Journal of Diabetes Science and Technology
Volume 5, Issue 3, May 2011

SYMPOSIUM

Building a Diabetes Screening Population Data Repository Using Electronic Medical Records

Wen-Jan Tuan, M.S., M.P.H.,[1] Ann M. Sheehy, M.D., M.S.,[2] and Maureen A. Smith, M.D., M.P.H., Ph.D.[1]

Abstract

There has been a rapid advancement of information technology in the area of clinical and population health data management since 2000. However, with the fast growth of electronic medical records (EMRs) and the increasing complexity of information systems, it has become challenging for researchers to effectively access, locate, extract, and analyze information critical to their research. This article introduces an outpatient encounter data framework designed to construct an EMR-based population data repository for diabetes screening research. The outpatient encounter data framework is developed on a hybrid data structure of entity–attribute–value models, dimensional models, and relational models. This design preserves a small number of subject-specific tables essential to key clinical constructs in the data repository. It enables atomic information to be maintained in a transparent and meaningful way to researchers and health care practitioners who need to access data and still achieve the same performance level as conventional data warehouse models. A six-layer information processing strategy is developed to extract and transform EMRs to the research data repository. The data structure also complies with both Health Insurance Portability and Accountability Act regulations and the institutional review board's requirements. Although developed for diabetes screening research, the design of the outpatient encounter data framework is suitable for other types of health service research. It may also provide organizations a tool to improve health care quality and efficiency, consistent with the "meaningful use" objectives of the Health Information Technology for Economic and Clinical Health Act.

J Diabetes Sci Technol 2011;5(3):514-522

Introduction

Statistics have shown increasing prevalence of diabetes mellitus among adults and children in the United States.[1] More than 40% of people with diabetes in the United States are not aware of their disease. Reasons contributing to the high prevalence of undiagnosed diabetes are likely complex and multifactorial and may include health system-related factors, such as screening, care management, and provider training.[1–3] It has become more important than

Author Affiliations: [1]Health Innovation Program, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin; and [2]Department of Medicine, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin

Abbreviations: (EMR) electronic medical record, (ETL) extract, transform, and load, (HIPAA) Health Insurance Portability and Accountability Act, (HITECH) Health Information Technology for Economic and Clinical Health, (IRB) institutional review board, (WCHQ) Wisconsin Collaborative for Healthcare Quality

Keywords: diabetes screening, information processing, outpatient research data repository, patient encounter

Corresponding Author: Wen-Jan Tuan, M.S., M.P.H., Health Innovation Program, Room 210-18, 800 University Bay Dr., Madison, WI 53705; email address tuan@wisc.edu

ever to have appropriate and accurate data that can effectively help researchers and health care professionals generate evidence-based knowledge to deliver necessary clinical support for best practice. Yet many challenges and barriers exist in collecting and shaping data to a proper format for research.[4] As part of continuing efforts to address this nationwide diabetes epidemic, this article introduces a data framework that encompasses a multilayer approach design for developing the diabetes screening population data repository using electronic medical information. The data framework can be quickly implemented by researchers and complies with both federal regulations and institutional review board (IRB) requirements. Further, it is consistent with the goals of the national Health Information Technology for Economic and Clinical Health (HITECH) initiative to develop a robust infrastructure to manage patient data and information. This database can serve as a means to increase diabetes case findings, but it can also serve as a model to improve quality for other chronic diseases.

Design Objectives

There has been a rapid advancement of information technology in the area of clinical and population health data management since 2000. Implementation of the electronic medical information system intends to provide accurate and timely data so that health care professionals can obtain the medical knowledge needed to make critical decisions that deliver high-quality health care.[5] However, with the fast growth of electronic medical records (EMR) and the increasing complexity of information systems, it has become challenging for researchers and organizations to effectively access, locate, extract, and analyze information. Majority of the current health care data repositories reside in complex, heterogeneous systems that are often originally designed for financial/operational purposes or tightly bound to specific data models created for other preexisting projects. Ambiguity in the design nature of these data repositories can impose additional barriers for researchers to gather and organize information in a format meeting their analysis needs.[6]

Despite technology complications, there has been great demand for integrating biomedical information from multiple data sources into a single clinical data warehouse.[7,8] These systems are evolving to meet research needs by implementing larger network systems, allowing access to patient records, and integrating ever more items of patient data. However, these data repositories do not always construct information adequate for health services research. The development of these systems can also be expensive and time-consuming.

This article introduces an outpatient encounter data framework designed to construct an EMR-based population data repository for diabetes screening research. The goal of the data framework is to provide researchers a template to extract and transform information from large and complex EMR systems to a disease-specific data warehouse with refined information that is transparent and meaningful to researchers on the project. The objectives of the data framework are to help researchers efficiently identify needed data elements, effectively repurpose data to address research questions, and access and use information while protecting patient privacy and confidentiality.

