Available at
http://pvamu.edu/pages/398/asp

Vol. 1, Issue 2 (December, 2006), pp. 141 – 151
(Previously Vol. 1, No. 2)

Applications and Applied Mathematics (AAM):
An International Journal

Data Management Plans: Stages, Components, and Activities

Abbas S. Tavakoli, College of Nursing, University of South Carolina, Columbia, SC 29208
atavakol@gwm.sc.edu

Kirby Jackson, School of Public Health, University of South Carolina, Columbia, SC 29208
kjackson@gwm.sc.edu

Linda Moneyham, College of Nursing, University of South Carolina, Columbia, SC 29208
ldmoneyh@gwm.sc.edu

Kenneth D. Phillips, College of Nursing, University of South Carolina, Columbia, SC 29208
kdphilli@gwm.sc.edu

Carolyn Murdaugh, College of Nursing, The University of Arizona Health Sciences Center, Tucson, AZ, 85721
cmurdaugh@nursing.arizona.edu

Gene Meding, College of Nursing, University of South Carolina, Columbia, SC 29208
gemeding@gwm.sc.edu

Received May 20, 2006; revised received October 13, 2006; accepted October 16, 2006

ABSTRACT

Data management strategies have become increasingly important as new computer technologies allow for larger and more complex data sets to be analyzed easily. As a consequence, data management has become a specialty requiring specific skills and knowledge. Many new investigators have no formal training in management of data sets. This paper describes common basic strategies critical to the management of data as applied to a data set from a longitudinal study. The stages of data management are identified. Moreover, key components and strategies, at each stage are described.

KEY WORDS: Data management, methodology

1. INTRODUCTION

Many new investigators have little experience in managing data sets beyond that obtained in the process of completing their dissertation research, which frequently involves small samples and cross-sectional data. The management of longitudinal data sets required for intervention research poses a challenge to new investigators. Data management requires special knowledge and skills that are usually obtained through supervised hands-on graduate or postgraduate experience. Without such experience, investigators are left to a trial-and-error approach or dependence on other team members to determine appropriate data management

strategies. Well-trained data management professionals on a research project should be viewed as a tremendous asset. Holding regular team meetings with the data management personnel is a way to assure that all investigators are familiar with the data sets and how they have been created. Team meetings allow for early decisions about problems with data collection and entry. Although information about data management issues is available (e.g., Burns, et al., 2001; Davidson, 1996; Hott et al., 1999; MacMurray, 1990; Polit et al., 2001; Roberts et al., 1997; Youngblut et al., 1990), little is written about the more practical aspects of the process. The purpose of this paper is to describe the stages, components, and strategies of successful data management using a longitudinal data set.

Data management involves preparatory, data organization, and data analysis/dissemination stages. Each stage is equally important to study outcomes. Specific activities are required at each stage to ensure the integrity of the data management process. Each stage is described using examples from a longitudinal data set. To facilitate discussion, background information about the data set is provided.

2. BACKGROUND

The longitudinal data set, referenced in the following sections, was generated in a study which tested a peer-based social support intervention designed for a population of rural women with HIV disease (Moneyham, 1999).The study is referred to as the Rural Women's Health Project (RWHP).The experimental study used a repeated measures design with data collection points at baseline, immediately following completion of the 6 months intervention, and at four months following intervention completion. The 280 study participants were recruited from 10 community-based HIV/AIDS service organizations serving rural areas of the southeastern United States. Study participants were randomly assigned to intervention and control groups. Intervention group participants received a total of 12 face-to-face peer counseling sessions over a period of six months, while the control group received the usual care provided by the agency by which they were recruited. Peer counselors were recruited at each local study site to implement the intervention. The data set was large and complex. In the design intervention group participants were nested by study site and peer counselor. There were multiple intervention points for each experimental group participant.

A comprehensive data management plan was designed to organize data handling processes in order to assure data integrity and security. The goal was to create a practical data management plan that incorporated specific data management activities, guidelines for documenting activities at each stage, and methods to avoid, detect, and resolve potential problems. The stages, components, and strategies of the plan are represented in Figure 1. The components outlined in each stage can be used as a step-by-step checklist to insure that any issues are addressed and properly documented.

