Journal of Biomedical Informatics 52 (2014) 231–242

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Modular design, application architecture, and usage of a self-service model for enterprise data delivery: The Duke Enterprise Data Unified Content Explorer (DEDUCE)

Monica M. Horvath [a,*], Shelley A. Rusincovitch [b], Stephanie Brinson [a], Howard C. Shang [c], Steve Evans [a], Jeffrey M. Ferranti [a,d]

[a] Duke Health Technology Solutions, Duke University Health System, Durham, NC, United States
[b] Duke Translational Medicine Institute, Duke University School of Medicine, Durham, NC, United States
[c] Duke Clinical Research Institute, Duke University School of Medicine, Durham, NC, United States
[d] Department of Pediatrics, Duke University School of Medicine, Durham, NC, United States

ARTICLE INFO

Article history:
Received 23 October 2013
Accepted 7 July 2014
Available online 19 July 2014

Keywords:
Cohort definition
Research query tool
Medical informatics applications
Information systems
Application development
System design and architecture

ABSTRACT

Purpose: Data generated in the care of patients are widely used to support clinical research and quality improvement, which has hastened the development of self-service query tools. User interface design for such tools, execution of query activity, and underlying application architecture have not been widely reported, and existing tools reflect a wide heterogeneity of methods and technical frameworks. We describe the design, application architecture, and use of a self-service model for enterprise data delivery within Duke Medicine.

Methods: Our query platform, the Duke Enterprise Data Unified Content Explorer (DEDUCE), supports enhanced data exploration, cohort identification, and data extraction from our enterprise data warehouse (EDW) using a series of modular environments that interact with a central keystone module, Cohort Manager (CM). A data-driven application architecture is implemented through three components: an application data dictionary, the concept of "smart dimensions", and dynamically-generated user interfaces.

Results: DEDUCE CM allows flexible hierarchies of EDW queries within a grid-like workspace. A cohort "join" functionality allows switching between filters based on criteria occurring within or across patient encounters. To date, 674 users have been trained and activated in DEDUCE, and logon activity shows a steady increase, with variability between months. A comparison of filter conditions and export criteria shows that these activities have different patterns of usage across subject areas.

Conclusions: Organizations with sophisticated EDWs may find that users benefit from development of advanced query functionality, complimentary to the user interfaces and infrastructure used in other well-published models. Driven by its EDW context, the DEDUCE application architecture was also designed to be responsive to source data and to allow modification through alterations in metadata rather than programming, allowing an agile response to source system changes.

1. Introduction

Many healthcare organizations have created integrated data repositories to organize and store data generated in the care of patients for secondary analysis [1–3]. Traditionally, these data have been provided to researchers following consultation with database analysts, who translate requirements into structured query language (SQL) queries that return data extracts. However,

Abbreviations: BI, Business Intelligence; CCOW, Clinical Context Object Workgroup; CM, Cohort Manager; CR, Chart Review; CT, Clinical Text; DEDUCE, Duke Enterprise Data Unified Content Explorer; DMC, Duke MAESTRO Care; EDW, enterprise data warehouse; EHR, electronic health record; EM, Export Manager; ETL, extract-transform-load; GQ, Guided Query; HbA1c, glycated hemoglobin; HIPAA, Health Insurance Portability and Accountability Act; i2b2, informatics for integrating biology and the bedside; IRB, institutional review board; ldl, low-density lipoprotein; MAESTRO, Medical Application Environments Supporting Transformation of Research and Operations; NCRR, National Center for Research Resources; NIH, National Institutes of Health; NINR, National Institute of Nursing Research; PHI, protected health information; RPDR, Research Patient Data Repository; SQL, structured query language; STRIDE, Stanford Translational Research Integrated Database Environment.

* Corresponding author. Address: Health Intelligence and Research Services, Duke Health Technology Solutions, DUMC Box 104501, Durham, NC 27710, United States.
E-mail address: monica.horvath@duke.edu (M.M. Horvath).

http://dx.doi.org/10.1016/j.jbi.2014.07.006
1532-0464/

this method is time-consuming, does not scale easily, and often does not allow clinician researchers or quality improvement personnel to intervene during data exploration in order to match project and research objectives.

A self-service query interface can alleviate these issues, allowing users to access data without first needing to understand the underlying database structure [1–3]. However, such an interface must ensure a consistent environment for data exploration and related activities by users and provide a robust framework for complex query concepts. Despite the importance of this topic, issues surrounding the design of a user-friendly query interface, query path, and underlying application architecture have received scant attention in the literature, likely reflecting the wide heterogeneity of available methods and technical frameworks [4–7].

In this paper, we discuss the design, application architecture, and implementation of Cohort Manager (CM), the keystone module within the Duke Enterprise Data Unified Content Explorer (DEDUCE) self-service data access portal, and its context in the overall application modular design. CM allows subject–area-spanning cohort definition, combination of cohorts from multiple queries, and data extract definition in a Web-based environment. We describe our framework for presenting complex query design options to users, the application architecture supporting this framework, and the product's usage patterns to date.