Methods

Construct

Health informatics literature has shown that patient encounters frequently serve as a core analytical unit in biomedical knowledge repositories or clinical data warehouses designed for disease management or quality-of-care measures.[9,10] The health information technology expert panel of the Agency for Health Research and Quality defines encounters as interactions between a health care provider and a patient for any form of diagnostic treatment and/or therapeutic event.[11] An encounter may consist of a series of health care activities related to medical evaluation, treatments, laboratory tests, and medication.[12] The service characteristics and temporary nature of the encounter can vary significantly by type of encounter, such as physician office visits, phone calls, and hospital stays. The complexity of the care delivery process has made it challenging for researchers to choose proper criteria to define encounters for research. In an effort to address encounters in an ambulatory care context, we specify an outpatient encounter as a face-to-face contact between a patient and a provider of health care on a specific date. The location of the encounter must be in an office or in an outpatient setting. This study also limits encounters to billable services rendered by physicians, physician assistants, or nurse practitioners.

In typical outpatient or physician office visits, an encounter involves one health care provider performing a set of diagnostic or treatment activities during the visit period. Although a patient may have multiple visits to a provider on a single day, all events that occurred in these visits are grouped into one single encounter. Ancillary services for which a performing provider does not exercise independent medical judgment in diagnosing or treating conditions are not considered encounters. For instance, a nurse assisting a physician by drawing a blood sample is not an encounter. Nonetheless, these

procedures are included in addition to other medical services as part of the single encounter.

Figure 1 presents a conceptual structure for the diabetes screening research data repository. In addition to the patient, provider, location, and encounter constructs, the data repository includes diagnostic/treatment-related components as well as other ancillary constructs, such as vital signs, laboratory orders, and medications. Records in the diagnosis and procedures constructs are directly connected to encounter data in either a one-to-one or one-to-many relationship. Although most of the data in the ancillary constructs can be also associated with their corresponding encounters in a one-to-many relationship, laboratory or prescription records ordered by providers outside a face-to-face contact (e.g., telephone call) or by providers from external organizations cannot be linked to the encounter data. These data can be linked through the patient and are included, resulting in a one-to-zero-or-many relationship. Details of these constructs and methods used to create them are discussed in the following subsections.

Patient

This study analyzed ambulatory care practices in a major Midwestern academic-based physician group from January 1, 2005, through December 31, 2007. We focused on ambulatory patient care because of its involvement with diabetes prevention and screening in asymptomatic people. Patients included in the study also actively interacted with their primary care providers in the context of ongoing care. In essence, the diabetes screening data repository consisted of patients who met the Wisconsin Collaborative for Healthcare Quality's (WCHQ) "currently managed" definition during the 3-year study period.[3] The method is property of the WCHQ and is used herein with their permission. Specifically, in order to be considered "currently managed" for a year, patients were required to have had at least two primary care encounters (internal medicine or family practice) in an outpatient, nonurgent setting in the past 36 months, with at least one of those visits in the past 24 months. In other words, for 2005, the patient needed to have at least two visits in the combined years 2003, 2004, and 2005, with at least one being in 2004 or 2005. If the patient did not independently meet this WCHQ definition for each of the three years, they were not included. Patients also had to be aged 20 years or older on January 1, 2005. Pregnant or diabetic patients and individuals who died during the study period were excluded from the diabetes screening data repository.

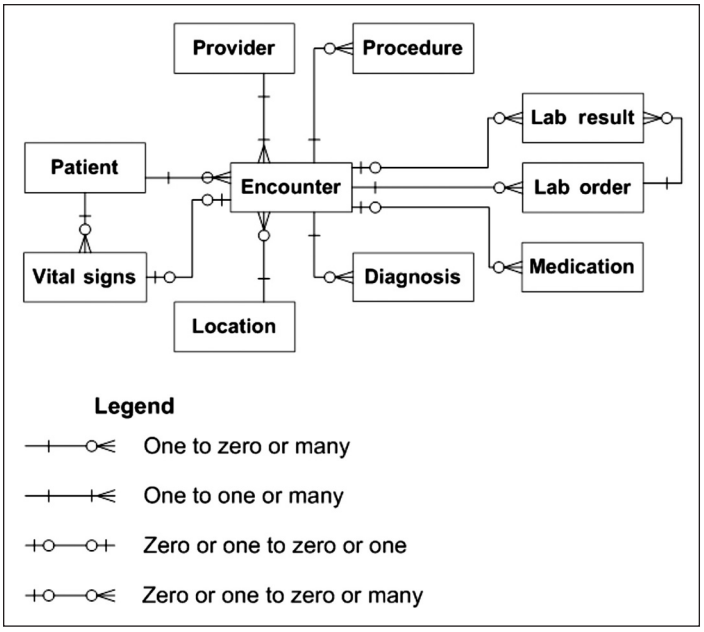

Figure 1. Conceptual structure of the outpatient encounter data framework.

Information Processing Principles

The EMRs used for the diabetes screening research reside in our enterprise reporting database developed by a leading software company in medical information systems.[13] Tables in the EMR source database are constructed based on a relational model designed for general operation and transaction purposes. Despite a variety of clinical data available, the source database is not effectively structured for health services research. The source database also contains patient-identifiable information and other private data. Those data elements may or may not be viewed by researchers, depending on their data use agreements or IRB approvals. The database is also physically located on a different network available only to a limited number of authorized staff.

