Published in final edited form as:
Nurs Res. 1990 ; 39(3): 188–189.

Data Management Issues In Longitudinal Research

JoAnne M. Youngblut, PhD, RN [Assistant professor],

Frances Payne Bolton School of Nursing, Case Western Reserve University, Cleveland, Ohio. At the time of the study, the author was a doctoral student at the University of Michigan, School of Nursing, Ann Arbor, MI

Carol J. Loveland-Cherry, PhD, RN [Assistant professor], and

The University of Michigan, Ann Arbor, MI

Mary Horan, PhD, RN [Professor and director]

Kirkhof School of Nursing, Grand Valley State University, Allendale, MI

Data management begins with the design of the study and includes such things as scheduling data collection, selecting a statistical package, deciding on the structure of the raw data, coding questionnaires, and entering and cleaning data preparatory to initiation of analysis. Although much has been written about various aspects of the research process, including design, instrumentation, sampling, attrition, and data analysis, little can be found about the practical aspects of data management, especially for longitudinal studies (Barhyte & Bacon, 1985).

Generally, researchers learn about data management issues and possible solutions by trial and error, by word of mouth from other researchers, or by problem solving using logic during the course of their study. The authors' premature infant project is illustrative of major data management issues. For this study, 120 families were recruited from two Level III neonatal intensive care units (NICUs). Data were collected in five waves over a period of 15 months from both parents in the home and from the baby's hospital record and outpatient clinic visits. The focus of the study is to investigate how parents' reactions to the premature birth of an infant affect family functioning and how their reactions and family functioning affect the infant's developmental progress.

Data Integrity

Maintaining the quality of study data is always of concern, but this is especially true in longitudinal studies. If data are to be collected on multiple instruments, from multiple respondents, at more than one time, and at more than one site, measures must be instituted to maintain consistency at the point of data collection and in subsequent coding and data entry. Consequently, several strategies were employed to enhance the quality of the data obtained.

First, interview forms were color coded by respondent and by interview time. Color coding helped the interviewer to sort forms for mothers and fathers while in the home and reminded her on which form each respondent's answers were to be recorded. Second, forms for each visit were collated ahead of time and placed in data collection packets to be distributed to interviewers at each site. Both of these measures made the interviewer's job easier and insured that a complete set of the correct forms were taken to each visit.

Finally, several record-keeping systems were implemented. Each site developed a system to keep track of when each family was to be seen so that family visits were performed according to study protocol. One of the sites used a tickler file similar to that used by community health nurses. Identifying data for each family were recorded on an index card which was then placed in the file under the month during which the next visit would be

done. The other site used a calendar to track visits. When a family was recruited to the study, all five of the family's visits were marked on the calendar during the appropriate months.

Flow sheets also were used for the data packets. These were marked when a packet was received, when it had been taken to the data entry firm, and when it was returned. Using the flow sheet made it easier to see the status of data coding and entry for each data collection period in a short amount of time.

Data Structure

Decisions about the best structure for the raw data need to be addressed early in a project. This allows computer column numbers to be printed on all instruments for direct data entry. There are several issues to consider, depending on the design of the study. One involves selecting a statistical package, another concerns defining the case, and a third involves handling the repeated observations.

Selecting a Package

If more than one program is available, the researcher should select a package that will meet most of the data analysis needs of the study. Doing so may save both time and money. This involves investigating differences among the packages and evaluating those differences. While almost all of the packages contain most of the statistical tests necessary for data analysis, some procedures are more complete or easier to use in one package and some packages have more options for a particular procedure or produce output that is easier to interpret. If a specific statistical procedure from another package is necessary, the raw data for the variables to be used can be written from the existing system file for one package (the file that the computer makes from the raw data and the instructions for reading the data) into a new data file and read into the other statistical package.

Familiarity of the project staff with the available packages is also important, as is the relative cost of performing specific analyses with each package. Finally, packages should be evaluated regarding the procedures necessary to perform data transformations. For some studies, the ease of creating new variables and recoding old ones may be the deciding factor.

