THE USE OF RELATIONAL DATABASES AS A TOOL FOR CONDUCTING CLINICAL STUDIES

Randolph A. Miller MD, Wishwa N. Kapoor MD, Jacqueline Peterson PA

Division of General Medicine, Department of Medicine,
University of Pittsburgh, Pittsburgh, Pennsylvania

Abstract

Many clinical investigators already have access to powerful university mainframe-based relational database management systems. Many of these programs, with minimal expense in time and money, can be employed to assist the investigator in conducting clinical trials. We describe our experience using one such general relational database system (System 1022) in completing several academic studies.

Introduction

A common problem in academic medicine is finding a workable scheme for organizing and implementing a clinical study. The goal of most clinical studies is to evaluate a group of patients vis-a-vis a defined set of parameters ('protocol') over a specified time interval. This paper describes our experience using System 1022, a general relational database program, for the overall management of data collection and analysis in such studies. We provide the description as a paradigm for others who do not have access to medically oriented computer-based record-keeping systems such as COSTAR (1), HELP (2), TMR (3) or the Regenstreif Medical Record System (4). Many university computer centers have already purchased easy-to-use relational database management programs. Often, these general database programs can be linked to powerful statistical analysis programs such as SPSS or SAS, allowing the completion of clinical studies with great ease. Using such a system, an investigator can design and carry out a clinical trial with only a few man-hours of programming time and a relatively small amount of time devoted to data analysis.

System 1022 was developed by Software House, a software company in Boston, and runs on DEC PDP-10, PDP-20, and VAX hardware (5). System 1022 can be run in either batch mode or interactively.

Our Purpose in Using System 1022 for Clinical Studies

A team consisting of a physician, a physician's assistant, and a medical secretary (or medical student in some instances) have carried out several retrospective and prospective studies at our institution using System 1022 (6-12). Among these, we have conducted a large study of patients with syncope in order to evaluate methods of diagnosis and to gain an understanding of prognosis. It was necessary to collect prospectively a large body of clinical information because the spectrum of diseases that cause syncope was known to be wide, and because the natural history of patients with syncope could be determined only through long-term follow-up. Data modules were created encompassing demographic information; description of any previous syncopal events; description of the particular events surrounding the current syncopal episode; description of symptoms of the syncopal episode itself; notation of the presence or absence of associated cardiovascular, neurological, or other diseases; findings on physical examination; results of laboratory tests; and data to be collected during follow-up visits. Since certain laboratory tests produce complex results, it was necessary to build separate modules to categorize their results: the 12-lead electrocardiogram, extended Holter or electrocardiographic monitoring, the electroencephalogram, computerized axial tomography (CAT scan) of the head, the echocardiogram, cardiac catheterization, and others fell into this class. By careful evaluation of a large collection of patients using these criteria, it was hoped that new or previously unrecognized associations between syncope and other diseases might become evident. It was further hoped that we could derive clinical prognostic indices useful for future patients presenting with syncope. Examples from this study will be used in describing the utility of System 1022 for data management by a clinical team.
Structuring Data Using System 1022

Well-designed database management systems allow their users to create and manipulate information-containing structures in an efficient and modular fashion. The relational database model offers significant advantages over currently available alternative database formulations (13, 14, 15). Before database systems were developed, it was necessary to write tedious application-dependent programs (commonly in BASIC, COBOL, or FORTRAN) for directly obtaining and manipulating data. Such programs were difficult to modify because changes in data structure required changes in programs. By contrast, data are usually program-independent in a database management system. In 1022, as in most database systems, a data definition language exists with which the user simply declares the structure of records within each dataset. Via the record declarations, the database system gains an 'understanding' of the structure of information being processed. Ullman states (15):

A major role of the database management system is to allow the user to deal with the data in abstract terms, rather than as the computer stores the data. In this sense, the DBMS acts as an interpreter for a (very) high-level language ... ideally allowing the user to specify what must be done, with little or no attention on the user's part to the detailed algorithms or data representation used by the system.

Figure 1 shows the 1022 data definition language (DDL) declarations for the demographic data module of our study. In 1022, each attribute (component of a record) is given a name and an optional abbreviation, is classified according to type (integer, real, text, or date), and is given a maximum field length. Furthermore, each attribute field within a record can be 'keyed' to expedite searching, or left unkeyed to avoid generation of space-consuming (but time-saving) key tables. Figure 2 shows the module describing previous syncopal events.

Data Entry and Editing Using System 1022

Once the data description language modules are loaded into the database system, little if any programming is necessary to conduct the clinical trial. It is a simple matter to OPEN a specific dataset and issue an ADD command; 1022 will then create a new (blank) data record and sequentially prompt the user for attributes by name until the fields of the record are completed. We have found it convenient to collect raw data (during the clinical study) using a data-gathering form created on a word processor. This form corresponds in exact variable order to the data definition files, so that a medical secretary can enter data from the form using the ADD command. Figure 3 shows the data gathering form that corresponds to the data description language file of Figure 2. It was easy to verify the accuracy of entered data by printing the stored data in a format akin to our data-gathering form and then manually comparing the two. When it was necessary to make corrections, the CHANGE and DELETE functions of 1022 allowed straightforward alteration of attribute values within single records or groups of records.

