Secondary use of clinical data: The Vanderbilt approach

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

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

\begin{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.
\end{abstract}

\section{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\textsuperscript{[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\textsuperscript{[2]} that defined challenges and stressed benefits of research-driven secondary use of clinical data. MacKenzie et al.\textsuperscript{[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.\textsuperscript{[4]} surveyed 17 institutions and observed a significant increase in the clinical repositories used for research since 2007. In 2013, Embi et al.\textsuperscript{[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\textsuperscript{[6]}, Massachusetts General Hospital\textsuperscript{[7]}, the Mayo Clinic\textsuperscript{[8]}, Columbia University Medical Center\textsuperscript{[9]} and Stanford Medical Center\textsuperscript{[10]}. Data exploration tools such as i2b2\textsuperscript{[11]} and Harvest\textsuperscript{[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.
Beginning in the early 1990s, Vanderbilt University Medical Center (VUMC) began a series of clinical informatics initiatives \cite{13,14} resulting largely in the elimination of paper medical records by 2004 \cite{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 \cite{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) \cite{20}, privacy and security \cite{21}, data mining and pattern discovery based on probabilistic machine learning \cite{22}, and personalized medicine \cite{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) \cite{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. \cite{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 \cite{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) \cite{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 \cite{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 anesthesia, 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
RD from StarPanel, with important metadata retained (e.g. medical record number, timestamp, document type, and document name). Documents are easily mined using simple keyword or regular expression searches. We use MedEx [26] for medication extraction from past medical histories, medication lists, prescriptions and refills, and problem lists. Semi-structured data are more granular and generally consist of multiple-answer forms, known in StarPanel as StarForms. Data from StarForms are parsed and stored in the RD as key-value pairs. More information on clinical documentation types available within Vanderbilt clinical systems is available elsewhere [19,27]. StarDocuments and StarForms are loaded weekly in the RD, after ORI-constructed programs perform careful cleaning procedures (e.g. removing html tags, extracting clinical note sections such as family history and problem lists, and removing duplicated elements). The RD database is stored on a secure IBM Netezza 1000-24 [28] database server housed in the Vanderbilt Data Center. The database is fully compliant with the administrative, physical, and technical provisions of the Health Insurance Portability and Accountability Act of 1996 (HIPAA) Security and Privacy Rules, and operates with oversight from the Vanderbilt Institutional Review Board (IRB).

2.4. Research data warehouse: the de-identified data layer (SD)

Many secondary use research projects can be performed in a de-identified environment, and in so doing minimize privacy risks to individuals. By removing potential identifiers (defined in the safe-harbor provision of HIPAA [29]) before storing in a similar data architecture to the RD, we have established the SD as a separate research-ready data warehouse (Fig. 1 Sec. C). De-identification methods include: (a) medical record numbers (MRNs) are replaced by research unique identifiers, generated by a one-way hash from the MRN; (b) dates are shifted backwards by a constant within each patient, a deterministically calculated number of days between 0 and 364, obfuscating the true service dates while preserving the time dependence between service dates, and (c) free text notes are stripped of the remaining protected health information (PHI) identifiers via De-ID [30], a product of Data Safety Software. Because data are de-identified, research protocols using only the SD as a data source are often evaluated as non-human subject protocols by the IRB. Nevertheless, access to the SD requires researchers to provide evidence of IRB approval or determination and a signed data use agreement. The SD is linked to an anonymized DNA biobank (BioVU) [25], which includes de-identified genetic biosamples (DNA) collected from leftover clinical blood samples. Genotyping data obtained from a variety of platforms (e.g. Infinium HumanExome BeadChip, HumanOmni5-Quad, and Human1M). The combination of DNA samples linked to de-identified clinical data provides a powerful resource for genomic research [31–35], while the SD alone can support a substantial breadth of research; the only substantial limitation includes time-sensitive epidemiological studies [36,37]. Dates are shifted in the SD as part of the de-identification process, so although time intervals within each medical record are preserved, cohort data cannot be linked to temporal events such as epidemics or natural disasters. The SD and BioVU methodologies have previously been described in detail [25].

2.5. Supporting informatics and biostatistical methods development

Building and maintaining large-scale, research-oriented data repositories requires methods and experts from numerous specialty domains. Data architecture and technical infrastructure specialists, medical experts, and clinical workflow experts (Fig. 1 Sec. D) are all needed to ensure proper contextual understanding of the data retrieved and stored in the research data warehouse. Informatics professionals with conceptual and applied knowledge of natural language processing are required to turn unstructured clinic notes and medication orders into structured information that can be easily stored and readily queried, through initiatives like MedEx [26]. For de-identified data warehousing initiatives like the SD, informaticians and computer scientists with knowledge and understanding of patient de-identification and re-identification risks are critical. Privacy and ethics experts are critical for input and assistance with governance and policy [21,37–41]. Finally, predictive modeling and machine learning experts are required to convert data into actionable information to feed forward into research protocols or into improvements in workflow within the clinical enterprise. Vanderbilt has strong informatics and biostatistics departments, resulting in availability of faculty with needs for
large data infrastructure, to support these sorts of methods research [27,42]. These relationships are synergistic. In exchange for access to large datasets required to build and maintain their personal research programs, these experts are happy to participate in SD/RC formative planning and iterative exercises designed to build and improve data warehousing services [20,21,38,42,43].