To ensure data validity and the protection of patient confidentiality, we propose a six-layer information processing model to serve as a guiding principle for developing data extracting, transforming, and loading criteria used to construct the diabetes screening data warehouse (see **Figure 2** for an illustration). Each layer represents unique functional features corresponding to a specific stage in the data extraction and transformation process. A layer consists of blocks sharing similar logical contexts specific to the layer. A single block may contain one or multiple data tables with individual medical records or aggregated data extracted from the source database. The composition of the blocks can vary by project, depending on the sample design and analytical

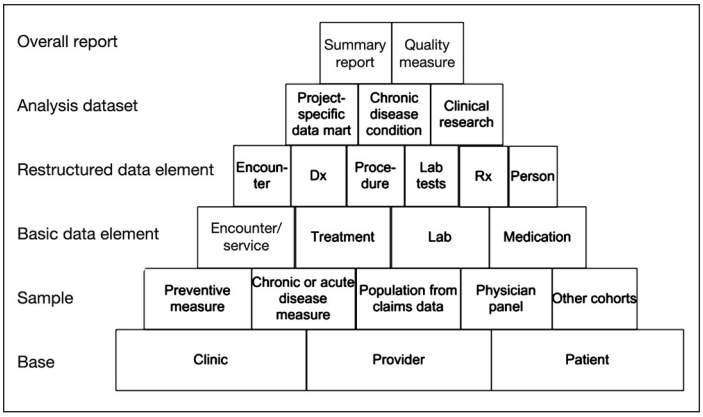

Figure 2. Information processing model. Dx, diagnosis; Rx, prescription.

needs of the project. Data are processed in a sequential fashion from the lower layer to the upper layer.

Data Repository Architecture

The architecture of the diabetes screening data repository consists of the six information processing layers mentioned earlier. The development starts with the first three layers, which focus on extracting and integrating source data into a core structure composed of all necessary data components for health services research. To protect patient confidentiality and meet analysis needs, data in the first three layers are further transformed into three subsequent layers, which comprise data marts designed for research or reporting purposes.

The first three layers of the data framework are composed of interconnected components, including a base layer, a sample layer, and a basic data element layer. The "base" layer contains information specific to clinic, provider, and patient constructs. These constructs hold unique operational or profiling information related to patients, affiliated health care professionals, and clinics in health service organizations. Patient sample lists, provider panels, and disease registries are maintained in the "sample" layer. Data in the sample layer have often been previously created for quality measures or other clinical management purposes. The "basic data element" layer contains both the detailed and the summarized information of the encounter data generated from the EMR. Tables in this layer also include linking variables and system-related data elements needed for constructing the encounter and other ancillary constructs.

The "restructured data element" layer constitutes actual data tables in the research data repository. All protected health information in this layer, with the exception of information allowed in Health Insurance Portability and Accountability Act (HIPAA)-compliant limited dates, is either removed or replaced by encrypted values that cannot be linked back to the source data. Additional constructs may be incorporated into this layer to meet specific project needs. **Figure 3** presents a logical model of key data tables in the restructured data element layer (lookup tables are not presented). Without sensitive information, data in this layer can be easily transported into different storage locations or accessed by researchers who are not permitted to view patient-identifiable information.

In the "analysis data set" layer, researchers can continue to construct project-specific data marts for their clinical research, using information in the restructured data element layer. Reconstructed data can be further organized into the "overall reporting" layer for enterprise reporting purposes.

Extract, Transform, and Load Processes

SAS/Base (SAS Institute, Inc., Cary, NC) is used as the primary extract, transform, and load (ETL) tool to construct the diabetes screening data repository. The design of the outpatient encounter data framework serves as a guideline to repurpose EMRs into a centralized data repository. The actual construction of the data repository should be platform and software independent.

Figure 4 shows an infrastructure topology involved in the ETL process, corresponding to the six information processing layers. In essence, patients' EMR data were stored in an enterprise-reporting database. The EMR data were systematically transformed to tables in the base, sample, and basic data element layers. A de-identification process was performed to replace patient and encounter identifiers with encrypted random numbers when data were moved from the basic data element layer to the restructured data element layer. Different encrypted numbers were also regenerated every time the restructured data element layer was updated. An auditing algorithm was also placed to ensure removal of patient-identifiable information at the end of the ETL process. All the ETL activities were performed in a secure ETL instance on a Windows server (Intel Xeon 5130 2.00 GHz dual-core cpu; 4 GB RAM; Win Server 2003R2 standard edition) behind a HIPAA-compliant firewall.

In our diabetes screening study, there were 51,970 patients who met the currently managed definition for each year between 2005 and 2007 in the EMR database. A total

