DATABASE MANAGEMENT FOR LARGE AND COMPLEX SURVEYS

Margo Brinkley, Lisa Morrissey, Robert Thornton, Research Triangle Institute

A. INTRODUCTION

This paper discusses database management for large complex surveys. This topic is general; however, since we are limited to a relatively short presentation and paper, we have decided to discuss the topic emphasizing a large Research Triangle Institute (RTI) project, the National Medical Care Expenditure Survey (NMCES). This project is sponsored by the National Center for Health Services Research (NCHSR) and the National Center for Health Statistics (NCHS). It is an ongoing survey and is producing a massive and complex database. We feel that this paper would be more meaningful if it addresses the subject of database management for large complex surveys through specific references to NMCES, rather than from a more general perspective relating to an undefined data collection or data management activity.

At the present time, NMCES is basically concerned with data collection rather than database development. Because of this, the ideas we have for managing and supporting this database for analytical efforts have not been implemented. Thus, they are to a degree our opinion, based on our experience with a number of large surveys, of what must be provided if an analyst is to effectively use the data being collected.

We would like to emphasize one fundamental point in this introductory section before we describe NMCES as briefly as possible in terms of background, data collection activities, and the basic software that has been used to support the project and before we discuss our approach to database management. The database we are to develop is a research oriented database as opposed to what might be considered a commercially oriented database. Personnel files, payroll files, parts inventories, social security files, driver license files, and motor vehicle registration files are examples of large databases which can be handled by operations systems, i.e., whose information needs can be easily anticipated at system design time. Thus, it is unlikely that major modifications will be necessary in order to continue meeting system objectives for these types of needs. A research database, on the other hand, routinely requires adhoc transactions which are not required in the normal commercial environment. For example, imputation, computation of sampling weights, and composite variables (often appended to the database), and a considerable amount of editing to correct data or eliminate unacceptable data are often necessary. Since a research oriented database is not individual specific in the sense that the objectives are not to obtain conclusions concerning a given individual, imputation can be a powerful tool in handling missing data. Commercial enterprises, however, rarely face the important problem of missing data and its effect upon analysis efforts.

Since this paper is being presented to a meeting of the research oriented American Statistical Association, many of the above points concerning research versus business oriented databases are already known. However, we have encountered many people whose exposure to large complex surveys and dynamic research databases is rare and who fail to understand that statistical analysis at one or more stages often becomes a "fishing expedition". This requires the database management staff to routinely provide analysts with different subsets of the database often requiring computation of new variables or samples of the basic files.

B. NMCES BACKGROUND

A considerable amount of medical data, including information on medical costs and expenditures, exists through a multitude of data collection activities. However, these data are fragmented and scattered among numerous private concerns such as insurance companies, hospitals, and government agencies. For the most part they are unusable because of their fragmentation and because in many instances they were collected to answer only a very specific set of questions. From a technical standpoint it would be virtually impossible to integrate existing medical data into a comprehensive database which could support analysis designed to answer general questions related to government policy in areas such as national health insurance or which could provide information about the accurate cost of medical care in the United States. Thus, one of the key purposes of the NMCES project is to develop a comprehensive database which can provide detailed data on medical expenditures for both the insured and uninsured portions of our country's population.

To develop this database a large survey was designed and implemented which consists of three basic components: a household survey, a medical provider survey, and a survey of providers of insurance coverage which involves both insurance firms and employers who subscribe to group plans. The largest survey is the household survey which consists of over 13,000 households (referred to hereafter in this paper as reporting units) and approximately 39,000 individual participants who are associated with the reporting units. The household survey is a panel survey consisting of six different data collection contacts with a reporting unit. The first two interviews were personal interviews conducted by field interviewers, the next two interviews were by telephone, and the fifth interview was again a personal interview conducted in the field. Following the last interview, a telephone interview to clear up inconsistencies, etc., will be conducted.

During the interviewing process, critical data were extracted and printed as a summary report. This report was sent back to each reporting unit and to the interviewer prior to an upcoming interview. This technique was used to obtain information which was missing or inconsistent on previous interviews (for example, the amount of a medical expenditure paid by an insurance company and the amount which had to be paid by a family). To supplement the information collected from the household, the other two surveys previously mentioned will be used to collect information from hospitals, clinics, physicians, insurance carriers, and employees. These surveys will provide a basis for assessing the accuracy of the data collected in the households as well as additional information which in general is not known by most survey participants. In addition to the data collection activities, the project has subprojects for coding of medical conditions in a form that makes them more useful for analytical purposes and for developing a data file of episodes which involve combinations of specific medical conditions, hospital visits, etc.

