi.org/10.1016/j.jbi.2014.02.003
1532-0464/

Beginning in the early 1990s, Vanderbilt University Medical Center (VUMC) began a series of clinical informatics initiatives [13,14] resulting largely in the elimination of paper medical records by 2004 [15]. Vanderbilt's current clinical framework consists of a variety of software systems, both off-the-shelf commercial solutions and applications developed in house. A centralized transactional messaging engine called the Generic Interface Engine (GIE) manages communication and information exchange between systems. This early adoption and integration of electronic clinical information systems have had significant impact in the domains of clinical care, patient safety, provider accountability, and improved documentation [16–19]. The end result of our early launch and continuously evolving clinical systems is an information-rich environment covering 2 million patients, with longitudinal records spanning more than a decade.

For a long time at Vanderbilt, the EHR system (StarPanel) and an enterprise data warehouse (EDW) were the two main repositories of clinical and billing data. The need for a dedicated research framework emerged because common data repositories represent only half of the solution. Researchers need secure and reliable access to data programmers and/or self-service tools to query data, and must understand the meaning and structure of data elements to avoid making naïve assumptions. This paper provides a description of Vanderbilt's approach to secondary use of clinical data and presents a set of practical "lessons learned" that could prove useful for other institutions considering assembling similar infrastructure and data access services.

2. Methods

2.1. Overview

The Office of Research Informatics (ORI) leads Vanderbilt initiatives involving the secondary reuse of clinical data for research. Working with faculty across the Vanderbilt Departments of Biomedical Informatics (DBMI) and Biostatistics, ORI contributes regularly to the support of new methods development. By providing data and infrastructure support to nationally recognized research initiatives in natural language processing (NLP) [20], privacy and security [21], data mining and pattern discovery based on probabilistic machine learning [22], and personalized medicine [23], ORI contributes to building and refining enterprise systems that rapidly inform and improve upstream clinical enterprise processes.

The ORI research support enterprise can be loosely described as a centralized collection of tools and services that are available to all research teams. We use an iterative model for tool development where the lessons learned are constantly and rapidly incorporated as new functionality. The hierarchical evolution enables tool support to be made available to research teams at no cost. Services, both technical and administrative, are available to research teams at low cost under a fee-for-service pricing model with billing through a centralized Vanderbilt Core Ordering and Reporting Enterprise System (CORES) [24]. Leveraging tools and services in an equitable fashion and asserting a fee-for-service model allows us to satisfy researchers' need for data access and to provide sustainable funding for the research enterprise, two of the main obstacles observed by MacKenzie et al. [3] at other institutions. Tools and service-level requests rely on a large-scale research data warehouse, which we have assembled through connectivity with our EDW team, ancillary clinical care systems throughout the medical center, registries, and other peripheral support systems. Since we depend on other systems for data, we face similar challenges as other institutions with regard to information quality and standards [3]. To mitigate these hindrances as much as possible we process, transform, organize and optimize our data for use across multiple tools and platforms. We maintain a fully de-identified research data warehouse called the Synthetic Derivate (SD) and a fully identified research data warehouse called the Research Derivative (RD). The RD can be thought of as a mirror of all of the clinical data collected at VUMC, but organized for research. The SD represents a "de-identified" version of the RD and is linked to an anonymized DNA biobank (BioVU) [25]. Fig. 1 below shows an overview of the clinical and research informatics environment at Vanderbilt. The specific sections (marked A–F) will be described in more detail.

2.2. Clinical enterprise

The epicenter of Vanderbilt's network of clinical information systems (Fig. 1 Sec. A) is a modern, web-based, EHR interface called StarPanel [15]. StarPanel facilitates clinical note generation, provider and user communication, and integrates patient specific data, in real-time, from a variety of clinical care systems such as the laboratory information systems, the inpatient registration system, the provider order entry system, a nursing documentation system, a barcode medication administration system, and various other ancillary systems like anesthesiology, cardiology, radiology, and trauma. While StarPanel has been carefully architected for rapid response time, and is designed to support the daily workflow of clinical care teams, it is not well suited for efficiently querying or extracting data across populations of patients. This limits its usefulness for research applications. In addition to StarPanel, most clinical data within Vanderbilt clinical systems are captured and stored in an EDW. As is the case with many institutional EDWs, data capture and organization is largely driven by business intelligence/reporting needs and long-term preservation goals. These business-driven architectures are usually not designed for supporting large research communities. EDW leaders and professional support personnel are also typically more concerned with institutional program goals (e.g. large-scale quality initiatives) than supporting individual research projects. Access and utilization of EDW data sources by independent research teams can be challenging without expert guidance.

2.3. Research data warehouse: the identified data layer (RD)

The RD (Fig. 1 Sec. B) is a database of clinical and administrative data that is well suited for research, quality improvement, and institutional projects requiring rapid, efficient extraction of clinical data on a defined cohort using specific tests or phenotypes as inclusion criteria to deliver identified datasets, recurring reports, and up-to-date counts of subjects meeting the inclusion criteria. The bulk of structured clinical data comes into the RD daily via the EDW, which has well established Extract, Transform and Load (ETL) pipelines from multiple sources of patient registration, clinical and billing information. In many cases, though, the EDW storage mirrors the production databases of the source systems, resulting in both record attribute redundancy and value limitations from a clinical perspective. To address this issue, we created our own ETL layer. The RD uses the same coding schemes used by the VUMC clinical systems and is an aggregation of different standards. As such, structured medication information uses the First Databank (FDB) coding standard, diagnoses use the International Classification of Diseases (ICD-9), and medical services and procedures use the Current Procedural Terminology (CPT). Our commercial laboratory information management system (LIMS) uses 2 letter combinations for lab codes, which our StarPanel EHR then maps to VUMC specific lab short names, thus allowing flexibility in situations where source vendor systems are replaced. The RD system uses laboratory short names as identifiers. Electronic notes and reports known as StarDocuments are stored as unstructured data via plain text documents. All note data are assembled in the

Please cite this article in press as: Danciu I et al. Secondary use of clinical data: The Vanderbilt approach. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.02.003