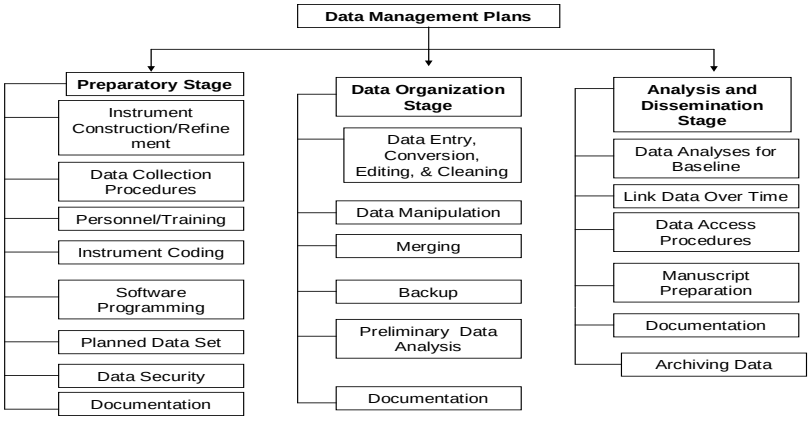

Figure 1. The Data Management Process

3. PREPARATORY STAGE

The preparatory stage takes place during the project startup period and includes instrument construction and refinement, data collection procedures, personnel training, instrument coding, software programming for data entry, planning for data set security procedures, and documentation.

3.1 Instrument Construction/Refinement. The structure and format of the research instrument are critical to data accuracy and completeness. Previously validated surveys should be constructed exactly as published to ensure they are valid. Instruments must not only include items to measure study variables, but also instructions to the data collectors (e.g., research assistant, interviewer, etc.) and study participants. Instructions for data collectors include information to ensure consistency in data collection procedures with accurate and complete responses from respondents. Instructions will vary depending on whether the instrument is self-administered or the data collector reads all items to the participant and records their responses. The latter method was used to collect the referenced longitudinal data set. Figure 2 provides an example of instructions for data collectors. The instructions are written in a bold font so as to be easily identified. Instructions to be read to participants also need to be easily identifiable, and may be written in a different font (e.g., *italic*). The major advantage of written data collection instructions is continuous reinforcement of data collection procedures.

Figure 2. Data Collector Instructions Included in Instrument

Rural Women's Health Project
INTERVIEW #1

Date: _____

Participant ID#: ___ ___ ___ ___

Interviewer ID#: ___ ___

Instructions for Data Collectors

Directions for interviewers are written in "bold" type.

Directions to be read to participants are written in *"italic"* type.

For open-ended questions, write the woman's own words in the space provided.

Each set of questions has a matching card that lists the answer choices. The cards are lettered alphabetically (e.g., "A", "B", etc.) and the directions tell you when each card is to be used.

Use a ballpoint pen to complete the questionnaire; do not use a pencil, as answers may not be readable. Write clearly and neatly.

Sometimes a woman <u>will not answer</u> a question because:

She does not know the answer;
The question does not apply to her situation; or
She does not want to answer the question.

<u>Write one of the following reasons for not answering in the margin next to the question</u>

"doesn't know" or "refused to answer" - (code as "8")
"does not apply" - (code as "9")

If the woman does not understand a question, read the question again slowly and clearly. Allow the woman enough time to answer. <u>Do not try to explain the question or suggest an answer</u>. If after reading the question a second time the woman still does not understand the question, write in the margin <u>"does not understand the question".</u>

The formatting of study instruments facilitates administration and helps to ensure that items are completed in the correct order and responses are accurately recorded. Missing responses are a common problem and occur for a variety of reasons. It is important for investigators to be able to assess the reason for missing responses. For example, participants may not understand a question, they may refuse to answer a question, or the question may not be applicable to the participant. Information about why an item is not answered is useful to evaluate the study instrument. For example, if participants do not understand a question, it may need to be revised. If a large number of participants refuse to answer an item, it may need to be deleted.

<u>3.2 Data Collection Procedures</u>. Written standardized procedures facilitate consistency in data collection across study participants and data collectors. The adage "garbage in, garbage out" illustrates the issues in management of raw data. The quality of data collected is foundational to the validity of study findings. Quality data collection requires a systematic approach and includes: 1) training data collectors; and 2) monitoring completeness and accuracy of raw data.

Several strategies will assure quality data collection in face-to-face interviews, particularly when multiple data collectors are used. Interviewers' training must focus on consistency in how questions are asked and how participants' responses are recorded. Consistency can be

facilitated by "scripting" the interview so that all directions for participants are written and are read to participants. Additionally, interviewers must be instructed not to "interpret" questions for participants. Variation in interpretations of questions introduces the interviewers' biases which can increase variance in responses due to error. When words are used that may be difficult to understand, or have multiple meanings, acceptable words that can be substituted are included on the instrument. Interviewers can be instructed that if a participant cannot answer a question to read the question slowly and clearly a second time and allow the participant adequate time to respond. If after the second reading, the participant cannot answer the question, the interviewer is instructed to record the reason the participant did not answer the question. As previously noted, reasons for not answering a question include: 1) lack of understanding; 2) refusal to answer; or 3) irrelevant or not applicable. Each non-response category is coded for data entry. Reasons for not answering questions are important information for investigators as they provide information about the clarity or appropriateness of a question.