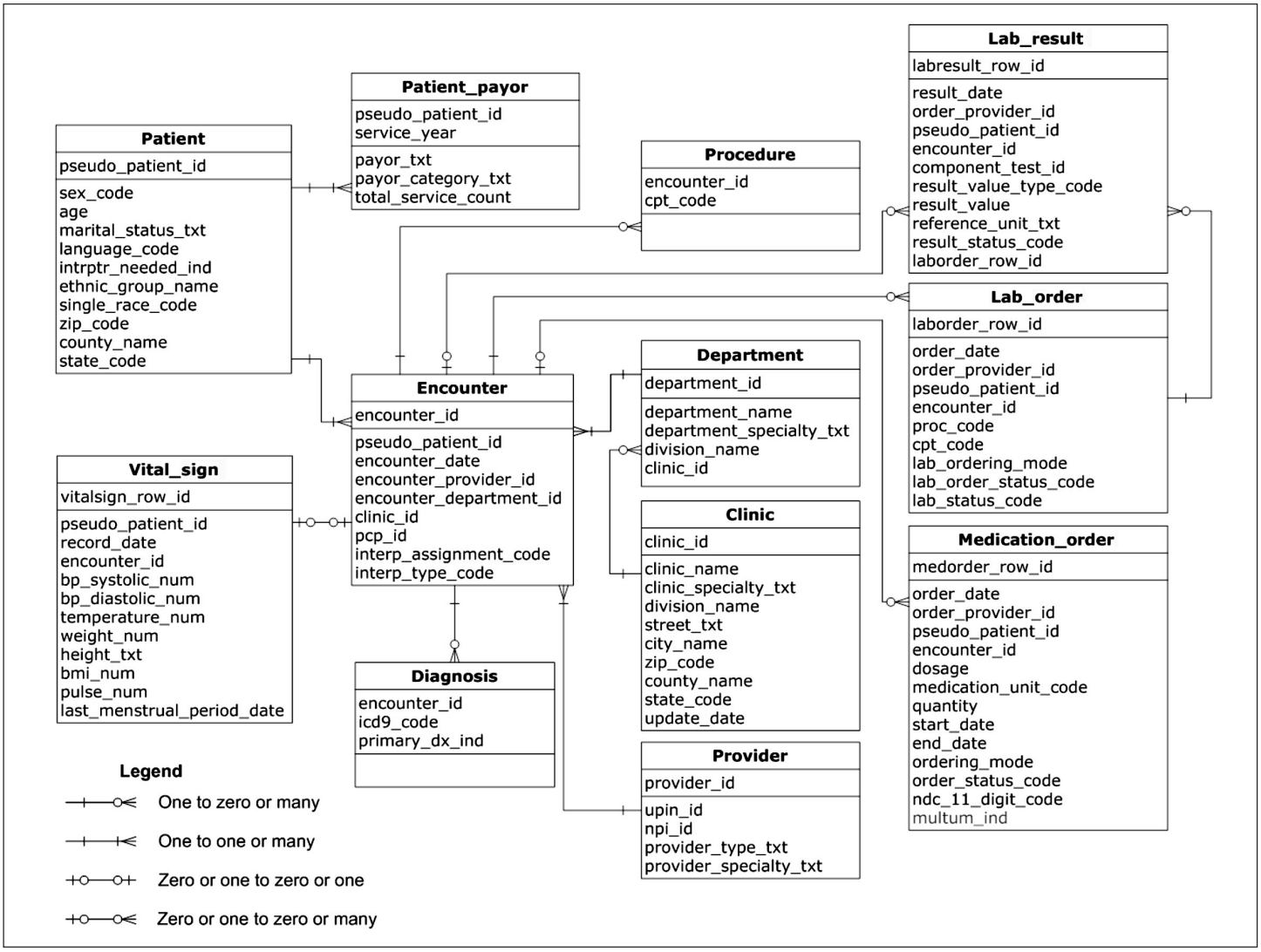

Figure 3. Logical structure of the outpatient encounter data framework.

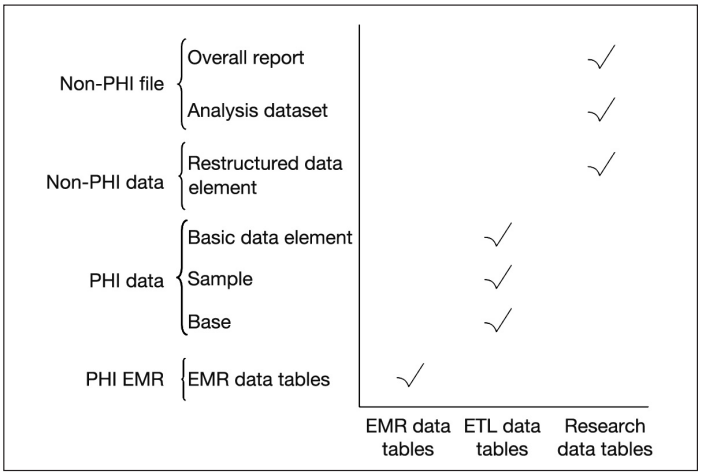

Figure 4. Infrastructure topology of the ETL process. PHI, protected health information.

of 4979 patients were removed because of pregnancy or preexisting diabetic conditions (see **Table 1** for the exclusion criteria), leaving 46,991 patients in the final study population. In order to determine prior diagnosis of chronic diseases, prediabetes data, and other medical history, patients' medical history, treatments, laboratory, and medication records generated between 2003 and 2007 were extracted from the EMR database. The ETL process of the first three layers was completed in approximately 2.5 h. The protected health data elements were kept in the secure ETL instance for data linking and validation purposes. Only authorized analysts are permitted to access this instance.

