ARTICLES

Managing Clinical Research Information: A Case Study in Information Access, Presentation, and Analysis

Raza Hashim, PhD; Thomas L. Lewis, MD; Stephen J. Rosenfeld, MD

ABSTRACT
Medical information systems at the NIH must meet the twin challenges of excellent patient care and innovative clinical research. This article describes the approach taken by the clinical center at the NIH to enhance its information systems to meet these challenges. The four key components of this strategy are the standard clinical desktop, the clinical data repository, the repository access package, and protocol mapping. Together, these tools provide significant new capabilities for patient care staff and clinical investigators.

KEYWORDS
• Clinical research
• Protocol management
• Clinical data repository
• Lexicons
• Data analysis and presentation
• Clinical guidelines

The Warren Grant Magnuson Clinical Center (CC) is a 325-bed research hospital that serves the intramural program of the National Institutes of Health (NIH). The CC represents nearly 50 percent of all NIH-funded clinical research center beds in the nation. With approximately one thousand active research protocols, the CC admits patients from around the nation and around the world for clinical trials and natural history studies, and all care is provided.

Note: We wish to acknowledge the following technical contributors: Michael Cai, Jeff Haver, Murali Kumar, Yenshei Liu, Nasir Majeed, Jon McKeeby, Mark Miller, Steve Moore, Michael Niv, Jason Schanuel, Mike Staley, and Jason Tsai.
free of charge. The CC’s mission is to further clinical research, provide superlative care to its patients, and to do both in the most efficient manner possible. An appropriate information technology infrastructure is critical to achieving these goals.

There is no clear demarcation between clinical research and patient care at the CC. Our information systems must meet the day-to-day information demands of physicians and nurses providing inpatient and outpatient services, as well as provide additional and aggregate data to support protocol management, data mining, and analysis of data from large groups of patients. Because every patient admitted to the CC is on a research protocol, these additional data must be available with the same ease of use as the traditional, patient-centric view of the medical record.

The present system has grown by accretion, not design. It consists of a comprehensive hospital information system (Eclipsys 7000) that has been in place in one form or another for twenty-five years. The functionality of this core system is expanded by ancillary systems such as a laboratory information system (SCC), blood bank systems, a cardiology system, and so on. Ancillary systems are interfaced with the HIS to differing degrees. Currently, approximately 90 percent of the medical record is captured in electronic form, including physician orders and nursing documentation.

Where possible, our systems have been designed around COTS (commercial off-the-shelf) products marketed for clinical care. Research and management needs were historically served through custom extraction processes. These processes were difficult to support and inflexible, and difficult to adapt to changing needs. To serve the clinical, research, and management communities better, we embarked on a repackaging of our core systems to provide a standard, expanded human interface to information sources (the standard clinical desktop), complete access to the longitudinal patient record for both clinical care and research (the clinical data repository), a set of tools for using these data (the repository access package), and a new tool for authoring, documenting, and managing protocols (the protocol mapping project). The following section describes the architectural and design criteria on which these four components are built.

Architectural and Design Considerations at NIH

We had to face several challenges while designing the information infrastructure at the clinical center at NIH. The positives included a medical record that is 90 percent electronic. Physicians are significant users of the electronic system, with 90 percent direct physician order entry. The four thousand staff members using our electronic systems are computer-literate and Web-aware and have high standards for system use and functionality. Our computing environment is heterogeneous with respect to operating systems and hardware platforms, including Macintosh (one of the largest nationally), PCs, and
multiple varieties of UNIX. An extensive archive of medical data is available in electronic form, including several hundred million observations collected over the past twenty-five years.

We also encountered several significant negatives. Although current patient care information is readily available in our hospital information system, it is not possible for patient care and clinical research staff to access any of the historical data on-line. The data format is complex and awkward to process with standard tools. Data cannot be interchanged easily among different hardware, software, operating system, or application platforms. Metadata describing the content, format, interpretation, and historical evolution of the clinical data are not available to either end users or application designers.

After carefully assessing the needs of clinical investigators and patient care staff in light of existing electronic patient care systems, we established the following architectural and design guidelines.

**Repository Content.** The data should be complete, comprehensive, consistent, reliable, and timely. All medical data generated by CC medical information systems should be available to authorized users anywhere (on-site, off-site) and at any time (seven days a week, twenty-four hours a day). The repository should be seamlessly integrated with the transaction-oriented patient care systems. Data should be available in a timely manner in the repository within some time interval after being generated; for example, lab results are available ten minutes after the lab technician certifies them in the lab system. Any part of the medical record generated using computer systems should be available to the physician and researcher in an electronic format. The data should be internally consistent and reliably available across system boundaries.

