Case Study: A Data Warehouse for an Academic Medical Center

Jonathan S. Einbinder, MD, MPH; Kenneth W. Scully, MS;
Robert D. Pates, PhD; Jane R. Schubart, MBA, MS;
Robert E. Reynolds, MD, DrPH

ABSTRACT

The clinical data repository (CDR) is a frequently updated relational data warehouse that provides users with direct access to detailed, flexible, and rapid retrospective views of clinical, administrative, and financial patient data for the University of Virginia Health System. This article presents a case study of the CDR, detailing its five-year history and focusing on the unique role of data warehousing in an academic medical center. Specifically, the CDR must support multiple missions, including research and education, in addition to administration and management. Users include not only analysts and administrators but clinicians, researchers, and students.

KEYWORDS

- Data warehousing
- Database
- Academic medical center
- Clinical research
- Case study
- Evaluation

Large organizations build data warehouses to "analyze what has occurred within the business across time" in order to obtain "a competitive edge in the marketplace."[1] Many healthcare organizations see data warehousing as a way to facilitate operational efficiency and informed administrative decision making. In the 2000 HIMSS Leadership Survey of more than eleven hundred healthcare professionals, 58 percent of respondents indicated that their organizations were currently creating data warehouses or planned to do so within the next two years.[2] However, the missions of academic medical centers extend beyond administrative analysis and decision making. The mission statement for the University of Virginia Health System, for example, refers to "advancement of medical and scientific knowledge" and "professional preparation of individuals dedicated to health care service"—missions that can be expressed more succinctly as research and education.

At the University of Virginia, we have created the clinical data repository (CDR)—a data warehouse that is intended to support the research and education functions of our academic medical center in addition to providing data for managers and administrators. In this article, we present a case study of the CDR, describing the system and discussing the premise that a data warehouse for an academic medical center should support the organization's academic missions.

History of the CDR

When the project began in 1995–96, the CDR, initially referred to as the "clinical research database," was intended to support and enhance clinical research at the University of Virginia by providing clinicians, students, and researchers with direct, rapid access to retrospective clinical and administrative patient data.[3] Reflecting this intent, the system was funded by the School of Medicine and housed in the Academic Computing Health Sciences group, which is distinct from the medical center's IT group. With considerable assistance and cooperation from data owners and stewards, legacy data from several different sources were loaded into a single relational database and periodically updated. Authorized users accessed the CDR through a standard Web browser and viewed or downloaded data to their personal computers for further analysis. Initially, emphasis was placed on getting the CDR running as quickly as possible and with a minimum of resources; consequently, extensive transformation of data to an enterprise data model was not performed.

The CDR project team consists of 2.5–3.0 FTEs (full-time equivalents)—one developer, one developer-database administrator, and portions of analyst, clinician, and administrative FTEs. To date, the costs of developing and operating the CDR have been approximately $200,000 per year, underwritten by the School of Medicine.

Over the course of the project, there have been significant enhancements to the user interface, incorporation of additional data sources, and the development of an integrated data model. There has also been increasing interest in using the CDR to serve a broader audience than researchers and to support management and administrative functions—"to meet the challenge of providing a way for anyone with a need to know—at every level of the organization—access to accurate and timely data necessary to support effective decision making, clinical research, and process improvement."[4] In the area of education, the CDR has become a core teaching resource for the Department of Health Evaluation Science's master's program and for the School of Nursing. Students use the CDR to understand and master informatics issues such as data capture, vocabularies, and coding, as well as to perform exploratory analyses of healthcare questions. Starting in Spring 2001, the CDR will also be introduced into the university's undergraduate medical curriculum.

System Description

Following is a brief overview of the CDR application as it exists at the University of Virginia.

System Architecture. The CDR is a relational data warehouse that resides on a Dell PowerEdge 1300 (Dual Intel 400MHz processors, 512MB RAM) running the Linux operating system and Sybase 11.9.1 relational database management system. For storage, the system uses a Dell Powervault 201S 236GB RAID Disk Array. As of October 2000, the database contained 23GB of information about 5.4 million patient visits (16GB visit data, 7GB laboratory results). Data loading into Sybase is achieved using custom Practical Extraction and Report Language (Perl) programs.

CDR Contents. The CDR currently draws data from four independent systems (see Table 1). In addition, a number of derived values (for example, number of days to next inpatient visit, number of times a diagnostic code is used in various settings) are computed to provide summary information for selected data elements. Data from each of these source systems are integrated into the CDR's data model.