Information in the basic data element layer was further de-identified and transformed into a centralized research

Table 1.
Exclusion Criteria: Prior Diagnosis of Diabetes and Pregnancy[a]

Condition	Definition	Criteria	ICD-9 codes
Diabetes mellitus	Diagnosis of diabetes mellitus by diagnosis code	Two ICD-9 codes per Hebert and colleagues[14] on two separate occasions within any 2-year time period 2003–2004	250.00, 250.01, 250.02, 250.03, 250.10, 250.11, 250.12, 250.13, 250.20, 250.21, 250.22, 250.23, 250.30, 250.31, 250.32, 250.33, 250.40, 250.41, 250.42, 250.43, 250.50, 250.51, 250.52, 250.53, 250.60, 250.61, 250.62, 250.63, 250.70, 250.71, 250.72, 250.73, 250.80, 250.81, 250.82, 250.83, 250.90, 250.91, 250.92, 250.93, 357.2, 362.01, 362.02, 366.41
Pregnancy	Diagnosis of pregnancy by diagnosis code	≥1 PP-1 code *or* both a PP-2 and PP-4 code in any year 2003–2007. PP-1 and PP-4 are ICD-9 codes, PP-2 CPT codes.	PP-1: 650, V27, V27.0, V27.2, V27.3, V27.5, V27.6; PP-2: 59400, 59409, 59410, 59510, 59514, 59515, 59610, 59612, 59614, 59618, 59620, 59622; PP-4: 640.01, 640.11, 640.21, 640.31, 640.41, 640.51, 640.61, 640.71, 640.81, 640.91, 641.01, 641.11, 641.21, 641.31, 641.41, 641.51, 641.61, 641.71, 641.81, 641.91, 642.01. 642.02, 642.11, 642.12, 642.21, 642.22, 642.31, 642.32, 642.41, 642.42, 642.51, 642.52, 642.61, 642.62, 642.71, 642.72, 642.81, 642.82, 642.91, 642.92, 643.01, 643.11, 643.21, 643.31, 643.41, 643.51, 643.61, 643.71, 643.81, 643.91, 644.21, 645.11, 645.21, 646.01, 646.11, 646.12, 646.21, 646.22, 646.31, 646.41, 646.42, 646.51, 646.52, 646.61, 646.62, 646.71, 646.81, 646.82, 646.91, 647.01, 647.02, 647.11, 647.12, 647.21, 647.22, 647.31, 647.32, 647.41, 647.42, 647.51, 647.52, 647.61, 647.62, 647.71, 647.72, 647.81, 647.82, 647.91, 647.92, 648.01, 648.02, 648.11, 648.12, 648.21, 648.22, 648.31, 648.32, 648.41, 648.42, 648.51, 648.52, 648.61, 648.62, 648.71, 648.72, 648.81, 648.82, 648.91, 648.92, 651.01, 651.11, 651.21, 651.31, 651.41, 651.51, 651.61, 651.71, 651.81, 651.91, 652.01, 652.11, 652.21, 652.31, 652.41, 652.51, 652.61, 652.71, 652.81, 652.91, 653.01, 653.21, 653.31, 653.41, 653.51, 653.61, 653.71, 653.81, 653.91, 654.01, 654.02, 654.11, 654.12, 654.21, 654.22, 654.31, 654.32, 654.41, 654.42, 654.51, 654.52, 654.61, 654.62, 654.71, 654.72, 654.81, 654.82, 654.91, 654.92, 655.01, 655.21, 655.31, 655.41, 655.51, 655.61, 655.71, 655.81, 655.91, 656.01, 656.11, 656.21, 656.31, 656.51, 656.61, 656.71, 656.81, 658.91, 659.01, 659.11, 659.21, 659.31, 659.41, 659.51, 659.61, 659.71, 659.81, 659.91, 660.01, 660.11, 660.21, 660.31, 660.41, 660.51, 660.61, 660.71, 660.81, 660.91, 661.01, 661.11, 661.21, 661.31, 661.41, 661.51, 661.61, 661.71, 661.81, 661.91, 662.01, 662.11, 662.21, 662.31, 662.41, 662.51, 662.61, 662.71, 662.81, 662.91, 663.01, 663.11, 663.21, 663.31, 663.41, 663.51, 663.61, 663.71, 663.81, 663.91, 663.93, 664.01, 664.11, 664.21, 664.31, 664.41, 664.51, 664.61, 664.71, 664.81, 664.91, 665.01, 665.11, 665.22, 665.31, 665.41, 665.51, 665.61, 665.71, 665.72, 665.81, 665.82, 665.91, 665.92, 666.02, 666.12, 666.22, 666.32, 666.42, 666.52, 666.62, 666.72, 666.82, 666.92, 667.02, 667.12, 667.22, 667.32, 667.42, 667.52, 667.62, 667.72, 667.82, 667.92, 668.01, 668.02, 668.11, 668.12, 668.21, 668.22, 668.31, 668.32, 668.41, 668.42, 668.51, 668.52, 668.61, 668.62, 668.81, 668.82, 668.91, 668.92, 669.01, 669.02, 669.11, 669.12, 669.21, 669.22, 669.31, 669.32, 669.41, 669.42, 669.51, 669.52, 669.61, 669.62, 669.71, 669.72, 669.81, 669.82, 669.91, 669.92, 670.02, 671.01, 671.02, 671.11, 671.12, 671.21, 671.22, 671.31, 671.32, 671.41, 671.42, 671.51, 671.52, 671.61, 671.62, 671.71, 671.72, 671.81, 671.82, 671.91, 671.92, 672.02, 673.01, 673.02, 673.11, 673.12, 673.21, 673.22, 673.31, 673.32, 673.41, 673.42, 673.51, 673.52, 673.61, 673.62, 673.71, 673.72, 673.81, 673.82, 674.01, 674.02, 674.12, 674.22, 674.32, 674.42, 674.51, 674.52, 674.62, 674.72, 674.81, 674.82, 674.92, 675.01, 675.02, 675.11, 675.12, 675.21, 675.22, 675.31, 675.32, 675.41, 675.42, 675.51, 675.52, 675.61, 675.62, 675.71, 675.72, 675.81, 675.82, 675.91, 675.92, 676.01, 676.02, 676.11, 676.12, 676.21, 676.22, 676.31, 676.32, 676.41, 676.42, 676.51, 676.52, 676.61, 676.62, 676.71, 676.72, 676.81, 676.82, 676.91, 676.92