C. DATA COLLECTION PROCESS FOR NMCES

While the purpose of this paper is to discuss data management of the database being developed and provide some general comments on management of similar databases, it is felt that a few brief comments on the data collection process

should be made. This process consists of both field interviews and telephone interviews in which the data are recorded on hard copy data collection forms. (There are numerous forms involved including the primary forms for the different surveys as well as various supplements and continuation pages required for unusual situations.) The data are mailed to RTI, logged in, reviewed in a manual edit, converted to machine readable form using a programmable key-to-disk system which provides some data editing capabilities, and is later machine edited using an IBM 370/165 computer system and then integrated into the database. From the database, critical data are extracted to become part of the summary database which is then used in subsequent interviews. In parallel with the data collection activities, other software is used to monitor field operations. This software is referred to as the Control System and was featured in one of the papers presented earlier in this session. At this time the majority of the effort that has gone into the NMCES project has been related to data collection, control of field operations, and routine edits of data which are made before the data are integrated into a database.

D. NMCES VOLUME AND BASIC FILE STRUCTURE

Two factors that must be taken into consideration when deciding the most efficient method of database management are the volume of data being collected and the natural file structure that results from the data collection procedures and forms being utilized. From the standpoint of volume, the initial NMCES database will be somewhere between 1.5 and 2 billion characters of data. In other words, depending on blocksize and other technical factors and assuming a density of 1600 bytes per inch, the data will span 30 to 40 reels of tape (2400 feet per reel).

The design of this database will not only be affected by its impressive size but also by the dynamic nature which must be built into the file structure. The database will be expanded with the insertion of imputation flags as well as with the addition of composite variables. Whole subfiles containing data from other related surveys and subfiles of statistical sampling weights will be integrated into the initial structure. It is expected that with ongoing analysis the database will continue to grow and require changes in the basic file definitions.

In working with NMCES data for the purposes of initial editing and report generation, the data were subdivided into segments containing various types of information. Each segment type defined a fixed length record preceded by a header of pertinent identification information. Examples of various segment types are hospital visits, conditions, medical provider visits and health insurance segments. A database system based on the merging of all segments of similar type would result in an aggregate of over 200 different files. Even though the linkage exists to integrate these records, the files basically lose their identity from the standpoint of reporting unit, participant, etc. In other words, the database becomes a collection of files associated with various types of medical and cost information where the particular household or individual is not the key file record identifier.

Inherent in the purpose of the survey is a necessity to link data according to individual participants. This ability is not easily achieved in a variable and complex database. The variability in this database is considerable from reporting unit to reporting unit and from participant to participant within reporting units. Every member of a reporting unit was interviewed; this varied from one individual in a reporting unit to as many as 23 individuals in the maximum case. In addition, all of certain types of data were collected on each member of a reporting unit. For example, a given member of a reporting unit could have no visits to a provider of medical services for a given survey wave and another member could have 50 to 100 visits. Thus, if thought of from the standpoint of the reporting unit and participants within, the amount of data collected for a given reporting unit or participant is highly variable. To further complicate the database, each of the household survey waves was different, making the total household survey a collection of six different surveys linked through common identifiers.

The natural or basic file structure which has developed from this survey is not easily handled from the data processing and computer aspects of the project. The basic file linkage is to an address resulting from a cruising and listing of primary sampling units and the subsequent sample that was selected. However, each of these lines in the sample, which is referred to a case number, can in an interview result in multiple reporting units. From a data processing point of view this means that multiple reporting units can link back to the same line number in the original sample. The next basic linkage is the particular survey, or wave, that was conducted. Within that, linkages are reporting unit, participant within reporting unit, book number (since a number of the interviews result in multiple questionnaires being filled out by the interviewer), and segment within book. Thus, if thought of as a hierarchical structure or as a totally integrated database, at least 8 to 10 key indexes are required to subset the database in a way that can be useful to an analyst.

To emphasize the level of detail necessary in selecting a file structure, we want to define one example of additional complexity referred to as split reporting units. For example, a reporting unit initially might consist of a husband and wife, who separate at a later point in the survey. A new reporting unit is then defined and the husband and wife treated as two different reporting units. At an even later interview the husband and wife may have reunited and are again considered as a single reporting unit whose identification is that of the particular address of the reporting unit being interviewed. In terms of database complexity, this means that participants do not necessarily remain in the same reporting unit throughout the survey. Thus, analysts must be prepared to deal at a reporting unit level with a varying number of participants who may leave and reenter a reporting unit over time.