**Secure and Universal Access.** The data should be securely available to all authorized users across systems. It should be possible to manage the security policy from a centralized location. Data should be available to any authorized user at any time and any place—that is, there should be Web access to order entry, results retrieval, and specialized RAP data presentation.

**Presentation Tools.** Effective presentation tools are needed for different kinds of data (time series, scatterplots, multimedia, and so on) to different kinds of users (doctors, nurses, administrators, financial managers, and so on). The current presentation tools do an inadequate job of data presentation to various users. It should be possible to browse patient data in both patient-centric and protocol-centric views. CDR design should support complex queries and browsing in a responsive and easy-to-use way.

**Data Interchange.** The data should be available across system boundaries in an interchangeable format in a secure fashion. Because most researchers do their data analysis using standard tools like Excel, SAS, or Mathematica, it should be possible to export CDR data to these tools. Data should be stored so that they are easily accessible by multiple operating environments and systems (VM, UNIX, NT, and so on) and institute-specific systems.
**Systems Design Technology.** The systems design should maximize the use of standards-based technology, including object-oriented design, modular components, relational database technology, HL-7, XML, Java, open systems, and the use of standard tools.

**Metadata.** Metadata should be explicitly represented in the enterprise environment to facilitate application development, data presentation, and data management. This allows accurate representation of data over extended time intervals, normalization of lab data that have been subject to changes in reference ranges, and improved systems design. Several authors have pointed out the importance of metadata and lexicons.\(^1,2\)

The next section describes the top-level navigation tool for our clinical information environment: the standard clinical desktop.

**Standard Clinical Desktop**

The standard clinical desktop provides a common gateway for access to data and applications for physicians, nurses, clinical investigators, administrators, and others with patient care and research responsibilities at the NIH. The intent is to have complete “one-stop-shopping” access to all relevant information resources and productivity tools throughout the clinical center. The desktop is icon-based, easy to learn and navigate, and compact. It simplifies management of our extensive user and application environment, provides for central management and deployment of applications, and enhances our ability to implement a secure infrastructure. It contains four categories of tools to assist our staff, which may be described generically as follows: tools to support the care of a specific patient, tools to access the medical literature and related data sources, clinical research tools, and personal productivity tools. The standard clinical desktop contains off-the-shelf components. The additional functionality it provides is the result of the ability to exchange information among these off-the-shelf components.

**Patient-Specific Tools.** Patient-specific tools are those used in the direct care and treatment of individual patients. The primary application is the CC medical information system, a comprehensive system used by physicians to write orders and retrieve results, by nurses to document care, and by ancillary departments to manage requested services and report results.

**Tools to Access the Medical Literature.** A variety of tools is used to access the medical literature and related medical databases. Some of these databases are national in scope; others are based in the clinical center intranet. A Web browser allows virtually unlimited access to World Wide Web–based medical resources. Among the nationally available databases are a desktop icon giving direct access to the National Library of Medicine Pub Med. In addition, there is a bookmark file to simplify access to a variety of medical literature sites with a single click. The electronic version of *Harrison’s Textbook of Medicine*, drug interaction databases, and drug fact sheets facilitate care by being instantly
available to users without necessitating a trip to the library thirteen floors down or stocking a copy of each on fifty nursing units and clinics. The CC intranet contains various local databases essential to the clinical research program of the NIH. These include a database containing consent forms for approximately one thousand active research protocols. Protocol consent forms on the Web are quickly available for printing and are easy and inexpensive to maintain in compliance with regulations governing human subjects research. Other applications include on-line training courses in biosafety, computer security, and nursing procedures. Various hospital policy and procedure manuals are available for reference.

**Clinical Research Tools.** The repository access package (RAP), a set of tools to assist physicians in analyzing data for patient care and research, is an increasingly prominent part of the standard clinical desktop. This application includes clinical and research presentation tools that go beyond those generally available for patient care but are of critical importance to the clinical investigation mission of the CC. This application is discussed at some length in a subsequent section of this article.

**Personal Productivity Tools.** Personal productivity tools (word processors, spreadsheets, presentation applications) are used universally by CC staff both to improve day-to-day activities and to support research. To send confidential patient information among medical providers, the desktop provides secure e-mail access using a Web-based SSL-enabled product.

The standard clinical desktop environment has been used throughout the center for the past four years and has proven immensely popular with our staff. Acceptance was almost instantaneous, thanks to both the ease of use and the wealth of applications available. Over three hundred workstations are now available throughout the hospital at inpatient nursing units and outpatient clinics. This approach has also proven to be a desirable metaphor for structured system enhancements.