Data Retrieval and Analysis Using System 1022

In conjunction with the simplicity of data declaration illustrated above, an important and practical aspect of relational database systems is the powerful data management commands they provide. In 1022, the FIND and SEARCH commands allow one to select subsets of records from a given dataset based on user-specified criteria. For example, in the demographics dataset in Figure 1, one could locate all persons age 65 or greater by issuing the commands, FIND ALL (to encompass the entire dataset initially) followed by FIND Age_at_Entry GE 65. (Note: GE means greater than or equal to). The MAP function allows one to form a bridge between separate datasets based on keyed attributes which they share in common. Because of this feature, a 1022 user can initially create functionally independent data modules which can later be integrated or subdivided using database programs. After issuing (the above-mentioned) commands to identify geriatric patients in the demographics module, one could examine the history of previous syncopal events in this subpopulation by issuing the command, MAP to PREVIOUS-SYNCOPE via pid pid. This command causes 1022 to jump from the currently selected subset (of the demographics module) to a subset of the previous-syncope module (Figure 2) via patient i.d. numbers that match those of the selected (geriatric) subset in the original demographics module. Issuing the command, PRINT MEAN Number_Previous_Events would then display the mean number of previous syncopal episodes in patients age 65 years or greater. In addition to the MEAN function, 1022 provides several useful numerical functions including TOTAL, VALUES (which gives the frequency distribution for all discrete values of a variable (attribute) occurring in the currently selected data subset), and STDEV, which gives the standard deviation for all values of a variable occurring in the data subset. Another useful 1022 command is DUMP, which allows the user to restructure a dataset. It can be used in two ways. First, if halfway through a study, one discovers that a crucial attribute has been omitted from the record definitions in a dataset, all is not lost. One can DUMP the records from the old dataset to a new dataset which contains space for the missing attribute field. Second, the DUMP command can be used to create files readable by other programs (e.g. BASIC or PASCAL). A user can design more complicated functions than those described above by using the Data Programming Language feature of 1022 (vide infra).
Thus one can obtain insight into diagnosis and prognosis by using a database system to analyze and summarize data from large numbers of patients. The following abstract, presented at the National meeting of the American Federation for Clinical Research in 1982 (8), is an example of the results obtained in our prospective syncope study:

We evaluated and followed prospectively 150 patients with syncope in order to one, determine how often a cause of the syncope could be established; two, to define the value of the various diagnostic modalities in the workup of these patients; and three, to determine the natural history of this symptom. Eighty-five patients were women and sixty-five were men with a mean age of the total group being 55 years. Ninety-seven patients were initially evaluated as inpatients and 53 as outpatients. Sixty-seven percent of the total group had more than one syncopal episode. Using specific criteria, a definite cause of syncope could be assigned in only 62 patients. In the remaining patients, a cause could not be assigned in spite of extensive testing including cardiac monitoring, EEG, head CAT scan, cardiac catheterization, cerebral angiography, electrophysiologic studies and glucose tolerance testing. In patients in whom a diagnosis was assigned, 60% were made on initial evaluation (including history, physical examination, and 12 lead electrocardiogram). Ambulatory or intensive cardiac monitoring revealed the diagnosis in 35% and other tests in 5%. Cardiovascular causes were responsible in 56% of those patients in which a secure cause of syncope could be established. During a follow-up period of 5-9 months, 8 patients died and 19 had recurrent syncope. Of 7 patients with renal insufficiency, 3 died. We conclude that the etiology of syncope frequently remains undetermined and when a specific etiology can be assigned, it is most often made by initial evaluation or prolonged cardiac monitoring. Follow-up revealed a significant mortality, particularly in those patients having co-existent renal insufficiency.

The 1022 Data Programming Language and Host Language Interfaces

Although it is possible to use 1022 without writing any programs, System 1022, like many of the better database systems, provides a Data Programming Language (DPL) for special applications. The 1022 DPL is an easy-to-learn, BASIC-like language. With it, we were able to create derived fields within records, such as current age (based on date of birth), time since last follow-up visit, and time from first visit to death. DPL programs were also used to generate reminders about patients missing critical information during the study, and to indicate when patients were overdue for follow-up visits.