Defining a Case

It is important to understand how statistical packages handle specific analysis procedures. For example, in the premature infant project, comparisons between mothers and fathers were planned using the paired *t* test. Some statistical packages will do paired comparisons across cases, while others expect paired comparisons to be within a case. Since the latter was true of the package selected for the study, the case was defined as the family rather than the individual.

Another reason for indexing the data by the family was the nature of the family and baby data. Data about family structure were collected from both parents together, rather than from each separately. In addition, there were data about the baby's hospital course, developmental progress, and the outpatient clinic visits that "belonged" to both parents. Indexing the data by the respondent would have required entering the family and baby data twice, at a significant increase in the cost of data entry.

While indexing the data by the family avoids the problems cited, it has its own problems. Defining the family as the case makes doing analysis with all parents, both mothers and fathers combined, more difficult. Some statistical packages have a routine to handle this while others do not. If the package chosen does not provide a solution, there are two ways to

remedy this problem. When the variables for mothers and fathers are exactly the same and the data for those variables are in exactly the same column(s) on different lines in the raw data file, then writing a second setup file (the file containing the instructions for the program to read the data) defining the case as the individual is possible. If the data are in different columns or if some of the data will be the same for both parents (such as the family and baby data in the current study), two setup files (one to read the mothers' data and one to read the fathers' data) are necessary. The two resulting system files can be merged to produce one system file with all parents as the reference group and with the individual as the case.

Another way to deal with this problem is to modify the system file. Using this approach, one would subset the existing system file, first excluding the mothers' variables and then excluding the fathers' variables. Since mothers' and fathers' variables have different names in the original file, variables that mothers and fathers have in common must be renamed so that the same name is now used in both system files. In addition, a new variable indicating sex of respondent must be created in each file. The two resulting system files can then be merged to produce a system file of all parents.

Handling Repeated Measures

While data analysis can be accomplished on either a mainframe or a microcomputer, the mainframe was most practical for the authors' study because of the large number of variables in each wave. The following suggestions are made with the mainframe user in mind. Therefore, some aspects of the discussion may not be considered relevant for researchers who are using a microcomputer for data analysis.

There are three basic options to consider in handling repeated measures data. First, all the data for all data-collection points can be entered as a single case in one file. However, care must be taken if data entry and analysis are to begin before all data are collected. For example, since subject recruitment in the authors' project occurred over a two-year period, data collection for Time 1 overlapped with data collection for the other four time periods. In this case, the number of data records for each subject varied across subjects at any given time. Adjustments in the setup file or the raw data file must be made so that the computer correctly reads the data. In addition, since the cost of data analysis on a mainframe is often dependent on the number of variables in the system file to be analyzed, using such large files can appreciably increase the total cost of data analysis.

A second option is to enter each wave of data for each family as a case in the same file. In order to do this, the variables collected in each wave must be the same, although having a few more or a few less variables in a certain wave would not affect the result. This method is especially valuable when the number of observations will differ for different cases. For example, if the study design is to interview families every week during the first month of their infant's NICU stay, some families will be interviewed once, while others may be interviewed four times depending on their infant's length of stay. The disadvantage is that in order to do comparisons across time, the system file must be subsetted and the variables renamed as described above.

A final option is to put each wave of data in a separate file. This approach was most satisfactory for the authors' premature infant project because the number and types of variables collected at each time period were drastically different, making a single file impractical. Also, comparisons of variables across time were planned. Using a separate file for each wave makes such analysis easier. System files created for each wave of data can be merged as needed without subsetting the data and renaming variables. If a single file had been used, procedures similar to those described above for combining data on mothers and fathers would then have been necessary.

Summary

Much of the existing literature about the research process has neglected data management. While design, instrumentation, sampling, and analysis are important parts of the process, paying attention to the issues surrounding data management is crucial to the success of the study. Data entry and analysis are facilitated when the details of data structure and management are decided before data collection begins.

Acknowledgments

The project to which this paper refers was funded by a grant from the National Center for Nursing Research, #R01-NR01390, to the second and third authors. The first author was supported in part by a National Research Service Award predoctoral fellowship, #F31-NR06152.

References

Barhyte DY, Bacon LD. Approaches to cleaning data sets: A technical comment. *Nursing Research*. 1985; 34:62–64. [PubMed: 3844165]