Figure 1 shows a screen view of three components of the standard clinical desktop, illustrating the ability to go from a patient diagnosis of pheochromocytoma to *Harrison’s Textbook of Medicine* for more detail and then look up the effect of relevant medications.

**Clinical Data Repository**

The clinical data repository (CDR) is a historical archive of approximately twenty-five years of clinical data (1976 to 2000) from the MIS system. These data were previously kept as a tape archive. The CDR project involves developing data models and writing cleansing scripts to convert the data from a proprietary data format to a relational format. The CDR is a relational database implemented in Sybase 11.5 for a Sun Ultra 4000 running Solaris 2.6. The repository is connected to the hospital system and various ancillary systems through an HL-7 gateway.
**CDR Data.** Structured data in the CDR consist of lab orders, lab results, microbiology reports, blood bank reports, vital signs measurements, and medications. Unstructured, or free text data, consist of radiology reports, bone marrow reports, cytology reports, surgical pathology reports, nuclear medicine reports, and EEG reports. The repository contains about 150 gigabytes of raw data. The microbiology and lab results portion of the repository is about 50 gigabytes in size and has about 80 million result records. We are currently working on getting other portions of the repository cleansed and on-line. New data from the hospital information system and various ancillary systems are fed to the repository over an HL-7 interface. Currently this includes ADT, lab orders, lab and microbiology results, medications, and radiology reports.

**CDR Design Considerations.** During the design of the CDR we had to take several factors into consideration, including metadata analysis and management, data cleansing, data modeling, volume of data and data management, and system performance.

**Metadata.** Although we had historic data, we did not have any metadata available. During the CDR conversion process, the metadata was gathered in a bottom-up fashion while processing the historic tapes. For example, in the case of lab data, reference ranges and test names changed over time. Therefore, it was necessary to create metadata structures to answer even simple questions.
such as, “Give me a table of serum calcium values on patient x for the date range y.” Note that serum calcium is present in the repository as several tests with different normal ranges. Moreover, the units to measure serum calcium have changed during the past twenty-five years. This metainformation is present in the data only implicitly. In order to make the repository useful for the end users, this information had to be made explicit using equivalence class mappings.

**Data Cleansing.** By “data cleansing” we mean the process of validating, editing, transforming, mapping, and deciding what data to keep as we move data from their structure and format in older systems to the new representation in the CDR. Sometimes the transformations are reasonably straightforward. For example, a test result may have been expressed in two different units (grams/liter and milligrams/milliliter) that are mathematically identical but visually different. Such cases must be identified and mapped to a newly defined metadata structure that uses the same units in representing the results, regardless of how the test result was originally displayed.

Slightly more complex is the case where the units are different, and not mathematically equivalent, for the same test. An example might be grams/deciliter and milligrams/milliliter. To normalize these results, it would be necessary to multiply the numeric result of the first test by ten if the goal is to have both results use the same units and be comparable on charts and graphs. Still other cases arise as the data are examined more closely. Many of these issues relate to the underlying database and architectural design of the system. Many systems used a fixed number of digits to store a numeric result, with the decimal point implied or defined external to the data. With this design, it is possible to have a valid numeric result that is either too large or too small to fit in the result field. Usually the numeric result is then placed in a free text comment. These must be parsed and extracted if the data are to be useful. The risk of error is substantial and custom programming is required.

**Modeling.** Both lab and HIS data were in hierarchical data structures that worked well for high-performance interactive systems but were not flexible enough for CDR user requirements. Although the data were present as attribute-value pairs in the hierarchical structure, the attributes were often overloaded and frequently contained miscellaneous processing instructions for the interactive system that necessitated extensive logic to extract the significant medical data. Data extraction involved mapping the hierarchical structure to HL-7-like attribute-value pairs followed by mapping to a newly designed relational model. The relational structure may then be easily queried and managed with standard database tools.

**Data Management.** Handling twenty-five years and 150 gigabytes of data in a UNIX and relational database environment provided significant challenges. Standard database design principles like normalized data models made the process of loading data time-consuming and impractical. Several techniques
were adopted to imbed time-stamp and batch information in keys so it was easy to load and back out data by year, or HIS tape, or batch, and so on. We maintained metadata tables to keep track of various statistics on the database in order to estimate the time it would take to do various queries.