[a] ICD-9, International Classification of Diseases, 9th revision; PP, pregnancy; CPT, Current Procedural Terminology.

data repository in the restructured data element layer. Different from data in the previous layers, the research data repository is physically located on a research center's Windows server (Intel Xeon 5450 3.00 GHz dual-core cpu; 8 Gbs RAM; Win Server 2007 standard edition) behind a university-wide firewall. In sum, 956,888 outpatient encounters of the 46,991 patients were extracted, along with 4,732,517 laboratory orders, 995,778 laboratory result records, 1,035,224 medication orders, and 1,064,390 procedures rendered from 1492 health care professionals.

A number of diabetes research-related variables (e.g., American Diabetes Association-designated risk factors, comorbidity variables, screening indicators) were also generated and added into the data repository. Detailed criteria used to construct these research variables are presented in **Table 2**. Data in the diabetes screening research data repository were available to any clinical investigators with explicit IRB approval. Users can continue to mine the data in the data repository in the restructured data element layer. They can also construct specific analytical data sets in the analysis data set layer

Table 2.
Criteria to Create Patient Risk Factors and Comorbidity Variables

Data element	Definition	Criteria	ICD-9[a] codes
Hypertension	Hypertension by diagnosis code	Two ICD-9 codes per Elixhauser and associates[15] on two separate occasions within any 2-year time period 2003–2007	401.1, 401.9, 402.10, 402.90, 404.10, 404.90, 405.11, 405.19, 405.91, 405.99
Cholesterol[b]	Hyperlipidemia or hypertriglyceridemia by diagnosis code or lab	Two ICD-9 codes per Segars and Lea[16] on two separate occasions within any 2-year time period 2003–2007 *or* low-density lipoprotein ≥160 mg/dl[17] or high-density lipoprotein <35 mg/dl or triglycerides >250 mg/dl[18]	272.0–272.9
Polycystic ovarian syndrome	Polycystic ovarian syndrome by diagnosis code	Two ICD-9 codes[c] on two separate occasions within any 2-year time period 2003–2007	256.4
Prediabetes	Impaired fasting glucose, impaired glucose tolerance, subclinical diabetes, or gestational diabetes by diagnosis code	Two ICD-9 codes[c] on two separate occasions from 2003–2004	Impaired fasting glucose: 790.21; impaired glucose tolerance: 790.22; subclinical diabetes: 790.29; gestational diabetes: 648.8
Vascular disease	Ischemic heart disease, stroke, or peripheral vascular disease by diagnosis code	Two ICD-9 codes per Chronic Condition Warehouse[19] (ischemic heart disease), Goldstein,[20] Tirschwell and Longstreth (stroke),[21] or Elixhauser and associates[15] (peripheral vascular disease) on two separate occasions within any 2-year time period 2003–2007	Ischemic heart disease: 410.00, 410.01, 410.02, 410.10, 410.11, 410.12, 410.20, 410.21, 410.22, 410.30, 410.31, 410.32, 410.40, 410.41, 410.42, 410.50, 410.51, 410.52, 410.60, 410.61, 410.62, 410.70, 410.71, 410.72, 410.80, 410.81, 410.82, 410.90, 410.91, 410.92, 411.0, 411.1, 411.81, 411.89, 412, 413.0, 413.1, 413.9, 414.00, 414.01, 414.02, 414.03, 414.04, 414.05, 414.06, 414.07, 414.10, 414.11, 414.12, 414.19, 414.8, 414.9; stroke: 430, 431, 434.00, 434.01, 434.10, 434.11, 434.90, 434.91, 435.0, 435.1, 435.3, 435.8, 435.9, 436, 997.02; peripheral vascular disease: 440.0, 440.1, 440.2, 440.20, 440.21, 440.22, 440.23, 440.24, 440.29, 440.30, 440.31, 440.32, 440.8, 440.9, 441.2, 441.4, 441.7, 441.9, 443.1, 443.81, 443.89, 443.9, 447.1, 557.1, 557.9, V434
Overweight	Body mass index ≥25 kg/m^2 or overweight or obese by diagnosis code	Two ICD-9 codes per Elixhauser and associates[15] on two separate occasions within any 2-year time period 2003–2007 *or* most recent recorded body mass index ≥25 kg/m^2	278.0, 278.00, 278.01