2.6. Translational use for the clinical enterprise

Knowledge dissemination and translation into clinical practice are the ultimate goals of informatics and biostatistical research. Initiatives supported by ORI (Fig. 1 Sec. E) have eventually become standard practice in clinical workflow. The Pharmacogenomic Resource for Enhanced Decisions in Care and Treatment (PREDICT) [23] is a prominent personalized medicine initiative that first leveraged BioVU genotyping and SD phenotyping information to facilitate predictive modeling that would ultimately change prescribing practices at the bedside. Quality improvement initiatives at VUMC have utilized the RD to efficiently measure patient outcomes resulting in the implementation of new clinical care processes. The most recent effort is an acute kidney injury (AKI) initiative that leverages the RD to identify risk factors for AKI, and implement new prescribing workflows in the computerized physician order entry system.

2.7. The research enterprise

Building comprehensive research data warehousing assets is a formidable task, but of little value to the larger research community unless accessible in an equitable and approachable manner. Our general approach (Fig. 1 Sec. F, Fig. 2) is to build: (1) self-service tools available at no – or low – cost for researchers; and (2) customized tools and data extraction services using a fee-for-service agreement with researchers to sponsor ORI programmers when existing self-service tools are not adequate to fulfill complex use cases. In working with the researcher-sponsored complex use cases, ORI programmers compile lessons learned and review methods that can later be abstracted and fed back into the no – or low – cost offerings.

Two web-based applications provide researchers with “self-service” access to the de-identified data warehouse: the Record Counter user interface (RC UI) and the SD user interface (SD UI). The SD UI and RC UI share common query functionality for data elements such as labs, medications, vitals, codes (ICD-9 and CPT) and registries such as the tumor registry. Users are able to generate ad hoc queries and view approximate record counts in real-time. The primary difference between the two applications is that the RC UI allows users to view only aggregated record counts (grouped by age, race, and sex), while the SD UI provides aggregated counts and the ability to save result sets, allowing review of record level data and requests for DNA data from BioVU. Results obtained via the SD UI can be exported in text (.txt) format for upload into biostatistics programs or can be sent directly to an automatically generated REDCap data mart. REDCap is web application that facilitates survey and database management in a secure manner and has been described in detail in [44]. The SD UI can be accessed only by investigators who have completed an approval process and received IRB permissions for use in their specific study and who have signed a data use agreement. The RC UI is accessible to the entire Vanderbilt community through our StarBRITE research portal [45] at no cost, and is often used for hypothesis generation, study planning and feasibility consideration.

Subject Locator is a web-based “self-service” tool that leverages the RD to facilitate identification of patients in the clinical enterprise who meet basic requirements for study inclusion in prospective studies and trials. Here, researchers leverage a tool very similar to the RC query generator to select specific criteria (e.g. labs, medications, vitals, ICD-9 and CPT codes) for use as rough inclusion/exclusion criteria for their study. Researchers also specify a set of clinics most likely to be effective for recruiting patients, based on medical conditions and a priori relationships with clinical providers. Study criteria and visit information is cross-referenced on a daily basis. The resulting list is provided to research teams for prospective consideration of new participants. Within the application, users manage lists of prospective subjects as they are investigated and potentially contacted by research teams. Access to Subject Locator is available to research teams working on individual projects that obtain IRB approval to use the tool. Furthermore, Subject Locator includes automated checks to ensure individual end-users have “key study personnel” designation within the Vanderbilt IRB system for the particular study.

Researchers interface with the RD primarily through the Vanderbilt Data Coordinating Center (VDCC) by way of fee-for-service cohort extraction, custom algorithm development and project management. A project manager verifies research study IRB status, and secures signed data use agreements as needed, and then works with the RD engineers and investigators to refine cohort definitions and study design. To meet the needs of the project, the RD team creates the dataset and delivers it to the requester in a secure and agreed upon format using either a secure transfer server or by uploading it into the requester’s REDCap project. A detailed flow diagram of the RD intake process and delivery is presented in Fig. 3 below.

3. Value proposition

At Vanderbilt, the usefulness of a dedicated infrastructure for supporting the secondary use of clinical data for research is becoming increasingly apparent for various stakeholders across the enterprise.
Researchers have data, tools and expertise available to them irrespective of their seniority level. Since its inception, the SD has provided infrastructure and support for various research projects that resulted in 87 papers published between 2010 and 2013 [21,23,26,27,32–37,39–42,45–115], with 61 first authors from 18 different departments. The diverse range of topics studied and reported in these papers (e.g. quality improvement, genetic associations (BioVU), pharmacogenomics, NLP, privacy, ethics, general informatics methods research) illustrates the utility of the SD for research purposes. The RD is a more recently developed resource, and even though the research projects it has supported have not yet matured into publications, it shows encouraging trends in usage. Between September 2010 and August 2013, we received 102 requests for data from 72 different investigators from 21 different departments.