**Performance.** There are three main users for the CDR: interactive, batch and bulk report processing, and researching. Each of these users puts different performance constraints on the repository. The interactive user looks at the data one patient at a time. At any time there are several interactive users accessing the repository. The report processing user is interested in obtaining single-patient or cross-patient reports over a set of patients. Usually, a few users in this category are on the system at the same time. The researcher is interested in cross-patient queries for a set of patients on a protocol or is looking at individual test values across protocols. In addition, the researcher may be executing a series of complex multipatient interactive queries where subsequent queries are dependent on the result of the current query.

A simple relational model with indexing could not perform adequately for these categories of users. We mapped the relational model into an object-based hierarchical blob repository—a collection of related data items, stored and retrieved together. The indexing was done using the relational tables. Instead of containing a result, each row in the table was a pointer to an object, stored on disk, that contained a category of information like a lab test with many results. Special procedures were written to keep the relational repository synchronized with the blob repository. The interactive and report users were pointed to the blob repository. Because both these categories of users want a particular type of data—the electrolyte panel over a date range, for example—the blob repository resulted in fetching all required data in few retrievals, one per test in the panel, as opposed to a result set retrieval from a relational database. Using this technique, the blob repository was able to support up to 150 simultaneous interactive users without significant performance degradation. The corresponding relational implementation could only support 20 simultaneous users because of the data volume involved in interactive queries.

We have developed several applications around the CDR that allow the users access to the data in the repository. Two of these applications are described next.

**Metadata Management and Report Design Application.** This application has been written to maintain the historical lab metadata and help in designing cumulative lab reports, which may contain lab results for the past twenty-five years. The application allows the user to define tests and then map tests into equivalence classes. Figure 2 shows the preparation of a chemistry report illustrating the metadata associated with such a chemistry report, including the management of multiple serum calcium tests. The highlighted line is the equivalence class for calcium. The members of this class are shown in the bottom right corner. This application allows users to define new reports by mapping
equivalence classes into reports. Once a report is interactively defined and named, it is possible to request the named report on any patient in the repository. In the previous system, defining new reports or making changes to existing reports was a very tedious process that required several hours and the involvement of IS staff.

**Discharge Diagnosis Entry and Reporting System.** The goal of this application is to integrate medical research diagnostic analysis and coding with the CDR. The application allows for edit and entry of discharge diagnoses through an interactive application. The data entered are checked for accuracy and consistency. Information previously entered for a patient is immediately available for review and comparison if needed. The system also provides standard reports for discharge and traditional medical record applications. The application was written in PowerBuilder. Some Perl scripts produce the bulk reports that are generated by this application. The system is being developed using Sybase 11.5 for a UNIX environment. This application has been in use for one year.

**Repository Access Package**

The repository access package (RAP) is an application that allows physicians to access both historical and current clinical data on active and inactive patients from the CDR. RAP provides clinicians access to the longitudinal patient
record, thus overcoming the limitation of the hospital information system and the paper record. It is possible for clinicians to maintain their own patient lists and browse specific portions of the medical record, like selected tests, over this list of patients. RAP has tabular and graphic viewers for data and has facilities to export data to other applications like Excel. RAP not only gives users the ability to view data graphically but also gives them the ability to drill up or down the data using a graphical paradigm. Several user-interface designers have pointed out the advantages of using a graphical presentation of the medical record.3,4

RAP is a three-tiered client-server application. The RAP client runs on any browser that supports Java 1.2. Because RAP is a Java application, it runs on UNIX, Windows, and Mac clients. The RAP middle tier is also written in Java 1.2. It works on a Sun Ultra 4000 running Solaris 2.6. It talks to front-end client applets using RMI and retrieves information for the database using JDBC. The server is multithreaded and can handle simultaneous connections from multiple clients. RAP also provides interactive access to the traditional reports that are a part of the patient’s paper record. These reports are generated from the data repository using report-generating programs written in Perl 5.

To date, we have developed viewers for searching, demographics, visit history, vital signs, radiology reports, and lab data. These viewers allow users to look at data on patients one patient at a time in a variety of graphical and tabular formats.

Patient Find Viewer. This viewer lets the user search the patient database and retrieve information about the patients using the associated viewers.

Visit and Protocol History Viewer. It is possible to view the visit and protocol history of a patient using this viewer in RAP.

Lab Selection and Viewer. This viewer allows the user to select data for a date range by lab type and display the data in tabular and graphical format. Before, it was only possible to select data by lab test number; yet, it was possible for the same lab test to have multiple lab numbers.

Microbiology Selection and Viewer. This viewer lets the user browse the microbiology data for a patient. The data are displayed in a tabular format.

Vital Signs Viewer. This viewer lets the user plot and view in tabular format the vital signs data for a patient. In the current version of RAP, this viewer has been temporarily disabled because of data reliability issues.