From a data collection point of view, to attempt this level of detail may initially appear to be excessive or unnecessary. However, from a substantive standpoint, split reporting units are very important in that they, in many instances, increase the cost of medical expenditures (i.e., additional insurance policies, which may be subscribed to and then dissolved). From the standpoint of field operations, we were required to maintain all combinations of splits and to notify each interviewer of all possible members of the original reporting unit. For data collection and control of survey operations, splits were an extremely complex problem. From a data processing perspective, analysis based on participant often required retrieval of data from more than one reporting unit, thus detracting from the efficiency of the basic hierarchical design.

Another expensive complexity results from the neccessity to update missing or incorrect data in the summary files. These files, due to the field data collection procedures, become expanded subsets of the database. For example, an individual may recall originally reporting multiple visits to one doctor when in reality more than one doctor was involved. Since the purpose of the summary was to expand and capture all of the data associated with critical data items, a visit record which in the initial database could represent three or four visits to the same doctor, in the summary was expanded to an individual line item for each visit. In such a case as the above

mentioned, the particular summary line item was corrected to reflect the proper information. Since the original record in the initial database was expanded in the summary to represent a number of different records, and some of these were changed during the interview, there is no linkage back to the database for the corrected or new records. Thus, from a data processing perspective the summary data file becomes an expanded subset of the original data file which cannot be linked on a one-to-one record basis to the initial database.

It is apparent that to support analysis efficiently and expediently the data must be accessible through both types of file structures defined above. For example, analysis covering cost per visit to a medical provider would best be done through retrieval from a collection of fixed length segment files. On the other hand, cost per person or episode cost analysis would be more efficient if based on the retrieval from a hierarchical file structure with the participant identification number as one of the key indices. Also, many types of survey methodologies could call for speedy retrieval of all data for a particular household throughout the six waves, as well as all summary data for that household. To summarize, flexibility in modes of accessing records and future integration of other survey data files must be taken into consideration in deciding upon management of this database.

E. DATABASE MANAGEMENT

1. Introduction

Since the future content and structure of the NMCES database (additional surveys may be undertaken to fill gaps in the initial effort) and other similar databases is unknown at this point in its development, a general answer to database management is difficult to achieve. We will, however, discuss two approaches which appear to have promise. In addition, we will describe our efforts at a preliminary database using the MARK IV File Management System.

This preliminary database was set up as a hierarchically structured file of variable length records. It provided analysts with data for initial tabulations and allowed analysts the opportunity to work with large complex files. Although MARK IV was adequate for production of the summary database and the summary reports, we formed the opinion in producing this initial release database that MARK IV could not be the central package under which we would structure the entire database. Use of MARK IV in this preliminary analysis effort presented several problems. The lack of basic statistical capabilities required that either special purpose subroutines be added to MARK IV or extraction of package compatible subfiles from the complex database be performed. Neither solution can be considered a cost-effective approach to providing analysis of the data. In addition, a good understanding of MARK IV and the data structure was required to perform various tasks. For example, due to the necessarily complex file structure, subfile extraction and basic editing procedures such as elimination of exact duplicate segments required a sophisticated knowledge of MARK IV. Another technical problem encountered in MARK IV usage was the maximum allowable logical record length of 32,767 bytes. In many instances, the data for one respondent exceeded that limit.

Also, in the process of developing these preliminary release files, we concluded that at least initially no database management system (including System 2000, IDMS, ADABAS, MOD 204, etc.) could serve as the central system for management of the complete NMCES database. The reasons for this conclusion include:

1. The cost (estimated to be $500,000 or more) to structure and load this large collection of data into a database system.

2. The cost associated with selecting the appropriate Database Management System (a task complicated by the divergence of opinion on database systems).

3. The possible need for reorganization and reloading of the database at a later date.

4. The necessity for having a database administrator and associated staff who are at least as costly as the technical staff proposed by RTI.

5. The many unknown applications for the database (a factor in the selection process) and the still unclear content of the database itself.

While all of the database systems have impressive features, use new concepts and techniques for dealing with data and data files, and allow non-programmers to access the data; we do not see at this stage any way to cost effectively structure the complete NMCES database (which we feel is somewhat typical of data collected in a complex survey) so that it is accessible to analysts without strong backgrounds in computing. Because of this we have made some planning decisions which are certainly different from the more structured database approach generally employed in a commercial environment.

