

Development of an Enterprise-Wide Clinical Data Repository: Merging Multiple Legacy Databases

Kenneth W. Scully, M.S., Robert D. Pates, Ph.D., George S. Desper, Alfred F. Connors, M.D.,
Frank E. Harrell Jr., Ph.D., Karen S. Pieper, M.S., Robert L. Hannan, M.D.,
Robert E. Reynolds, M.D.

University of Virginia Health System, Charlottesville, VA

We describe the development of a clinical data repository whose core consists of four years of inpatient administrative and billing data from the mainframe legacy systems of the University of Virginia Health System (UVAHS). To these data we have linked a cardiac surgery clinical database and our physician billing data (inpatient and outpatient). Other databases will be merged in the future. A relational database management system (Sybase) running on a dedicated IBM RS/6000 minicomputer was employed to assemble 2.5 Gigabytes of core data describing approximately 100,000 hospital admissions over the four year period. To enable convenient data queries, the system has been equipped with a custom-built WWW user interface, which generates Structured Query Language (SQL) automatically. We illustrate the rapid reporting capabilities of the resulting system with reference to patients undergoing coronary artery bypass graft surgery (CABG). We conclude that this information system: a) constitutes a convenient and low-cost method to increase data availability across the UVAHS; b) provides clinicians with a tool for surveillance of patient care and outcomes; c) forms the core of a comprehensive database from which clinical research may proceed; d) provides a flexible interface empowering a wide variety of clinical departments to share and enrich their own clinical data.

INTRODUCTION

In recent years, strict budgetary constraints induced by the widespread introduction of managed care have resulted in renewed urgency for information availability across the health care enterprise. Information systems are required by hospital administrative staff for planning and policy, and by clinicians for evaluation of process and outcomes of clinical care.

Legacy hospital information systems -- while providing adequate support for patient care and billing -- are poorly suited for meeting the information needs of hospital management and staff. Data requests from hospital staff often cannot be serviced in a timely manner. Also, those hospital and clinical departmental systems installed at various times over the last twenty years are rarely compatible. Thus, creation of a dataset suitable for analysis often demands time-consuming data manipulation such as conversion and merging. Solutions to such problems include development of software to merge¹ or interface between² disparate clinical information systems.

Our approach is to create a core Clinical Data Repository (CDR) by loading in batch four years of administrative and financial data into Sybase RDBMS. These data represent 100% of inpatient admissions. Although a similar strategy has been adopted by others^{3,4}, we feature the following: a) assurance of patient and physician confidentiality by removal of identifying information and the adoption of disguised internal identifiers; b) a custom built WWW user interface which generates automatically Structured Query Language (SQL) in response to point-and-click action by the user; c) a demonstration of how the CDR may be linked to departmental clinical databases.

Thus, the major advantage of the system is to enable the user to submit interactive and rapid queries to data from previously disparate systems, now at a single location and in a single data format. The CDR is accessible over the University Intranet using Netscape. It is closely associated with the University of Virginia Health System (UVAHS) real-time data systems but is sufficiently flexible to enable regular independent enrichment of its clinical data content. We illustrate how the CDR may be used as a starting point for both clinical research and administrative reporting.

METHODS

Hardware and Software

Our choices were driven by requirements for economy and rapid software development. We currently use Sybase Version 11.0 (Sybase Inc., Emeryville, CA) on a dedicated IBM/RS6000. The machine is running AIX UNIX operating system version 4.3 and is equipped with 320 MB memory.

Data loading into Sybase is achieved using custom written Practical Extraction and Report Language (Perl -- <http://www.perl.com/perl/index.html>) programs⁷ enhanced for interfacing to Sybase (Sybperl -- http://reality.sgi.com/pablo/Sybase_FAQ/Q9.4.html -- <http://www.mbay.net/~mpeppler>) (See Figure 1).