[a] ICD-9, International Classification of Diseases, 9th revision.
[b] International System of Units conversion factors: to convert cholesterol to mmol/liter, multiply by 0.0259; triglycerides to mmol/liter, multiply by 0.0113.
[c] Only a single ICD-9 code exists, no validation references available.

or the overall report layer. For instance, the diabetes-related variables, combined with the encounter data, laboratory information, and medication data able to be extracted using the data framework revealed that the American Diabetes Association diabetes screening guidelines had much better case-finding ability than the United States Preventive Task Force guidelines.[3] The information gained in this process can be used to improve diabetes case finding in our system as well as nationwide and demonstrated the strength of this database design for diabetes screening.

Discussion

There have been a number of data warehouse models proposed for biomedical or clinical research. Some models are designed for specific subject areas, such as oncology[10,22] and orthopedics.[23] Others are intended for bioinformatics purposes or decision support tools, such as the clinical research chart in the integrating biology and the bedside system (i2b2).[9,24] Each design has unique benefits and drawbacks. For instance, the i2b2 system allows data from different sources to be easily stored in modules or cells on a hive structure. It has been shown to be a robust tool for estimating initial cohort sizes. However, information in the i2b2 system is required to be stored and retrieved at a given encounter level but not at the level of a series of atomic observations that could have been charted together to convey a specific clinical fact or provider–patient session.[25]

There is an increasing demand for an intuitive data architecture for health services research, given the growing role of health services research within medical communities and the public sector.[26] The outpatient encounter data framework is developed on a hybrid data structure of entity-attribute-value models, dimensional models, and relational models. The data framework preserves a small number of subject-specific tables essential to key clinical constructs in the data repository. This approach enables atomic information to be maintained in a transparent and meaningful way to researchers who need to access data. The characteristics of the atomic observations are also often very different among clinical constructs. Storing data into separate construct-based tables (e.g., laboratory results and medication orders) not only ensures that critical research information can be quickly and easily identified, but also allows subject-based *ad hoc* queries to achieve the same performance level as conventional data warehouse models.[27]

Furthermore, maintaining the atomic data with different natures in separate constructs allows the data repository to preserve original spatial relationships between different charted fields. This feature is particularly important for researchers interested in grouping and analyzing a series of atomic events at a chart level. The hybrid structure also provides researchers greater flexibility to integrate additional constructs unique to their studies in the data repository. Those constructs can be created by independent ETL processes without affecting the existing table structure and data on the tables.

The data framework adopts the best practices, such as anonymization and security rules, for assuring ethnic standards and regulatory requirements.[28] Because there is no patient-identifiable information beyond that allowed in the limited data set, the final research data repository will be available to more researchers without jeopardizing patient confidentiality. Our next development task is to create standardized algorithms and modularized programs so that users without extensive database knowledge can also quickly pull information from the data repository into flat files for statistical analyses or reporting purposes. For instance, researchers can run add-variable modules to include variables into their research data sets. Similarly, they can use reporting modules to create overall reports or cohort modules to generate sample populations. We are also planning to include a metadata library allowing users to quickly search for data elements and identify data quality.

Conclusions

Since 2000, a growing amount of research has begun using EMRs because of their rapid accessibility, rich data content, standardized formats, and cost-effective integration.[29,30] The outpatient encounter data framework provides organizations and researchers with a tool to improve health care quality and efficiency, consistent with the "meaningful use" objectives of the HITECH Act. The data framework can be implemented by researchers without extensive information technology backgrounds or extensive resources or funding. It does not involve elaborate data warehouse platforms, Web interfaces, or complex security systems. In the future, we hope to expand the outpatient encounter data framework to a generalized data framework that can account for encounters that occurred in inpatient, home care, telephone, and online settings.

Funding:

This research was supported in part by the University of Wisconsin Department of Medicine, as well as the Health Innovation Program and the Community Academic Partnerships core of the University of Wisconsin Institute for Clinical and Translational Research, and grant 1UL1RR025011 from the Clinical and Translational Science Award program of the National Center for Research Resources, National Institutes of Health. Additional funding for this project was provided by the University of Wisconsin School of Medicine and Public Health from The Wisconsin Partnership Program.

Acknowledgment:

We thank Colleen Brown, B.A., for manuscript review and technical assistance.