System 1022 offers the further flexibility of host language interfaces to the FORTRAN and COBOL languages, as well as to statistical packages such as SPSS and a specialized National Cancer Institute statistical program. As an example, we used logistic regression analysis in developing a clinical index to predict the value of prolonged electrocardiographic monitoring in patients with syncope. The following abstract describing that particular study was presented at the Midwest Regional Meeting of the American Federation for Clinical Research in 1982 (10):

Our previous prospective study of the diagnostic evaluation and follow-up of patients with syncope has revealed that after initial history and physical exam, prolonged electrocardiographic monitoring is the most useful diagnostic test in assigning a cause of syncope. The purpose of the present study was to define those clinical predictors which help distinguish patients in whom electrocardiographic monitoring was diagnostic of the etiology of syncope. Of 294 patients evaluated prospectively, a diagnosis was established using specific criteria in 103 patients. History, physical exam, and initial EKG established the diagnosis in 60 patients. Electrocardiographic monitoring established a diagnosis in 28 patients and other tests in 15 patients. A diagnosis was not established in 101 patients. We compared the clinical characteristics of 28 patients in whom electrocardiographic monitoring was diagnostic (Group I) to 101 patients in whom a diagnosis was not established (Group II). Univariate analysis revealed that Group I patients were more frequently greater than 55 years of age (62% vs 54%, p = 0.008), had a higher incidence of history of ventricular arrhythmias (28% vs 4%, p = 0.0001), and had more often a history of congestive heart failure (21% vs 8%, p < 0.05). Group I patients more often had an initial systolic blood pressure of less than 110 mmHg (20% vs 5%, p = 0.02) and more often had an abnormal but nondiagnostic EKG (89% vs 64%, p = 0.01). The data were subjected to logistic regression analysis and an effective rule for maximizing the diagnostic value of cardiac monitoring was derived. The rule consisted of monitoring all patients who were 55 years old or older or who had a history of ventricular arrhythmias. By this rule 86% of patients with diagnostic monitoring would have been selected for testing.

Conclusion

In summary, using a relational database system (1022), we have found it possible to conduct several retrospective and prospective studies involving the collection of hundreds of data items on each of several hundred patients. All aspects of the studies, both clinical and data management, were carried out by a team consisting of only a physician, a physician's assistant, and a medical secretary.

We recommend this tool for conducting clinical studies to all interested parties who have access to relational database systems at university or other computing facilities. The method is becoming increasingly popular among faculty members at the University of Pittsburgh School of Medicine.

Figure 1: Data Definition Language declaring Patient Demographics

| ATTRIBUTE Protocol ABBREV prot INTEGER LENGTH 3 KEYED |
| ATTRIBUTE Patient_ID ABBREV id TEXT LENGTH 9 KEYED |
| ATTRIBUTE Date_of_Entry ABBREV entry DATE LENGTH 8 |
| ATTRIBUTE Last_Name ABBREV lasta TEXT LENGTH 25 KEYED |
| ATTRIBUTE First_Name ABBREV firsta TEXT LENGTH 10 |
| ATTRIBUTE Date_of_Birth ABBREV birth DATE LENGTH 8 |
| ATTRIBUTE Age_at_Entry ABBREV age INTEGER LENGTH 3 KEYED |
| ATTRIBUTE Patient_Gender ABBREV sex TEXT LENGTH 1 KEYED |
| ATTRIBUTE Race ABBREV race TEXT LENGTH 35 |
| ATTRIBUTE Home_Age ABBREV add TEXT LENGTH 35 |
| ATTRIBUTE Home_Age_age ABBREV age INTEGER LENGTH 10 |
Figure 2: Data Definition Language for Previous Syncopal Episodes

```
ATTRIBUTE Patient_ID ABBREV ptid TEXT LENGTH 8 KEYED
ATTRIBUTE Number_of_Previous_Events ABBREV npe INTEGER LENGTH 4 KEYED
ATTRIBUTE Age_At_First_ADMISSION ABBREV ad INTEGER LENGTH 2
ATTRIBUTE Prior_ADMISSION_sync ABBREV pas INTEGER LENGTH 2
ATTRIBUTE Prior_Events_Same ABBREV prsam INTEGER LENGTH 1
ATTRIBUTE Dx_Prior_Sync ABBREV prdx INTEGER LENGTH 3
ATTRIBUTE Dx_Prior_Sync_2 ABBREV prdx2 INTEGER LENGTH 3
ATTRIBUTE Previous_Situation ABBREV prsit INTEGER LENGTH 2
```

Figure 3: Data Gathering Form corresponding to Figure 2

**PREVIOUS SYNCOPAL EVENTS**

- # Previous Events (npe)
  (Not including index event)
- # Events in Last Year (nly)
  (Total including index event)
- Age at First Event (ad)
- # Previous Admissions for Syncope (pas)
- Characteristics of Prior Events (prsam)
  (0 - No prior event; 1 - same as index; 2 - different diagnosis; 9 - unknown)
- Diagnosis of Prior Syncopal Events (Last diagnosis (text) and code (prdx#), only codes to be input)
  prdx1 text-dx1
  prdx2 text-dx2
- Previous Events Situational (prsit) (Refer to situational codes in previous section)

References