Radiology Report Viewer. This viewer lets the user look at radiology reports for a patient. This viewer too has been temporarily disabled because of data reliability issues.

Standard Reports Viewer. This viewer lets the user see the standard NIH reports, such as the discharge diagnosis unit index and the cumulative chemistry report, from within the RAP application.

Figure 3 shows the graphical view of lab data for a patient with aplastic anemia. Trends are easy to discern. For precise quantitative information, the
Panel A
This panel illustrates the primary user interface elements of RAP, including frames, tabs, and lists.

Panel B
This panel illustrates the graphing tools available in RAP. Each of the graphs shown is resizable, and the time scale can be modified by explicitly selecting values or by click-and-drag. This view allows the researcher or clinician to evaluate long-term trends across visits, such as the rising MCV (lower left panel) or fluctuation in platelet counts (lower center panel).
user can select a particular section of the graph and then drill down. All other graphs synchronize to the time dimension of the selected graph. The researcher can easily switch between the graphic and tabular view of data. It is also possible to export the data to an analysis tool like Excel for more data analysis.

Protocol Mapping and Management

Every patient admitted to the CC must be entered on a protocol. Protocols are used to specify detailed therapeutic interventions or to provide a structured framework for studying the natural history of a disease. Although protocols bear some resemblance to clinical guidelines, the language used to describe them must be rich enough to capture the entire clinical course of a patient. As far as possible, contingencies should be anticipated and incorporated into the protocol, and the structure should be extensible on an ad hoc basis to incorporate unforeseen contingencies. This level of description is necessary in a research environment because analysis of variances may lead to critical insights.

Protocols as described here are a natural organizational structure for the medical record that complements the traditional patient-centric view. They can provide the basis for both ordering and documentation, and promote research functions to the same level as patient care functions in the system. Such research functions include management and analysis of protocol-based cadres of patients for accrual tracking, periodic safety review, and efficacy analysis, activities that would otherwise require periodic database extractions or the maintenance of a parallel research information system.

Formally building protocols into our information system infrastructure marries high-level research functions with the daily delivery of care. Protocols have always been central to patient care at the NIH. Typically, a principal investigator writes a relatively general description of a research study, including the scientific background, criteria for eligibility and response, statistical power, benefits and risks to the patient, and so on. This document is reviewed for its scientific merit, ethical issues, and costs. When approved, it forms the basis for a plan of care that is embodied in customized screens in the hospital system (for ordering) and a set of data collection instruments (case report forms, spreadsheets, customized databases, and so on). Each protocol is developed individually, and each set of ordering or recording customizations is also developed individually. The protocol mapping initiative standardizes each procedural step of this process without limiting its content. Using object-oriented tools and data structures, protocols can be built from scratch using a palette of components or by changing only what is necessary in an existing protocol. The benefits to the research effort are clear: the process of writing a protocol is streamlined, and the use of
standard components allows secondary research questions to be asked across protocols.

A pilot version of protocol mapping has been successfully completed, and we are currently expanding the project to include encoding of all new protocols and documenting a selected set of protocols. Issues we are now addressing include interfacing protocol maps to the core hospital information systems using an HL-7 compatible interface, establishing a strategy for standardizing terminology across protocols, and exporting protocol representations using XML and guideline interchange format (GLIF).

Conclusions

This article describes our approach to the coherent management of complex computer systems for patient care in a clinical research environment. The standard clinical desktop has proven itself over the past four years as an organizational and navigational paradigm for clinical systems by bringing multiple applications together. The data repository is now available for both on-line and batch research queries, replacing the old off-line batch system. It is also serving as a framework for gathering metadata about our current and historical systems. The repository access package provides an interactive look at the patient’s medical record and serves as an evolving tool set for clinical researchers. The protocol mapping tools will serve as a protocol-centric tool for organizing clinical trials as it organizes both the management of the trial and clinical data collection.

We consider the four proven components described here to be the essential elements of our clinical information management framework. With this framework in place, we are now in a position to ensure that our information technology infrastructure evolves, both our core hospital information system and all other current and future components as well. This open and modular design allows us to add new functionality and replace existing pieces of the current infrastructure without causing major disruptions or loss of current capabilities.

References

About the Authors

Raza Hashim, PhD, is chief of the Informatics Research and Development Group at the National Institutes of Health clinical center.

Thomas L. Lewis, MD, is a consultant in clinical informatics and former chief information officer at the National Institutes of Health clinical center.

Stephen J. Rosenfeld, MD, is hematologist and deputy chief information officer for Medical Informatics at the National Institutes of Health clinical center.