Patients can benefit greatly from a large research data warehousing program through initiatives like personalized medicine, clinical trial opportunities, and quality improvement programs that rapidly translate research into practice to transform clinical care. Centralized data warehousing resources also enable tighter data control and enhanced security and privacy of patient data.

For clinicians, translational projects drive opportunities for optimization of workflow with just-in-time information, enabling evolution of clinical practice through initiatives like personalized medicine.

In a strong research environment, with data and tools readily available, institutional leadership benefits from increased grants and contracts. Strong institutional assets also provide an advantage in recruiting new faculty and in enhancement of reputation for the center.

Finally, quality improvement and research often have similar needs for longitudinal data to create 360-degree views of single patients, as well as cohort identification; having a multi-purpose resource with a lean cost structure benefits both endeavors.

4. Lessons learned

We present below a series of lessons learned during the process of creating infrastructure, policy and organizational support for research data warehousing which are closely in line with the CRI trends observed by Embi et al. [5].

4.1. Leveraging clinical enterprise data sources

The primary role of clinical users is caring for patients, and technology must support and complement this mission [116]. As a direct consequence, the resulting data might be incomplete from a research standpoint, in different formats or missing altogether, and need to undergo a careful cleanup and transformation process before they can be used for research. Incompleteness, inconsistency, and inaccuracy are major challenges also observed at other institutions [9,117] and in industry [118]. Understanding the clinical significance of the data and the way they are coded in clinical settings is a major and necessary task in reusing clinical data. Ultimately, it is the responsibility of the research enterprise to process and maintain data in a meaningful and scalable manner.

4.2. Research data warehouse framework

In an environment where grant funding is increasingly competitive, researchers need quick, reliable and reproducible cohort identification mechanisms and the capacity for retrospective clinical research. Regulatory issues and policy around access to clinical data are often complex, and investigators sometimes need assistance with understanding privacy requirements. The availability of research and informatics support teams meets this need and eases investigator burden by providing specific data expertise and extraction skills. Data sharing also facilitates the development and validation of informatics methods including NLP, de-identification approaches and large-scale phenotyping. Collaborative initiatives such as the Electronic Medical Records and Genomics (eMERGE) initiative [66], the Strategic Health IT Advanced Research Projects (SHARPn) framework [7] or the Stanford Translational Research Integrated Database Environment (STRIDE) platform [10] promote sharing and testing of new methodologies in diverse clinical populations and environments, which ultimately strengthens the national research informatics enterprise.

4.3. Supporting informatics and biostatistical methods development

The anatomy of the research domain at leading academic medical centers is extremely complex, and successful projects often require a diverse range of expertise during different phases of their life cycles. While rigorous and organized documentation of the different methods, algorithms and data is a necessity, the research enterprise relies heavily on human experts to help build these repositories and ultimately advance the research enterprise. Developing and maintaining a team with deep clinical and research domain knowledge across the vast array of areas of study, though necessary, present significant challenges.

4.4. The research enterprise

Self-service tools are a viable solution for the research enterprise because they effectively scale for increasing data demand. The ultimate measure of utility of these tools is the ability to efficiently complete studies and projects as well as user satisfaction. Establishing and engaging a user group community for software tools early in the development process is ideal. User groups are invaluable to all the parties involved for face-to-face help, questions, instruction, and feedback. Building confidence and understanding of both the data and self-service tools empowers the end user, and relieves pressure on the internal support team. Building such a community after years into the project is not an easy task, but once the research community has become familiar with the tools available, they will generate an increasing demand and requests for more functionality. Visualization and graphical tools breed ideas and creativity to assist investigators in understanding and seeing trends within the data.

Human-support is also necessary for assisting investigators in a number of ways. The typical researcher from an academic medical department may not have sufficient awareness of institutional data to formulate a question that can be easily translated into an actionable dataset. Data sharing also facilitates the development and validation of informatics methods including NLP, de-identification approaches and large-scale phenotyping. Collaborative initiatives such as the Electronic Medical Records and Genomics (eMERGE) initiative [66], the Strategic Health IT Advanced Research Projects (SHARPn) framework [7] or the Stanford Translational Research Integrated Database Environment (STRIDE) platform [10] promote sharing and testing of new methodologies in diverse clinical populations and environments, which ultimately strengthens the national research informatics enterprise.

5. Conclusion

Secondary use of clinical data can play a critical role at large academic medical centers. Building a dedicated research infrastructure
at Vanderbilt has enabled us to better serve the research community, advance informatics and biostatistical methods development, and ultimately use evidence-based results to change clinical practice.

Acknowledgments

We gratefully acknowledge assistance from the StarPanel team led by Dr. Ing. Dario Giuse; the Enterprise Data Warehouse, lead architect Eric Griffin; the Biomedical Language Processing Lab, led by Dr. Joshua Denny; the Health Information Privacy Lab led by Dr. Brad Malin; and administrative and technical groups within the Vanderbilt Institutional Review Board.

References
