Database management for clinical trials

by JOHN M. LONG and JOSEPH R. BRASHEAR

Hyperlipidemia Program
Minneapolis, Minnesota

ABSTRACT

The Authors describe how they used a standard database management system (System 2000) and a computer utility to build a sophisticated medical records system in support of a national multi-clinic clinical trial. Privacy, protocol adherence, quality control and other key elements of an ethical clinical trial were satisfied at a fraction of the development cost for the more traditional approach of building a customized system. The authors feel that old lessons learned in other areas regarding the balance of manual to automated systems and the use of standard software are being re-learned for clinical trials. Neither a medical setting nor a clinical experiment changes the basic issues of good systems design. The possibility of using clinical trials as a test bed for developing medical information systems is also proposed.

In this paper we will describe our experiences in the development of a data processing system for a national, multi-clinic, controlled clinical trial. It is hoped that the description will demonstrate that controlled clinical trials are neither unique nor specialized users of medical information systems. In fact, clinical trials could become a vehicle for testing new concepts in information system design for medical applications.

The clinical trial is a controlled experiment designed to test the effects of a medical treatment on human subjects. It is the last step in the medical research process. It is the test to determine whether a drug, surgical procedure, or medical device is safe and effective for use in day-to-day medical practice, as well as a medical research device to test alternative hypotheses.

A controlled clinical trial requires detailed analysis of the medical well-being of large groups of patients. The subject population is well defined. Major aspects of their health care are carefully controlled. Clinical data is collected in a standard manner according to a pre-defined trial protocol. Patients are followed (indeed pursued) over long time spans to insure complete data collection and statistical reliability.

Our clinical trial, funded by a grant from the National Heart, Lung and Blood Institute, is formally known as the Surgical Control of the Hyperlipidemias, Secondary Prevention Trial. We simply refer to it here as the Hyperlipidemia Clinical Trial or Study.

The Hyperlipidemia clinical trial is a definitive test of the lipid hypothesis. It is designed to seek an answer to the question of the effect of maximal cholesterol reduction on patients with known atherosclerotic heart disease. One thousand patients are being randomly assigned to either a treatment or a control group. The patients will be followed for five years after randomization. Periodic clinic visits are scheduled to ascertain the extent of atherosclerosis and to detect the occurrence of any other medical problems. Trial protocol requires a highly restrictive recruiting and screening process to obtain 1,000 patients for randomization. It is anticipated that clinical data will be collected from at least 10,000 patients during the screening phase of the trial. Twenty-two different forms are used for data collection. They range in size from a single page log-in report to a 31 page medical history and physical form. Obviously both patient management and data management are primary concerns of the trial.

In addition to these physical parameters, we were faced with a set of environmental constraints. Form design and patient recruiting had already started when computer systems personnel were hired. We expected to be inundated with forms at any moment, and certainly before systems and procedures were fully established. Additionally, in the rush to recruit, incomplete attention had been given to the practical day-to-day aspects of how to handle data and readily analyze the end results of the trial. We were forced into early recognition that actual computer processing represented only one of three major data processing systems areas. We were forced by practical circumstances to relearn this old lesson which should not have been forgotten.

Given these parameters and constraints, a system had to be developed that could be operational quickly at reasonable cost. The data processing system is a necessary tool for a clinical trial. As in all of health care, the computer is not the paramount concern and should not become a major expense item. In addition, the entire system had to be responsive and easily modified. A medical researcher simply cannot anticipate all of the effects of his procedures. Medical ethics require constant surveillance for unanticipated events and possible procedure termination.

Other constraints on system design were no different
than those facing a systems analyst in any application. Privacy and security of data must be considered whenever data is collected about individuals. Economics dictate that novel developments in supporting disciplines cannot compromise the integrity of the user and his application. Neither a medical setting nor a clinical experiment change the basic issues of good system design.

Trial management was faced with a classical computer system dilemma. There are four obvious alternatives in building a data processing system: develop a custom system, still a tantalizing alternative for a systems analyst; share customized hardware and software with other similar applications; develop a generalized hardware/software system specifically for clinical trials, or use a computer utility and available standard software. The Hyperlipidemia Clinical Trial chose the fourth alternative.

We are using a shared CDC Cyber 74 facility at the University of Minnesota and System 2000 as our database management system. This approach has proven a good one. Starting in May of 1975, we had control of documents and protocol adherence via time-sharing terminals by the end of that year. The basic system was fully operational in May of 1976. By this we mean that we were able to enter clinical data in a controlled operation, provide data accountability, and retrieve reports based on that data. Improvements have continued and some work remains to be done in data analysis and statistical reporting, but data processing, including programming or systems analysis, is not responsible for any delays. We have been able to generate all reports very shortly after they have been defined. The basic system was developed in less than one year at a cost of less than $60,000. Of course, some of these savings are lost in higher operating costs which we will discuss later.