The custom (password-protected) WWW user interface was implemented using HTML, JavaScript, and the 'C' programming language. Advantages of this Netscape access to the CDR are at least fourfold: a) Netscape is platform-independent; b) it is familiar to many users; c) it requires no special software installation or maintenance at user workstations; d) it includes data download capability. Regarding the latter issue, data subsets resulting from queries may be downloaded directly to Microsoft Access or Excel for analysis (see Figure 4 below).

Owing to the variety of platforms and software used by the staff at a teaching medical center, a choice of data query and display methods is offered. Those users unfamiliar with Structured Query Language (SQL) may use the WWW user interface to build their queries. More experienced users may configure their workstations to access the CDR directly using Microsoft Access. Alternatively, those authorized users familiar with SQL may submit their own custom queries to the system (see Figure 1).

Repository Design and Protection of Data Confidentiality

A Steering Committee of senior clinicians has been established to guide the development of the CDR and to propose policies on its utilization and access. CDR access will require prior authorization and an assigned user identification that will permit automated tracking of use of the system. The essential features of the information system are displayed in Figure 1. The data are partitioned into two Sybase DBs -- "Production" (accessible by all authorized users of the system) and "Secure" (restricted access). A data dictionary consisting of

two Sybase tables and which serves as a driver for the WWW user interface system (see below) has been incorporated into the "Production" database. Regarding the "Production" database: a) all medical record numbers (MRNs) and physician numbers are replaced by disguised identifiers allocated internally by the loading software; b) patient personal data elements (e.g. address, date of birth) are dropped. The tables associating the disguised with the genuine numbers are located in the "Secure" database. In this way, all data pertaining to a patient or a clinician are grouped appropriately, while the confidentiality of the individual is preserved.

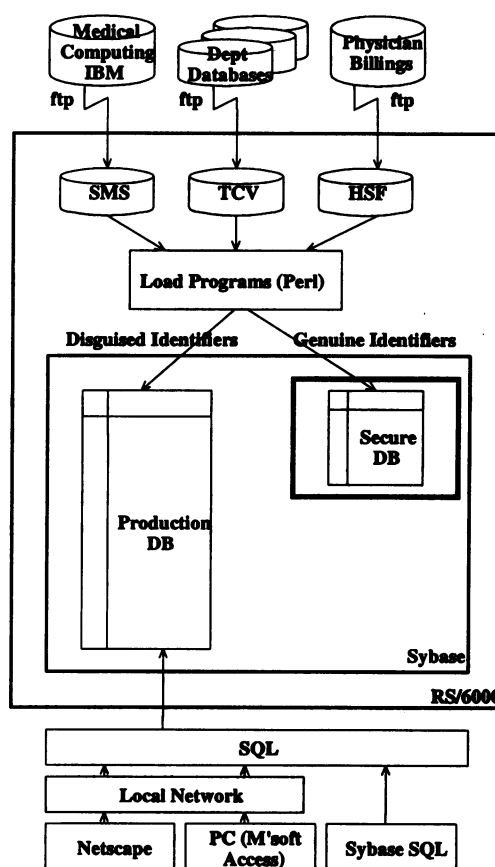


FIGURE 1 - System Overview. The upper sections show data sources, loading software and Sybase databases, and the lower sections depict the query, download and display systems currently supported. Key : SMS=Shared Medical Systems(Malvern, PA); TCV=Thoracic/Cardiovascular departmental data; HSF=Health Services Foundation (Physician billing data); DB=database; SQL=Structured Query Language; PC=Personal Computer.

Creation of the Data Repository

All data enter the Sybase RDBMS system via load programs written in Perl and SybPerl (see Figure 1). A separate program exists to load each ASCII data source in batch mode. During the load process, sensitive data pertaining to individuals is disguised or omitted (see above) and the data are enriched by the addition of calculated fields describing patient readmission and comorbid conditions.

True to the original conception of the CDR as a retrospective system (as opposed to a real-time patient care tool) the system is currently updated on a quarterly basis. This approach was adopted to produce a useable system in an economical and timely fashion, but plans exist to establish direct HL-7 links to the hospital information systems to enable more frequent updating of the CDR.