References:

1. Ciemins EL, Coon PJ, Fowles JB, Min SJ. Beyond health information technology: critical factors necessary for effective diabetes disease management. J Diabetes Sci Technol. 2009;3(3):452–60.
2. Narayan KM, Boyle JP, Thompson TJ, Sorensen SW, Williamson DF. Lifetime risk for diabetes mellitus in the United States. JAMA. 2003;290(14):1884–90.
3. Sheehy AM, Flood GE, Tuan WJ, Liou JI, Coursin DB, Smith MA. Analysis of guidelines for screening diabetes mellitus in an ambulatory population. Mayo Clin Proc. 2010;85(1):27–35.
4. Chan KS, Fowles JB, Weiner JP. Review: electronic health records and the reliability and validity of quality measures: a review of the literature. Med Care Res Rev. 2010;67(5):503–27.
5. Ledbetter CS, Morgan MW. Toward best practice: leveraging the electronic patient record as a clinical data warehouse. J Healthc Inf Manag. 2001;15(2):119–31.
6. Sen A, Sinha AP. Toward developing data warehousing process standards: An ontology-based review of existing methodologies. IEEE Trans Syst Man Cybern C Appl Rev. 2007;37(1):17–31.
7. Weber GM, Murphy SN, McMurry AJ, Macfadden D, Nigrin DJ, Churchill S, Kohane IS. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. J Am Med Inform Assoc. 2009;16(5):624–30.
8. Sahama TR, Croll PR. A data warehouse architecture for clinical data warehousing. Presented at: First Australasian Workshop on Health Knowledge Management and Discovery (HKMD 2007), Ballarat, Australia.
9. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010;17(2):124–30.
10. Rossille D, Laurent JF, Burgun A. Modelling of a case-based retrieval system for oncology. Stud Health Technol Inform. 2003;95:565–70.
11. U.S. Department of Health and Human Servics; Agency for Healthcare Research and Quality. Data element details: encounter. *http://ushik.ahrq.gov/dr.ui.drData_Page?system=mdr&Search=xxKEYIDxx&KeyOrgID=32&KeyRID=85872000&Referer=DataElement*. Accessed November 30, 2010.
12. Parmanto B, Scotch M, Ahmad S. A framework for designing a health-care outcome data warehouse. Perspect Health Inf Manag. 2005;2:3.
13. UW Health. *http://www.uwhealth.org/about-uwhealth/main/10730*. Accessed December 15, 2010.
14. Hebert PL, Geiss LS, Tierney EF, Engelgau MM, Yawn BP, McBean AM. Identifying persons with diabetes using Medicare claims data. Am J Med Qual. 1999;14(6):270–7.
15. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. Med Care. 1998;36(1):8–27.
16. Segars LW, Lea AR. Assessing prescriptions for statins in ambulatory diabetic patients in the United States: a national, cross-sectional study. Clin Ther. 2008;30(11):2159–66.
17. Grundy SM, Cleeman JI, Merz CN, Brewer HB Jr, Clark LT, Hunninghake DB, Pasternak RC, Smith SC Jr, Stone NJ; National Heart, Lung, and Blood Institute; American College of Cardiology Foundation; American Heart Association. Implications of recent clinical trials for the National Cholesterol Education Program Adult Treatment Panel III guidelines. Circulation. 2004;110(2):227–39.
18. American Diabetes Association. Standards of medical care in diabetes--2009. Diabetes Care. 2009;32 Suppl 1:S13–61.
19. Iowa Foundation for Medical Care. Chronic Condition Data Warehouse user manual, version 2.0. West Des Moines: Iowa Foundation for Medical Care; 2007.
20. Goldstein LB. Accuracy of ICD-9-CM coding for the identification of patients with acute ischemic stroke: effect of modifier codes. Stroke. 1998;29(8):1602–4.
21. Tirschwell DL, Longstreth WT Jr. Validating administrative data in stroke research. Stroke. 2002;33(10):2465–70.
22. Wah TY, Sim OS. Development of a data warehouse for lymphoma cancer diagnosis and treatment decision support. WSEAS Trans Inform Sci Appl. 2009;6(3):530–43.
23. Lin S, Lee YC, Hsu C. Data warehouse approach to build a decision-support platform for orthopedics. Int J Biosci Biotechnol. 2010;2(1).
24. Meystre SM, Deshmukh VG, Mitchell J. A clinical use case to evaluate the i2b2 hive: predicting asthma exacerbations. AMIA Annu Symp Proc. 2009:442–6.
25. Deshmukh VG, Meystre SM, Mitchell JA. Evaluating the informatics for integrating biology and the bedside system for clinical research. BMC Med Res Methodol. 2009;9:70.
26. Einbinder JS, Scully KW, Pates RD, Schubart JR, Reynolds RE. Case study: a data warehouse for an academic medical center. J Healthc Inf Manag. 2001;15(2):165–75.
27. Chen RS, Nadkarni P, Marenco L, Levin F, Erdos J, Miller PL. Exploring performance issues for a clinical database organized using an entity-attribute-value representation. J Am Med Inform Assoc. 2000;7(5):475–87.
28. Karp DR, Carlin S, Cook-Deegan R, Ford DE, Geller G, Glass DN, Greely H, Guthridge J, Kahn J, Kaslow R, Kraft C, Macqueen K, Malin B, Scheuerman RH, Sugarman J. Ethical and practical issues associated with aggregating database. PLoS Med. 2008;5(9):e190.
29. Dean BB, Lam J, Natoli JL, Butler Q, Aguilar D, Nordyke RJ. Review: use of electronic medical records for health outcomes research: a literature review. Med Care Res Rev. 2009;66(6):611–38.
30. Prokosch HU, Ganslandt T. Perspectives for medical informatics. Reusing the electronic medical record for clinical research. Methods Inf Med. 2009;48(1):38–44.