An interesting and significant body of work has been done in peripheral areas as well, ranging from data security and privacy to the social psychology of medical data processing. We have also been able to be truly responsive to our users. A great deal of time has been devoted to gaining input and support regarding reporting formats. The generality of standard software has permitted changes with little disruption to the automated portions of our system. Overall, we feel that we have been able to emphasize those areas where emphasis rightly belongs. Rather than expend all our resources on computers and software, we have emphasized data content and editing, meaningful data analysis, and effective computer generated reports.

Our medical records database design was based on three primary concerns. First, we needed ready access to large volumes of clinical data. Secondly, we are required to monitor adherence to the trial protocol by every patient from each clinic. And finally, we had to control data flow through a complex editing and certification process. These concerns led to consideration of three separate System 2000 databases. The Main, or scientific, database would contain all of the clinical information collected for every patient, that is, the patient’s individual medical record. The Administrative database would hold the information needed to control data collection and adherence to trial protocol. A Locator database would be used to monitor the flow of data forms through the entire system.

These functions are analogous to those found in a traditional circulating library. The Main database is equivalent to the card catalog. We use it to determine what data we have on each subject. The Locator database is the analogue of the checkout or circulating system, tracing the current location or user of individual holdings.

Function or method of use dictated the structure of the three databases and resulted in three different designs within this rather loose restrictions of System 2000. The simplest final structure was that of the Locator database. The automated version was abandoned. The checkout and circulation control function appears to operate better as a manual procedure.

The Administrative database is on-line and is accessed through a time-sharing terminal. Information is added or updated as forms are received. An entry is made for every data form which arrives from a clinic. This is a fundamental step in imposing control over forms flow and protocol adherence. The database has a tree structure with three hierarchical levels. All data elements, except a text memo, are inverted. That is, they are all key elements and may be used to qualify retrieval or access operations. Complete inversion has given us the capability to retrieve formatted administrative reports with minimal programming effort.

The Main database, holding the medical records, is also maintained on-line, but is updated periodically in a batch processing mode from punched cards. Although extensive editing and consistency checks are made in the update programs, it is expected that manual control of the input data will make it relatively error free. Entry into the Main database is the last step in the manual editing and checking chain. This database is also structured as a three level, hierarchical tree but the structure is unique. Our approach was dictated by storage costs and limitations, rather than access modes. You will recall that the clinical data is submitted on a large number of complex forms. A straightforward approach would be to define a data set or repeating group with unique data elements for each form. This approach rather quickly leads to realization of System 2000 overhead costs and limitations on numbers of unique elements. The design used was to designate the third level as a two element data set containing a question number and an answer. Access to information in the database requires detailed knowledge of each form and explicit definition of data by question number. This has not been a severe restriction and is offset by reduced computer operating costs. In fact, requiring knowledge of the data form has turned into an advantage. Requests for reports have been unusually brief and rational.

By using a standard software package, System 2000, we were able to minimize development time and cost. But, obviously, the system is not an optimal system for this particular application. For example, the ratio of total database size to actual information is greater than five to one. We are paying four indexing characters for each information character. Part of this cost is attributable to full
inversion, but a fair amount is pure overhead. For example, the cost to index a date is totally exorbitant in System 2000. First, dates must be represented as ten character data elements. Further, if a date is a key element, 20 characters are required to index each unique value. System 2000 also imposes a second indexing scheme for multiple occurrences of each key value. An application requiring data set selection with date of birth as one of the criteria becomes very expensive. The Administrative database has five dates designated as key elements. Given that there is a need to index dates, we could devise a coding scheme which would minimize file overhead. But we would have to give up the straightforward System 2000 natural language capability to manipulate dates on storage or retrieval. We do not believe the increased storage efficiency would offset the resulting increase in programming costs.

We could cite a number of similar instances of balancing or optimization which have been considered. There are undoubtedly many more that have not been recognized as yet. The major point is, that by selecting a software package, we have the time now to consider optimization. The trial data system is operational. We enter data and produce reports on a production basis. We have a happy customer and the resulting leisure to consider system enhancements.

An additional benefit from use of System 2000 is reduction in continuing programming staff and costs. Reports are produced using combinations of System 2000 natural language and FORTRAN or COBOL. Data elements for a report are selected and retrieved from the database using a simple “LIST” command. The selected data sets are ordered and written onto a sequential, FORTRAN compatible, temporary file. The final report is produced either by a simple formatting program or by a package such as SPSS. The “LIST” command is easily constructed, and complex selection criteria can be readily changed. All report programming to date has been a part-time activity for a senior programmer. Report production costs have averaged less than 50 cents per page.