The completeness and accuracy of the recorded responses are also monitored. Data collectors are instructed to check all items to see if every item has a recorded response before completing the interview. A second line of quality assurance is having other research team members review data collection instruments for completeness and accuracy for a second or even a third time. After all the checks of raw data are completed, the data collection instruments are ready for data entry.

3.3 Personnel/Training. When multiple data collectors are used, the interviewer training needs to be designed to establish consistency in data collection procedures. Written procedures are developed and included in a training manual to ensure consistency in approach. During training sessions, the procedures are used as a checklist to guide interviewer practice in role-playing data collection with observation and feedback from the investigators. Early data collection efforts are closely supervised to ensure compliance with procedures and to allow for individual feedback on performance and quality of data collection.

3.4 Instrument Coding. During the preparatory stage, data collection instruments and individual items are assigned a code name for ease of data entry and management. Code names should be meaningful and easy to remember. Coding and naming conventions should be standardized for files, variables, programs, and other entities in a data management system. For example, in the RWHP, individual data files were developed for the three different data collection time points and named in the order in which they were collected: RWHP1, RWHP2, and RWHP3. To assure brevity, all variable names were limited to eight characters or less. A coding manual was written that matched all variable names with variable labels and codes.

When variables are measured across multiple data points using the same measures, the variable names must reflect the different time points of data collection as well as different versions of instruments that might have been used. In Table 1 the variable names are described for variables measured at three different data collection points. As noted in the table, the variable name was slightly modified to reflect the time period when the variable was measured.

Table 1. Examples of Variable Names

Variable Description	Variable Name Baseline	Variable Name Time2	Variable Name Time3
What county do you live?	**County**	**Bcounty**	**Ccounty**
Do you have a paying job?	**Payjob**	**Bpayjob**	**Cpayjob**
Have you been told you have AIDS?	**Taid**	**Baid**	**Caid**
Coping Scale (54 items)	**Cope1 – Cope54**	**Bcope1 – Bcope54**	**Ccope1 – Ccope54**
Social Support Scales (19 items)	**Ss1-Ss19**	**Bss1 – Bss19**	**Css1 – Css19**
Depression Scales (20 items)	**Cesd1-Csed20**	**Bcesd1-Bcsed20**	**Ccesd1-Ccsed20**

Consistent rules must be used for coding variables. For example, when a participant does not provide a response, codes are used to distinguish reasons for the non-response. In the authors' data set, "Refused to Answer" was coded 7, "Does not Know" 8, and "Does not Apply" 9. For two digit variables, the responses were coded as 97, 98, and 99, respectively. To make sure that these types of answers are not coded as actual values, for instance an age of 99, three digits were used to avoid confusion. Such coding decisions were made prior to the start of data collection. Additionally, completed instruments were independently checked by the project coordinator and data manager for completeness and accuracy and codes were assigned for missing data prior to data entry.

3.5 Software Programming. Software programs such as SPSS (SPSS, 1990: SPSS, 1994), DBMS (Conceptual Software, 1999), and SAS (Aster, et al., 1991; Burlew, 1998; Cody, 1999; Delwiche, et al., 1996; DiIorio, 1991; Elliott, 2000; SAS Institute, 1990; SAS Institute, 1995; Spector, 1993) allow for the entry, transfer, and analysis of data. Prior to data entry, data fields must be determined to assure accurate data entry. SPSS (Version 12.0, Statistical Package for the Social Sciences) (SPSS, 2004) software includes a spreadsheet-type data entry window that was used to set up and enter data for the authors' data set. The maximum possible number of columns were assigned to each field such that when entries were out of the range of possible columns, a star symbol (*) appeared in the frequency tables that indicated an error in data entry. For example, one column was assigned for data entry for responses to items for a measure of coping for data entry since all possible responses were limited to one digit. If during data entry two digits were entered in error, the star symbol was observed in the frequency table.

After entering the data in SPSS, Database Management System (DBMS) software was used to convert the SPSS dataset to SAS datasets. The DBMS copy utility can be used to transfer data between 70 different databases and other applications packages, while giving the user the ability to customize the output data file. The DBMS copy utility can also be used to generate HTML output such that an input matrix can be entered into an HTML table.

The Statistical Analysis System (SAS) software is an integrated software package for data management, analysis, and reporting. Descriptive and inferential statistics programs were written using SAS version 8.0. SAS was selected to analyze the data for two important reasons: 1) SAS is a flexible package that can accommodate a very large number of variables, and 2) SAS allows control over statistical modeling algorithms.

It is important to label all data programming steps in order to create a data set history. Table 2 provides examples of labels used by the authors for each programming activity.