Creation of the WWW User Interface

In order to ensure that the repository is used by a wide variety of authorized hospital personnel, we created an interface that generated SQL following point-and-click actions by the user (see Figure 2). Briefly, the interface consists of a collection of HTML pages (e.g. one for each of the four panels of Figure 2) in which are embedded JavaScript functions which communicate with a set of Common Gateway Interface (CGI) 'C' programs via the WWW server. The behavior of the CGI software is determined by the data dictionary tables in the "Production" Sybase DB.

RESULTS

Repository Queries Using WWW Interface

To build a query the user is presented with a WWW page consisting of four frames (see Figure 2). The frame on the bottom left consists of a list of data categories in the database. The user selects a data category by clicking on its name which will produce a list of individual data elements in the frame at bottom right. Online help for each data element may be obtained by selection of its underscored name. Each data element to be included in the query result may be selected by clicking its associated "Select" button. Once selected, the name of the data element is added to a list on the left of the center frame (see Figure 2).

The rows of the query result may be constrained by clicking on the "Condition" buttons for each data element. This action will invoke a new window in the center frame that will enable the user to set conditions for the data element concerned. For example, the user may enter a list of DRG codes if the DRG

"Condition" button is selected. In Figure 2, the user has selected all patients who underwent coronary artery bypass graft surgery (CABG).

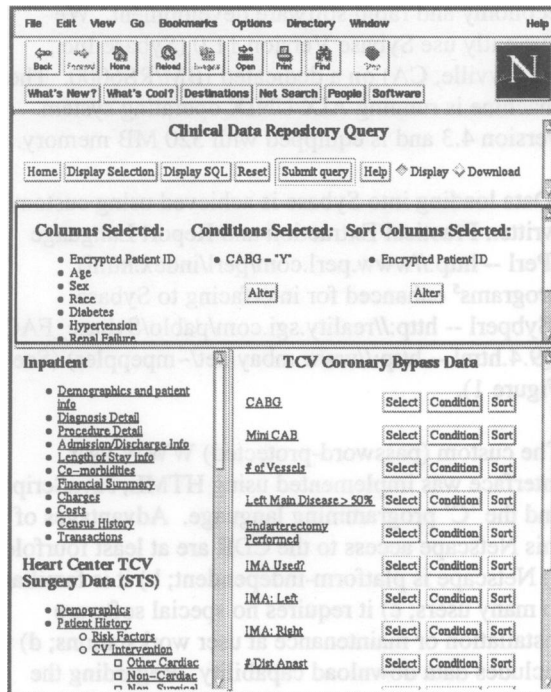


FIGURE 2 - SQL-Generating WWW Interface. The center frame describes the query about to be submitted to produce output shown in Figure 3.

Pt.ID	Age	Sex	Diab	Hyper-tension	Renal Failure	EF	Total LOS	Post Op LOS	# Re-adm	Days Next Adm	# of Vessels
588	63	M	Y	N	N	50	9	9	0		Triple
11056	46	M	N	Y	N	40	9	5	1	18	Triple
13502	55	M	N	Y	N	62	4	3	0		Triple
18437	52	F	Y	Y	N	62	67	65	0		Triple
25369	59	M	N	N	N	64	6	6	1	5	Double
32430	70	M	N	Y	N	60	19	18	0		Single
32962	51	M	N	N	N	35	5	5	0		Triple
33615	67	M	Y	N	N	45	9	4	0		Triple
34635	73	M	N	Y	N	66	5	5	0		Triple
34842	68	M	N	Y	N	45	11	6	1	8	Triple
44951	63	M	N	Y	N	35	11	3	0		Triple
48971	59	F	Y	Y	N	50	9	9	3	2	Triple
51865	62	F	Y	Y	N	57	10	7	0		Triple
51215	56	M	N	N	N	67	4	4	0		Triple
53382	66	M	N	Y	N	59	4	4	0		Triple
54570	55	M	N	Y	N	45	8	7	0		Triple
54580	60	M	Y	Y	N	60	5	5	0		Triple
54599	65	M	N	Y	N	45	5	4	0		Triple
54614	53	M	Y	Y	N	64	5	4	1	57	Triple

FIGURE 3 - Typical Query Result

Similarly, the sort order of the query result may be manipulated by selecting the "Sort" buttons for each data element. The user may view the SQL statement at any time during its construction by selecting the "Select SQL" button in the top frame. When the user has completed formulation of the query, it is submitted to the WWW server by selection of the "Submit Query" button, also in the top frame.