So far our decision to use a computer utility and standard software appears to be a good one. The bottom line, total cost or net profit, has been our measure. However, we have encountered enough problems along the way to confine our appraisal to one of cautious optimism. We have alluded to some of these problems earlier. We would now like to share them further.

Many of our development and operational problems can be grouped and labeled as environmental. We do not and cannot fully control our hardware or software environment. The trial is only one user of a university computer system and an indirect customer of the software house which produced System 2000. Even though a clinical trial might become the largest single user of this particular computer utility, it would still be a single user.

Production schedules for the Hyperlipidemia Trial are now largely based on University classroom assignments. Updating the Administrative database is referred to as a conversational mode operation. During the summer months or winter holidays, this is a factual statement. Response times to individual entries or inquiries have been typically less than five seconds and often only one or two. However, during mid-quarter or final week, response times are measured in minutes.

The batch mode update to the scientific database requires over 110K of memory and as much as 60 seconds of CPU time. The priority scheme at the University optimizes throughput for student FORTRAN jobs. Obviously, a major update run by our project requires enough resources to reduce run priority considerably. An update run submitted in the morning during final examination week will test the computer system’s mean time between failures.

These response time problems are a major frustration to the trial systems staff and management. But the bottom line is our major measure. We pay only for time used. The trial does not have the fixed expense attendant to a dedicated computer system. We should not expect the side benefits of a dedicated system. We have however, relearned another old lesson. Update runs are scheduled according to system loading. Batching data for a scheduled update appears to relieve the perceived pressure to rush individual forms to the terminal. We are experiencing better human editing and checking of data forms and a lower rate of rejection by the computer system on heavy system load days.

The elapsed time problems with batch updates have forced us to give close attention to file backup procedures. We are also looking into the use of update and suspense files as an adjunct to the central database. We should be able to reduce our system resource requirements and vulnerability considerably, by reviving these techniques. They appear to be compatible with our System 2000 database approach, and we do not anticipate major reprogramming costs for implementation.

So far we have discussed our implementation of one alternative approach to construction of an information processing system for a clinical trial. The use of standard computer resources and application of tested systems analysis techniques permitted attainment of trial goals in minimum time with minimal cost. This approach is certainly not new nor is it unique to clinical trials. Also, there is no doubt that another alternative could have been implemented to produce another optimal system balance. Indeed, other clinical trials have taken alternative routes to system implementation and have succeeded according to their measures.

The variety of paths taken to attain a common goal by different clinical trials points to a major side benefit which could be significant to all medical information processing. Clinical trials require systems to handle clinical records for ambulatory care, longitudinal analysis of clinical data, computerized medical record summary systems, and archival storage of clinical data. In fact, they need most of the tools we are developing for health care applications. There is no doubt in the minds of the computer community or the medical community that they are necessary tools. There is however, an undercurrent of doubt or concern over introduction of these tools into an existing system of health care delivery.
We contend that the controlled clinical trial is an ideal environment for development and testing of these new tools. A clinical trial, by definition, is a controlled environment. The patient population is stable, the trial has an endpoint in time, and the necessity for data collection is established. Further, installation of new procedures in a clinical trial would be a non-disruptive, parallel operation. A trial is not in the main stream of day-to-day health care delivery. A trial does, however, require most of the procedures used in the main stream.

Finally, a clinical trial is a research procedure. Both the patient population and the medical personnel are conditioned to accept new practices. Clinical trial medical staff are usually research oriented and more receptive to use of computers. This last may be the most important point. A prudent health care practitioner must be a conservative when his patient is involved. This attitude leads directly to conflict when something new is proposed for patient care. This type of conflict can become irrational and irreconcilable and lead to rejection of the new idea. Unfortunately, the psychological process described is evident throughout the health care industry today.

This potential test-bed facility, the controlled clinical trial, is not a new practice. Drug trials have been run by the pharmaceutical companies for years. The National Heart, Lung and Blood Institute has funded at least six major trials in the last fifteen years. These have required from 1500 to over 10,000 patients for each trial. Anticipated trial duration ranged from three to six years or more. Of more importance to the computer community, the clinical trial is here to stay. Recent legislation requires that all new medical devices, such as pacemakers, undergo a clinical trial before being released to the public.

We seldom find a potential testing ground for computer development which has broad generalized needs, is truly non-disruptive, and is securely financed. Controlled clinical trials meet all these criteria and more. We must stop regarding trials as different or unique experimental procedures and begin to exploit them as a resource.