In addition to the current contents listed in Table 1, users and the CDR project team have identified additional data elements that might be incorporated

Table 1. Contents of the CDR

Type of Data	*Source of Data*	*Description*	*Available Dates*
Inpatient, outpatient visits	Shared Medical Systems	Patient registration and demographic data, diagnoses, procedures, unit and census information, billing transactions, including medications, costs, charges, reimbursement, insurance information	Jul 1993–Jun 2000
Professional billing	IDX billing system	Physician billing transactions from inpatient and outpatient visits, diagnoses, and procedures	Oct 1992–Jun 2000
Laboratory results	HL-7 messages from SunQuest Lab System	Laboratory test results	Jan 1996–Jun 2000
Cardiac surgery	Cardiac surgery outcomes data (defined by Society of Thoracic Surgeons	Clinical details for thoracic surgery cases	Jul 1993–Jun 2000

into the CDR, including microbiology results, discharge summaries (and other narrative data), outpatient prescribing information, order entry details, and tumor registry information. As of October 2000, we have just finished incorporating death registry data from the Virginia Department of Health into the CDR. These data will provide our users with direct access to more comprehensive mortality outcomes data than are contained in local information systems, which generally are restricted to an in-hospital death indicator.

User Interface. The user interface runs in a standard Web browser and consists of a data dictionary, a collection of common gateway interface (CGI) programs implemented using the "C" programming language, and JavaScript-enabled HTML pages. Structured query language (SQL) statements are generated automatically in response to point-and-click actions by the user, enabling submission of ad hoc queries without prior knowledge of SQL. The SQL queries are sent to the CGI programs that query the database and return results in dynamically created HTML pages. The entire process is controlled by the contents of the data dictionary, which is used to format SQL results, set up HTML links for data drill-down, and provide on-line help. Data may be downloaded immediately into Microsoft Excel or another analysis tool on the user's workstation.

Query Formulation. Most CDR users use the Guided Query function to retrieve data. This process involves three steps:

1. Define a population of interest by setting conditions, for example, date ranges, diagnostic codes, physician identifiers, service locations, and lab test codes or values.
2. Submit the query, specifying how much data the CDR should return (all matching data or a specified number of rows).
3. After the CDR returns the population of interest, use the Report Menu to explore various attributes of the population on a case-by-case or group level. Custom reports can also be defined, and the results of any report can be downloaded into Microsoft Excel, Access, or other analysis tool.

Generally, the query process requires several iterations to modify the population conditions or report options. In addition, "browsing" the data may help the user generate ideas for additional queries. We believe that it is helpful for end users to go through this query process themselves—to directly engage the data. However, many users, especially those with a pressing need for data for a meeting, report, or grant, prefer to use CDR team members as intermediaries or analysts. To date, we have attempted to meet this preference, but as query volume increases, our ability to provide data in a timely manner may fall off.

Security. A steering committee of clinicians guided the initial development of the CDR and established policies for its utilization and access. Only authorized users may log onto the CDR. To protect confidentiality, all patient and physician identifying information has been partitioned into a "secure" database. Translation from or to disguised identifiers to or from actual identifiers is possible but requires a written request and appropriate approval (for example, from a supervisor or the human investigations committee). All data transmitted from the database server to the user's browser are encrypted using the secure Netscape Web server, and all accesses to the database are logged. In addition, CDR access is restricted to personal computers that are part of the "Virginia.edu" domain or that are authenticated by the university's proxy server.

Evaluation

Understanding user needs is the basis for improving the CDR to enable users to retrieve the data independently and to increase usage of the CDR at our institution. Thus, assessing the value of the CDR—how well we meet our users' needs and how we might increase our user base—has been an important activity that has helped guide planning for changes and enhancements and for allocation of our limited resources. Efforts to evaluate the CDR have included several approaches:

