Journal of Biomedical Informatics 44 (2011) 266–276

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

The DEDUCE Guided Query tool: Providing simplified access to clinical data for research and quality improvement

Monica M. Horvath [a,*], Stephanie Winfield [a], Steve Evans [a], Steve Slopek [a], Howard Shang [a], Jeffrey Ferranti [a,b]

[a] Duke Health Technology Solutions, Duke University Health System, Durham, NC, United States
[b] Department of Pediatrics, Duke University School of Medicine, Durham, NC, United States

ARTICLE INFO

Article history:
Received 19 February 2010
Available online 2 December 2010

Keywords:
Translational research
Medical informatics
Clinical informatics
Medical records systems
Computerized health care evaluation mechanisms
Hospital information systems

ABSTRACT

In many healthcare organizations, comparative effectiveness research and quality improvement (QI) investigations are hampered by a lack of access to data created as a byproduct of patient care. Data collection often hinges upon either manual chart review or *ad hoc* requests to technical experts who support legacy clinical systems. In order to facilitate this needed capacity for data exploration at our institution (Duke University Health System), we have designed and deployed a robust Web application for cohort identification and data extraction—the Duke Enterprise Data Unified Content Explorer (DEDUCE). DEDUCE is envisioned as a simple, web-based environment that allows investigators access to administrative, financial, and clinical information generated during patient care. By using business intelligence tools to create a view into Duke Medicine's enterprise data warehouse, DEDUCE provides a Guided Query functionality using a wizard-like interface that lets users filter through millions of clinical records, explore aggregate reports, and, export extracts. Researchers and QI specialists can obtain detailed patient- and observation-level extracts without needing to understand structured query language or the underlying database model. Developers designing such tools must devote sufficient training and develop application safeguards to ensure that patient-centered clinical researchers understand when observation-level extracts should be used. This may mitigate the risk of data being misunderstood and consequently used in an improper fashion.

1. Introduction

Although the adoption of health information technologies (HIT) such as electronic health records (EHRs) and computerized physician order entry (CPOE) has been identified as critical for improving the nation's health care [1–3], the invaluable clinical data gathered by HIT, which could spur research to an unprecedented degree, has often gone untapped [4]. The Recovery Act of 2009 included $19.2 billion in funding intended to encourage physicians and healthcare organizations to implement EHRs and make "meaningful use" of the collected data by exchanging information and reporting clinical quality measures [5]. An additional $1.1 billion in funding will be administered through the US Department of Health and Human Services. These grants will support comparative effectiveness research evaluating methods used to diagnose, treat, and monitor clinical conditions [6].

But despite clear imperatives for making information in health system databases accessible for secondary data analysis [7,8], a mature approach to combining, analyzing, and leveraging these resources has yet to emerge. Within many organizations, data reside in separate silos or in proprietary databases and are often captured in incompatible formats. Even when laborious manual queries are used, the resulting extracts are often incomplete because the source systems were not designed with domain-spanning research in mind. Such functionality requires the adoption of structure and standards [9,10].

In this manuscript, we report on a "research portal" developed by the Duke University Health System (DUHS)—the Duke Enterprise Data Unified Content Explorer (DEDUCE). This user-friendly data extraction system is envisioned as a multiple-tool environment that will expedite access to clinical data stored in the organizational data warehouse, supporting grant applications, research projects, and quality improvement (QI) activities. Here we report on the conceptual design and development of the first tool in DEDUCE, Guided Query (GQ), which uses business intelligence (BI) tools to allow users to obtain both an aggregate report and a raw data extract based on query parameters.

2. Background

Research portals developed in response to the need to access and combine diverse sources of data from clinical and research

* Corresponding author at: Duke Health Technology Solutions, 2424 Erwin Road, DUMC 2718, Durham, NC 27705, United States. Fax: +1 919 668 3241.
E-mail address: monica.horvath@duke.edu (M.M. Horvath).

1532-0464/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2010.11.008

domains, both within and across institutions. Two fundamental portal types are described in the literature: translationally-based and clinical practice-based. Although both include clinical data and provide user interfaces, the goals of the inaugural database designs differ. Translationally-based platforms such as REDCap [11], Slim-Prim [12], MMIM [13], and TraM [8] start with data collected for research purposes (clinical or basic science) and integrate these domains in a user-accessible repository. These tools may use federated queries to derive data for specific disease states across a national set of hospitals, enable data sharing for multicenter translational projects, and create a framework for the input of new research data and subsequent curation [8,13–15]. Clinical practice-based portals, on the other hand, use patient care data from clinic and hospital databases without predefining the research project or domain. The focus in this context is on ensuring that all reasonable data elements regarding a patient's healthcare encounter are standardized and accessible.