2. Materials and methods

2.1. Duke Medicine enterprise data warehouse

The Duke Medicine enterprise data warehouse (EDW) stores and interrelates data generated in the care of the approximately 4.3 million patients in our health system, comprised of three hospitals and a wide network of affiliated outpatient primary care and specialty clinics. Retrospective data extend back to 1996. The EDW is a dimensionally modeled, standards-based database organized into multiple high-level subject areas such as demographics, encounters, provider orders, procedures, diagnoses, lab results, medications, vitals, radiology reports, and pathology reports. A series of extract-transform-load (ETL) processes integrate data from source systems to ensure consistency and quality and to minimize redundancy. The EDW is supported by a team of more than 25 staff, and has been described previously [8].

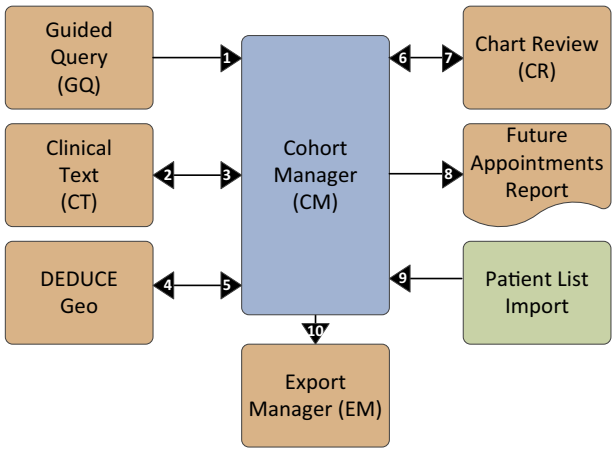

Fig. 1. The DEDUCE modular design for self-service enterprise data delivery. The Duke Enterprise Data Unified Content Explorer (DEDUCE) is comprised of a set of modules, each providing a distinct functionality for enterprise data interaction. Cohort Manager (CM) is the keystone module and provides self-service cohort definitions and joining functionality without the need to use structured query language (SQL) on the underlying enterprise data warehouse (EDW). Other DEDUCE modules interact with CM as illustrated above. Key to numbering: 1. A Guided Query (GQ) result set is imported into CM as a new cohort. 2. A given CM cohort is used to launch associated Clinical Text (CT) reports. 3. Patients from a given set of CT reports are imported into CM as a new cohort. 4. A given CM cohort is used to launch geospatial visualization in DEDUCE Geo. 5. Represents future state where patients identified using Geo visualization will be imported into CM as a new cohort. 6. The Chart Review (CR) functionality is used to select a set of patients and import into CM as a new cohort. 7. A given CM cohort is used to launch the CR dialogue and functionality. 8. A given CM cohort is used to launch the Future Appointments Report. 9. A user's patient listing is imported into CM as a new cohort. 10. A given CM cohort is used to launch the Export Manager module.

2.2. Rationale for a simple, self-service "Guided Query" tool at Duke Medicine

Individual consultation with data analysts had typically been required for obtaining clinical data extracts for research and quality improvement tasks at our organization, but the number of requests exceeded available resources, and wait times for service increased. To address this issue and support a sustainable mechanism to facilitate access to data, we developed a Web-enabled business intelligence environment called DEDUCE Guided Query (GQ), which allows researchers and quality improvement personnel direct access to EDW data. Specifically, DEDUCE GQ allows users to apply filter criteria to one clinical subject area at a time in a wizard-like setting to obtain both aggregate reports and predesigned raw data extracts from the EDW, with protected health information (PHI) available as appropriate [8]. DEDUCE v1.0, with GQ as the centerpiece, was deployed in August 2008 within the Cognos Business Intelligence (BI) platform (version 8.2; IBM, Armonk, NY, USA).

2.3. Development of DEDUCE as a modular design for more sophisticated enterprise data delivery

DEDUCE GQ was intentionally designed to be simple and easy to use; for those reasons, it was not configured to allow user-defined "join" logic. Through feedback from users, we quickly realized that more sophisticated self-service querying functionality was desirable. For example, a researcher may wish to identify all patients with a given chronic disease, such as type 1 diabetes, who have presented to a health-system clinic within the previous month. The user may realize that although this chronic condition is persistent for a patient (as is sex or race), it would not be captured on a patient-based, demographic basis. It would seem logical to use GQ to search for all clinic encounters within the past month and then identify those with appropriate diagnosis codes from the billing context. However, this strategy would miss diabetic patients who might have presented at Duke Medicine, but who did not have the relevant ICD-9-CM code applied during their most recent encounter – a potential scenario if, for example, the patient presented for a lab test or urgent care visit where their diabetes was not pertinent to the care provided. A better approach would be to individually identify two distinct, encounter-based patient cohorts: (a) all patients who had *ever* had a diagnosis code for type 1 diabetes; and (b) all patients who had presented at a Duke clinic in the past month. Joining these two cohorts to retain only those individuals common to both groups (by applying the Boolean "and" operator) would create the correct dataset and have the effect of using information associated with encounters as "patient-based" query attributes.

To allow for this more sophisticated type of self-service data querying, we reevaluated our overall approach and conceptualized an environment of interconnecting modules that, taken together, comprise the DEDUCE enterprise data delivery framework (Fig. 1).

This modular conception has been the blueprint for subsequent releases of DEDUCE. DEDUCE was extended as a Web-based