Figure 3 displays an example of results obtained from the CDR on submission of the query outlined in Figure 2 above. Patients selected were all those that were flagged in the departmental data as having undergone coronary artery bypass graft surgery (CABG). The CRD currently contains Thoracic/Cardiovascular (TCV) surgery patients discharged between 1/94 and 3/97 inclusive at UVAHS (n=2300) and of those all CABG cases were selected in this example query (n=1917).

Note that the data elements selected to produce the example shown in Figure 3 have been drawn from two databases (SMS/administrative: disguised patient identifier, age, sex, total length of stay (LOS), postoperative LOS, number of subsequent readmissions, days to next readmission; departmental TCV surgery: comorbidities diabetes, hypertension and renal failure, left ventricular ejection fraction, number of vessels operated upon). This illustrates one of the advantages of the RDBMS approach. Data from a variety of sources may be drawn together in configurations not possible employing current hospital information systems.

System performance varies depending upon complexity of the query and computer network traffic -- but results of querying some 100,000 hospital admissions and related clinical tables are generally obtained within 3-4 seconds when using indexes, or a few minutes when scanning the database sequentially.

Reporting Capabilities of the System

From the SQL query result screen shown in Figure 3, data may be downloaded directly into Microsoft Excel at the user workstation by selection of the "Download" button. Figure 4 graphs data downloaded from a query similar to that outlined in Figures 2 and 3 above -- the difference being that a wider range of clinical risk factors were selected from the CRD in this latter case.

An issue of interest to both administrative and clinical staff is patient length of stay (LOS). Figure 4 shows mean total LOS (derived originally from SMS/Administrative data) stratified by specific risk

factor (derived originally from TCV surgery data) for all CABG patients between 1/94 and 3/97 inclusive (mean age = 64 years, mean total LOS = 9 days, median total LOS = 7 days).

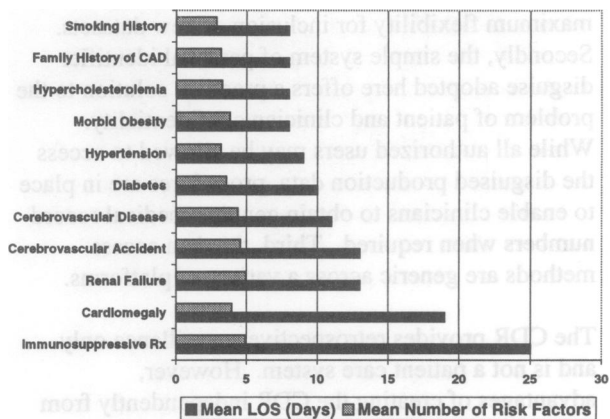


FIGURE 4 - Mean Total LOS and Mean Number of Risk Factors for CABG patients stratified by specific risk factor.

The LOS data shown in Figure 4 (i.e. the black bars) are complicated by the fact that the many of the 1917 patients have contributed to averages in multiple risk factors, resulting in overlapping categories. For this reason, we have included the mean number of risk factors for each specific risk category shown (gray bars). The data suggest that -- although patients with risk factors associated with higher LOS tend to have a greater number of risk factors overall -- it is the presence of the specific risk factors themselves which determine the total length of stay.

DISCUSSION

We demonstrate the creation and use of a clinical data repository comprising 4 years of inpatient data at a 700-bed teaching health facility. The work has been performed within an eighteen-month period, using 2 FTE's, a Sybase license (\$8,000), and an RS/6000 (\$20,000).