While translational portals have been relatively well documented in the literature, there is a notable lack of publications describing the conceptual design, deployment, and operationalized use of clinical practice-based portals. Most of the available information on such portals has been presented in forums such as conferences, proceedings, news articles, or electronic white papers, making a comprehensive discussion of differing feature sets and deployment methodologies challenging. Vanderbilt's Synthetic Derivative research application is one such tool advertised as containing both structured clinical data and care narratives (e.g., nursing notes; surgical reports) on 1.7 million patients, as derived from their health system's EHR [16,17]. Its website (http://www.mc.vanderbilt.edu/victr/pub/message.html?message_id=182) suggests that only de-identified data are available, and a recent conference presentation indicates that ICD9 codes, labs, vital signs, medications, CPT codes, and demographics are available as query criteria. However, the exporting capabilities are unclear and we infer that the tool is designed for the needs of the patient-centric researcher seeking to define a cohort and not optimized for QI personnel looking for "cohorts" of observation-level data (e.g., all lab results of a particular type). Similarly, the Stanford Translational Research Integrated Database Environment (STRIDE) provides self-service research access to a clinical data warehouse that supports two hospitals and numerous clinics [18]. Users can search for patients using criteria including demographics, ICD-9/CPT codes, lab results, pharmacy orders, and information held within narrative clinical reports. STRIDE also provides research access to a tumor tissue databank, thus integrating translational data with its clinical foundation. Yet according to its Web site, STRIDE does not yet release protected health information (PHI) and researchers must collaborate with informatics staff to discuss the extraction of clinical data for research purposes. Based on the only formal report available to date, it is unclear whether STRIDE permits the extraction of observation-level data needed for QI investigation [18].

Partners Healthcare system has published sporadic short reports on its research portal, the Research Patient Data Repository (RPDR), which is designed to aid cohort identification for research studies, support grant applications, and enable outcomes research for two medical centers and four community hospitals [19,20]. This tool has two distinct functions: (1) a query tool that returns aggregate numbers of patients based on complex queries generated from a user-friendly, "drag-and-drop" interface; and (2) a data acquisition tool allowing researchers to obtain detailed extracts including PHI, when authorized by an IRB protocol. Various inpatient and outpatient data elements are available, including demographics, encounter data, diagnoses, medications, procedures, labs, radiology/pathology reports, and discharge notes. However, as with STRIDE, it is unclear the extent to which observation-level data can be extracted independently of patient cohort definition. Recently, some RPDR features were incorporated into SHRINE [21], which uses a federated model to access the clinical databases of three large health centers. The SHRINE prototype, however, functions in a test environment using an enterprise dataset that is not refreshed. SHRINE is one of a growing number of tools that uses the open-source Informatics for Integrating Biology and the Bedside framework (i2b2; http://www.i2b2.org) sponsored by the NIH Roadmap National Centers for Biomedical Computing. This platform bridges clinical and scientific domains by providing open-source software tools for concomitant data collection and management. Aimed at clinical investigators, bioinformaticists, and software developers, i2b2 application modules can be integrated using a variety of Web services and XML messages [14,15,22]. The i2b2 framework has been a fixture at many healthcare informatics and data warehousing conferences where organizations discuss research query tools.

Although these clinically-based research portals offer aggregate counts and raw data, the emphasized goal is to define a highly specific patient cohort that suits the needs of a physician-researcher. However, there are myriad QI questions that require investigation of observation-level data, such as lists of medication or laboratory orders [23], and the query procedure should be designed around these needs. Such investigation will become increasingly important to comply with new "meaningful use" mandates from the Recovery Act [5]. We view the lack of focus on obtaining a specific, defined "cohort" of encounter-, process-, or observation-level data as a major gap in currently reported applications. Our objective in developing DEDUCE was to build an access model that simultaneously served both patient- and encounter-centered needs by creating a user-friendly gateway to various axes of patient care. The DUHS comprises two community hospitals and an academic facility, the Duke University Medical Center (DUMC); the DUMC itself includes a teaching hospital and more than 150 affiliated outpatient clinics. We recognize that in order to serve all user types from these settings, DEDUCE may ultimately require multiple access environments. Since there are relatively few formally published descriptions of how organizations have developed and deployed clinically-based portals, we share here the experiences of the DUHS in developing the underlying DEDUCE framework and releasing our first DEDUCE tool—Guided Query (GQ).

3. Business need and design requirements

The need for this application grew from an increasing number of requests to health system data warehouse personnel for research and QI data extracts. At this point, the Director of the DUHS Data Warehouse Group as well as the Associate Chief Information Officer for Enterprise Analytics and Patient Safety collaborated to create a suite of self-service tools for clinical data extraction. In doing this, an advisory steering committee was convened to shape the functional design (see Section 4.2). At the onset of the design process, the committee designated principal functional requirements for the long-term development of a comprehensive, multiple-environment research portal:

1. DEDUCE should be a flexible, self-service application for data extraction by clinical researchers and QI professionals across the DUHS. DEDUCE will include a number of tools or "user environments" catering to different user levels.
2. DEDUCE should incorporate multiple domains of patient care and be flexible enough to eventually include evolving knowledge domains from the translational arena, including genomics and proteomics. We did not want to predefine research areas, as has been done with translationally-based research portals.