2. Database Support Staff

We intend to provide a database support team (currently estimated to be four to five people) which is knowledgeable about the data and the collection procedures and which has been exposed to the types of errors and inconsistencies which usually appear in data files. The staff will have backgrounds in mathematics, statistics, and computer science and will be capable of providing an analyst with information from the database at essentially any level of detail desired. By this we mean the staff will have a complete understanding of the computer system, the diagnostic codes generated and the ability to work at the Assembler Language level if necessary; a basic understanding of mathematics and statistics that will allow them to easily interface with experts in these areas; and a working knowledge of the database.

3. Database Structure

In general, we view the database as having three major divisions. The first consists of data that will be removed from the initial database because of poor quality (either the data collection procedures were incorrect or the quantity of missing data was too great). Such data will in a sense be archived and unless supplemented by subsequent data collection efforts will probably never be used. The second database will consist of data which are considered of sufficient quality to provide significant analytical results. These data, from a computer programming perspective, may be the data which eventually require the most processing. Adequate data will exist for imputation and will be clean enough to provide analysts with results that can be considered accurate. The third database is what can be thought of as the critical database, i.e., the summary of data updated during each interview, other key information from the household survey, data from the medical provider and insurance surveys, as well as composite variables that have been computed and added to the data. This database will at least initially be the database that is most heavily used and will be maintained at a minimum of two different computer facilities (IBM based). It is possible,

since different users of the database will use different computer facilities, that different versions of the database (possibly even complete) will be implemented at other facilities. We plan to maintain the archival version of the database on magnetic tape since it probably will be accessed infrequently. The second portion of the database will be maintained on both magnetic tape and direct access devices. The third or critical portion of the database will probably be maintained on on-line direct access devices. At the present time, we intend to access the database through special purpose software developed in FORTRAN using an extensive subroutine library developed at RTI over the past few years and written primarily in IBM Assembler. The analysis of data will be primarily package oriented using existing statistical packages such as SAS, SPSS, BMD, etc., and also simulation and modeling packages.

F. CONCLUSIONS

We would like to conclude by reemphasizing two points. The first is the basic difference between a research oriented database which results from a complex large-scale survey and a large commercial database (much larger even than the NMCES database) where both data records and input and output transactions can be explicitly defined prior to creation of the database and its associated software support system. The latter situation is much easier to handle from a programming perspective regardless of whether in-house software is developed to maintain the database or it is maintained through a commercially available database management system. Once the database has been established and the associated software implemented, day-to-day maintenance including updates and report generation, etc., can normally be performed by clerical staff who have no background in computing. This is not the case with a research oriented database. In the case of NMCES, the most used part of the data, that which was referred to earlier as the critical database and would be maintained on-line, will change at least once a week and possibly even more often. Some of these changes will be simple, and others will require that the complete database be accessed to develop new records in order to group data in a way that is usable to an analyst. Because of a dynamic nature of the critical database we feel that professional programmers are essential as the database maintenance staff and we are somewhat skeptical of the ability of a commercially available database management system to function in this environment. We do plan, however, to experiment with some of the commercially available systems with that part of the NMCES database which will be maintained on on-line storage. Since the data will be stored on-line and must be frequently accessed, a database management system could be the most cost-effective way to process this part of a database (particularly when the database has become more static). In addition, the data processing support for NMCES will be a contract which will be renewed periodically. The use of a commercially available software system provides a degree of documentation that does not normally exist when special purpose software systems are used. Thus, it would be easier for the government to change contractors if the database was accessible through a popular commercial package.

The second point is concerned with management of a large complex database. On many projects, funds are only allocated for data collection, editing, basic tabulations, and documentation of the database. At that point the data is considered "clean" and can be provided to analysts in the form of release files. However, there are often many complex tasks that should be performed that are not initially obvious and that require the skills of professional programmers in conjunction with an analyst. By this we mean tasks such as imputation, composite variable computation, recomputation of sampling weights, development of subsets which require complex programming to access data and create records which do not exist in the basic structure. In general, we mean providing an analyst with anything required from the database and assuming cost implications are made known. We do not feel that the basic objectives of NMCES, and other surveys which generate similar large complex databases, can be achieved unless funds are available to establish a database management support team made up of professional programmers who in addition to having technical backgrounds have extensive knowledge of the data. Only time will tell if the database support staff can simplify the database management function to the point that nonprofessionals can perform the majority of the tasks that are required.