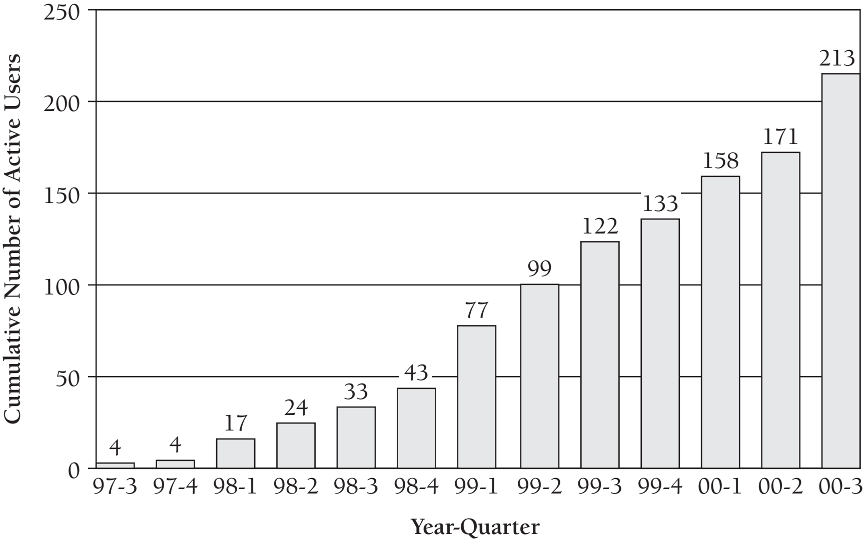
- Monitoring user population and usage patterns
- Administering a CDR user survey
- Tracking queries submitted to the CDR and performing follow-up telephone interviews

Usage Statistics. Voluntary usage of an IS resource is an important measure of its value and of user satisfaction.[5] However, usage of a data warehouse is likely to be quite different than for other types of information resources, such as clinical information systems. A clinical system is likely to be used many times per day; a data warehouse may be used sporadically. Thus, although we monitor system usage as a measure of the CDR's value, we believe that frequency of usage cannot be viewed in isolation in assessing the success of a data warehouse.

Since the CDR went "live," more than 300 individuals have requested and obtained logon IDs. As of September 30, 2000, 213 individuals had logged on and submitted at least one query. This number does not include usage by CDR project team members and does not reflect analyses performed by team members for end users. Figure 1 shows the cumulative number of active users (those who submitted a query) and demonstrates a linear growth pattern. The number of active users in any given quarter has been fairly constant or slightly increased.

CDR User Survey. In September 1998, we administered a survey to understand why users had chosen to adopt the CDR.[6] For a theoretical framework, the survey used Everett Rogers' diffusion of innovation theory, which describes the stages that potential adopters pass through when exposed to an innovation like the CDR.[7] In keeping with Rogers' framework, we asked questions about information content, ease of use, output format, accuracy of the

Figure 1. Cumulative Number of Active CDR Users



Note: An *active user* is a person who logged on to the CDR and submitted at least one query.

underlying data, system response time, and user work styles. Key findings included the following:

- Initial use was best explained by a user's proficiency in pertinent computer applications, familiarity with standard coding conventions, and an understanding of how data are recorded at our institution.
- Compatibility with an individual's work style and skills was strongly associated with satisfaction and continued use of the CDR. "Ease of use" was also associated.
- The reason most often cited for not using the CDR was "not enough time," pointing to the time constraints of busy health professionals as an important factor to consider in system design and training.
- The CDR's use of encrypted patient and physician identifiers was viewed as a barrier to usage by 44 percent of respondents.
- Most users were satisfied with the accuracy of data provided by the CDR and believed it would become an important resource at the University of Virginia.

Tracking Queries and Telephone Survey. To obtain a better understanding of the kinds of questions CDR users are asking, we periodically review SQL statements submitted to the CDR and contact users to clarify what they were looking for and whether they succeeded. CDR team members also keep logs of queries performed on a consultative basis. The ensuing list [http://www.med.virginia.edu/cdr/publications/usage.html] has also proven useful for internal marketing and for justification of CDR funding. Beginning in October 1999, users identified from an audit of usage logs were telephoned and a semi-structured interview administered. During the initial eleven-month period of data collection, 90 unique CDR users conducted 460 CDR sessions (defined as a single day of use). We examined over 100 queries in detail. The results thus far suggest some patterns.[8]

Who uses the CDR? CDR users represent a diverse group that includes physicians, other clinicians, health services researchers, analysts, managers, teaching faculty, and students. Most have attended a training class, are infrequent users, and rely on user support from CDR personnel for complex queries.

What kinds of questions are users asking? Generally, users are interested in identifying a specific population with the goal of examining treatment efficacy or utilization of resources. Table 2 provides examples of user queries.

How do users attempt to answer their questions? User questions frequently required information that crossed multiple source systems—inpatient, outpatient, professional billing, and clinical laboratory results. The “guided query” was the most popular interface choice (as opposed to writing SQL statements directly). Users were generally able to develop simple queries on their own but tended to have difficulty with more complex queries, prompting them to seek assistance from CDR personnel.