To the administrative and financial data core we have linked an inpatient departmental clinical database whose demographic and administrative data elements were assembled in an independent fashion. Following detailed inspection of those unlinked data remaining after automated matching of the MRN, admission date and discharge date of the 2 datasets, 2298 from a total of 2300 patient admissions recorded in the departmental TCV database (i.e. 99.9%) were eventually linked to their financial data

counterparts. This suggests that the quality of the departmental data is high and inspires confidence in the system overall.

The advantages of the CDR are clear. First, the use of a modern RDBMS such as Sybase affords maximum flexibility for inclusion of new datasets. Secondly, the simple system of personal identifier disguise adopted here offers a practical solution to the problem of patient and clinician confidentiality. While all authorized users may be allowed to access the disguised production data, procedures are in place to enable clinicians to obtain genuine medical record numbers when required. Third, the data access methods are generic across a variety of platforms.

The CDR provides retrospective surveillance only and is not a patient care system. However, advantages of creating the CDR independently from the patient care system include a) the patient care system is not impacted by queries made to the CDR; b) the data may be disguised upon loading into the CDR, thereby allowing a greater range of users access to the data. It is anticipated that the CDR will be helpful in such issues as data sharing with other hospitals, and response to data inquiries from third party payers and other institutions.

Currently, however, it should be noted that the clinical data of the CDR are sparse for some purposes. For instance, ICD-9-CM codes are the principal means by which patient populations may be identified. In the near future, however, we intend to load matching outpatient visit information, medications and laboratory data. More clinical departments will also be encouraged to include their datasets (including the hospital cancer registry) and further enrich the clinical data content of the system.

Figure 4 above illustrates how authorized hospital personnel may relate clinical resource utilization to critical aspects of patient care such as clinical outcomes (crucial for patients and heart surgeons) and financial outcomes (crucial for the survival of the institution). Thus, the system may be used to monitor the process and outcomes of patient care, and to identify opportunities for improvement of patient care. Table 1 lists the categories of studies that will be performed at UVAHS with the CDR in its current state of development.

In summary, we believe that the CDR system will potentiate the institution's intention to optimize cost-efficiency while maintaining or enhancing quality.

TABLE 1 - Using the CDR to Monitor and Improve Patient Care

Process of Care
- Clinical Pathways
- Structural Changes
- Care Guidelines
Use of Resources
- Length of Stay (LOS) - (Intensive Care)
- LOS (Hospital)
- Readmission
- Cost
Feedback to Physicians
- Patient Outcomes
- Complications
- LOS
- Resource Use

As the legacy systems of the UVAHS are gradually replaced, this system will emerge as the data repository of new institution information systems. The CDR therefore reflects an institutional philosophy of support for a dynamic data repository, with emphasis on provision of retrospective data analysis for hospital personnel.

References

1. Marrs KA, Steib SA, Abrams CA and Kahn MG. Unifying Heterogeneous Distributed Clinical Data in A Relational Database. Seventeenth Annual Symposium on Computer Applications in Medical Care. 1993; 644-648.
2. Gendler SM, Friedman BA, and Hendricks WH. Using Hub Technology to Facilitate Information System Integration in a Health-Care Enterprise. Am. J. Clin. Pathol. 1996; 105(4 Suppl. 1): S25-32.
3. Ruffin M. Developing and Using a Data Repository for Quality Improvement: The Genesis of IRIS. Jt. Comm. J. Qual. Improv. 1995; 21(10): 512-520.
4. Prather JC, Lobach DF, Hales JW, Hage ML, Fehrs SJ and Hammond WE. Converting a Legacy System Database into Relational Format to Enhance Query Efficiency. Nineteenth Annual JAMIA Proceedings. 1995; 372-376.
5. Wall L, and Schwartz RL. Programming Perl. O'Reilly & Associates, Inc. 1991. ISBN 0-937175-64-1.