Discussion

An academic medical center has multiple missions—patient care, teaching, and research—that must be carried out in a cost-effective and efficient manner. There is no particular reason that a data warehouse designed primarily with one purpose or constituency in mind (for example, administrators) cannot be used to support other functions, such as research and education. However, for several reasons, this crossover may not occur. The customers and primary purpose of any information system do inform its design and operation. Traditionally, a healthcare data warehousing uses administrative and financial data primarily to facilitate business goals and may thus operationally reside in the IS group or finance department. The data model may be optimized for financial and performance reporting, and the user interface may emphasize predefined reports targeted at administrators and decision makers. Clinicians and researchers may not have ready access to the system or to the analysts that function as its intermediaries. In addition, privacy and confidentiality policies and methodologies may not be set up to serve a research or teaching function.

At the University of Virginia, with the CDR we have attempted to address some of these issues to support our academic customers. Given the argument that conventional data warehouses may not support all of an academic medical center’s missions, it is fair to ask whether a system like the CDR, which is

Table 2. Examples of CDR Queries

Category	*Query*
Clinical Research	
Case finding	Identify patients for a clinical trial of a new medication to treat associated bone mineral loss in primary biliary cirrhosis
Exploratory analysis	Assess perioperative changes in neuroendocrine function in patients undergoing pituitary microsurgery
Grant preparation	Identify clinical details for two hundred recent stroke patients as part of a large multicenter research grant studying the occurrence of seizures following stroke
Quality Improvement	
Outcomes evaluation and assessment	Assess mortality of children with severe neurodevelopmental disabilities who receive tube feeding
Treatment patterns and protocols	Evaluate before-and-after results (admissions, length of stay) for a clinical practice guideline for DVT and PE
Medical management	Compare the frequency and cost of arterial blood gas tests in ventilated patients treated with and without weaning protocols
Practice Management	
Utilization analysis	As part of a laboratory utilization project, analyze lab results to determine probabilities of abnormal values. Analysis to be used to develop decision rules for test ordering.
Cost-containment study	Assess whether the cost of treating patients who had radical head-neck surgery is more expensive than a three-drug chemotherapy regimen for patients with particular head-neck cancers.
Feasibility study	Review data for patients with LOS > 25 days, particularly diagnoses and payer class, to assess need for a long-term care facility.
Marketing	Analyze procedure data for patients outside UVA's primary service area to determine why these patients from far away come to our institution. Results will be used to guide a marketing campaign.
Education	
Informatics	Explore different ways of identifying patient populations with pneumonia (exercise to understand coding and ICD-9-CM).
Clinical clerkship	Analyze how many patients admitted with influenza are treated with antiviral medications. Report results on rounds.

designed to support research and teaching, can effectively support managers and administrators. At present, CDR users do include UVA's Medical Management Program, as well as other administrators, but for the most part, administrative users access other databases, including an administrative repository operated by Health System Computing Services. For these administrative users, the CDR has been useful in several situations:

- When clinical data, mainly laboratory results, are required.
- When the user needs data immediately and can directly access the CDR.
- When an integrated view of the data is required, that is, across inpatient-outpatient settings or across hospital-physician billing systems (currently not available elsewhere).
- When a user has an operational question but is not "plugged into" typical administrative data sources. An example of a question is this: How many patients in the General Medicine clinic received pulmonary function testing last year, and how much did this cost?

We have also seen, however, that certain characteristics of the CDR may limit its utility to administrative users. The CDR provides few predefined reports. In part driven by our limited resources and in part by our research focus, we have concentrated on supplying the user with data rather than reports or answers. Many administrative users do not wish to download and analyze data. In addition, the encryption of patient and physician identifiers, required for our research and teaching function, has proven to be an obstacle to some users. It should be noted that, technically, it is not difficult to offer both encrypted and unencrypted views of the data,[9] but our current charter and policies do not allow this. Finally, new data are loaded into the CDR on a quarterly basis; thus, on average, the data are about a month and a half behind. In our user survey, several potential administrative users have indicated that this delay is too long. Each of these limitations could be addressed, in some cases with a change in policy and in other cases with the infusion of additional resources into the CDR. Likewise, with the incorporation of clinical data such as laboratory results and with changes in operating procedures to better support academic users, a conventional, administratively focused data warehouse could probably fill an academic role.

Lessons Learned

Next, we consider several aspects of the CDR project to illustrate some decisions that may be relevant to other data warehousing projects. Some of these decisions were made with full awareness of the implications for users of the system; the significance of other choices has only become apparent in hindsight.

Expectation of Direct Interaction with the Data. The CDR was created with the expectation that end users, including clinicians and researchers, would interact directly with the system, formulating queries and critically appraising the data they return. Potential advantages to this type of interaction include timeliness of data retrieval, opportunity to iteratively revise and modify queries and browse patient data, and a more complete understanding of the strengths and weaknesses of the coded data contained in the CDR (and the opportunity to provide feedback to source systems regarding how the data might be improved). We continue to believe in these benefits but have found that many users prefer to have CDR personnel formulate and submit queries for them. These users are busy people who need to get specific information from the CDR, usually in a very timely manner, and do not wish to take the time to learn how to use the system themselves. This is a paradox: people do not want to take time to learn to use a system when they do not have a clear need, but neither do they have the time when they do have the need. We hope that as we continue to improve the Web user interface and enhance the CDR data model to better integrate disparate source data, more users will perform queries themselves. At present, we attempt to respond to requests for consultation services in a timely manner (and offer to provide training at a later time).

Incremental Approach. The CDR system was designed and implemented rapidly with a minimum of resources. This approach has required some tradeoffs. Initially, there was limited integration of data across source systems. For example, to identify cases with a particular diagnosis required setting separate conditions for diagnoses in inpatient, outpatient, and professional billing data. The CDR has since moved to a data model that more tightly integrates the source databases. The user interface has also gone through several major and minor iterations (currently, we are on version 2.2) to make things easier for users and to take advantage of improvements in the data model. These tradeoffs—getting started quickly with limited resources versus data integration and ease of use—must be carefully considered when building a data warehouse or any information resource.

Organizational Issues. As discussed previously, the CDR began as a resource for clinical researchers. It was not envisioned as a nonresearch resource. We have successfully served nonresearch users and have come to believe that the distinction between "research" and "administrative" uses of a data warehouse is not always clear. Despite this belief, however, the CDR's research orientation and location in the School of Medicine have affected priorities for system development, availability of resources (limited), and potential administrative users' perception of the system.

The implication for institutions considering building data warehouses and interested in supporting all three of the academic medical center's missions is the following: consider carefully where in the organization to house the system and how it will be operated to ensure that administrative, research, and teaching functions can all be supported. One option is to create a single data warehousing group, probably housed in a central IS or informatics group. Or create separate groups and separate warehouses, one for academic and one for administrative users. In between these two extremes, however, are a range of possibilities such as a shared database but separate user interfaces (and separate support analysts), or a single administrative warehouse and a separate research repository that is fed from the warehouse and augmented with additional clinical details. The choice of approach will play a major role in determining which users in an academic medical center have access to timely, relevant data for making business decisions and conducting research.

References

1. Kachur, R. J. *The Data Warehouse Management Handbook.* Upper Saddle River, N.J.: Prentice Hall, 2000.
2. "The Eleventh Annual HIMSS Leadership Survey," sponsored by IBM. [http://www.himss.org/survey/2000/survey2000.html]. 2000.
3. Scully, K. W., and others. "Development of an Enterprisewide Clinical Data Repository: Merging Multiple Legacy Databases." Paper presented at the annual symposium of the American Medical Informatics Association, 1997.
4. Reynolds, R. E., and Knaus, W. A. "Clinical Data Repository Enhancements." Internal memorandum, University of Virginia, Jan. 30, 1998.
5. Slack, W. V. "Assessing the Clinician's Use of Computers." *MD Computing,* 1993, 10(8), 357–360.
6. Schubart, J. R., and Einbinder, J. S. "Evaluation of a Data Warehouse in an Academic Health Sciences Center." *International Journal of Medical Informatics* (forthcoming).
7. Rogers, E. *Diffusion of Innovations*. (3rd ed.) New York: Free Press, 1983.
8. Schubart, J. R., and Einbinder, J. S. "Evaluation of a Data Warehouse: Understanding User Needs." Paper presented at the annual symposium of the American Medical Informatics Association, 2000.
9. Halamka, J. D., and others. "Managing Care in an Integrated Delivery System via an Intranet." Paper presented at the annual symposium of the American Medical Informatics Association, 1998.

About the Authors

Jonathan S. Einbinder, MD, MPH, is assistant professor of clinical informatics in the Department of Health Evaluation Sciences at the University of Virginia, director of data administration at the University of Virginia Health System, and project director for the CDR.

Kenneth W. Scully, MS, is technical director and database administrator for the CDR.

Robert D. Pates, PhD, is a developer for the CDR.

Jane R. Schubart, MBA, MS, is a lecturer in clinical informatics in the Department of Health Evaluation Sciences at the University of Virginia.

Robert E. Reynolds, MD, DrPH, is vice president and interim chief information officer for the University of Virginia and professor of informatics in the Department of Health Evaluation Sciences.