3.6 Security Procedures. One of the most critical components of data management is data security. In the RWHP, several strategies developed prior to data entry, were implemented to maintain data security. To ensure anonymity, participants were assigned an identification number that was entered on all completed data collection instruments. The instruments were stored in a locked file box until they were transferred to the project office. In the project office, instruments were stored in locked file cabinets accessible only to project staff. After instruments were checked by the project manager for accuracy, completeness, and proper coding, they were delivered to the data manager for entry. The location of completed instruments was logged at all times. Once data entry was completed the instruments were returned to the project manager for permanent locked storage. Backup copies of all data files were created on a regular basis as an additional security strategy. Electronic copies of the system code, data, and other related files were stored in the main server of the College, with additional copies stored on a zip disk.

3.7 Documentation: The heart of effective data management is in the documentation. Documentation includes information about recruitment sites, data collectors, and participant progression through the study. A tracking system is necessary to document each participant's progression and completion of required data collection and intervention components. In addition, a log helps to document all decisions related to data collection, including coding, missing data, drop outs, etc.

Additional documentation includes information about coding and recoding of items, variable names, subscale construction and calculation of scale and subscale scores, and any other changes. Ongoing documentation and tracking is an easy and natural task if it is done as the study proceeds. Data managers need to have enough time budgeted to the project to successfully plan, execute, and document data management. If time is not budgeted for these tasks, it is likely that unnecessary mistakes will be made, resulting in flawed results.

4. DATA ORGANIZATION STAGE

4.1 Data Entry, Conversion, Editing, and Cleaning. Accurate data cleansing is critical to the project's success. If done improperly, the results of the study can be skewed. Data entry needs to be performed by well-trained and responsible individuals. Data must be entered with attention to detail and some individuals are better at this than others. In the RWHP, training in data entry included entering sample data and resolving of any issues that occurred. Consistency in data entry is best achieved by one rather than multiple individuals, and as the number of persons involved in data entry increases, the chance for error also increases. However, systematic bias may be an issue with only one data entry individual.

After data entry is accomplished, several activities then need to be performed to check the quality of the data. In our study, the SPSS file was converted to a SAS data set by using DBMS software. Then, the SAS program (RWHP1PRM.SAS) was used to generate the permanent data set.

The data were then examined for accuracy (RWHP1DIS.SAS). The values of interest were then compared to the ultimate reality behind the data. For example, we would not expect a person's age to be less than 0 or greater than 100. Any age that was outside the expected

range was judged invalid. Double entry of data is one alternative to establish data accuracy. However, complete double entry is costly and time consuming. Another method that is often used is drawing a random sample of the cases for double entry. In order to make certain the data had been entered correctly, 10% of the data were re-entered and compared with the original questionnaires. A high degree of accuracy (99%) was observed when we double-checked data entry of the random sample of cases. The identification number and corresponding items were printed and the entered values were compared to the responses on the original questionnaire. The SPSS files were corrected and then converted to SAS datasets. The process was repeated until we were confident that data were clean. All data collection time periods underwent the same process.

4.2 Data Manipulation. Prior to initiating statistical analysis, data manipulation also needs to be completed. Data manipulation includes recoding and creating new variables and creating scales and subscales. New variables are created when it is useful to collapse categories to create fewer response options to achieve more meaningful results. For example, the six response options for the variable "marital status" (single, separated, divorced, widowed, married, living with partner) were collapsed into two marital status categories for analysis, "single" (including single, separated, divorced, and widowed) and "couple".

4.3 Data Merging. In a longitudinal study with multiple data collection points in time, each time point is entered into separate data files. The files will then need to be merged to allow change in variables across time. Checks are implemented to validate that the files were merged properly. The data sets are linked by participant identification numbers.

4.4 Data Backup. A high priority is the creation of backup electronic copies of all files. Electronic copies of the system codes, data, and other related files were stored in the main server of the College. Additional backup files were stored on a zip disk. Hard copies of questionnaires were kept in a locked cabinet in the project office.

4.5 Documentation. During data entry, the data manager documents all of the item codes and recodes, variables names, and the creation of scales and subscales, and any other changes to the data. Additionally, all steps taken to transform, convert, or manipulate data, as well as file mergers must be documented.

5. ANALYSIS AND DISSEMINATION STAGE

5.1 Preliminary Data Analysis. Preliminary analysis of the data is a valuable tool that needs to be included prior to analysis in order to test the research hypotheses. The preliminary analysis can detect various issues that are not specifically related to quality of data, but may be important in making any inferences based on the data. In addition the preliminary analysis allows interim reports for dissemination to project staff. Often, in the process of writing reports major inconsistencies or unrealistic results are first discovered. However, care must be taken to ensure that preliminary analysis does not bias further data analysis. Any changes in data collection instruments or procedures must be documented and the time of implementation of the changes specified exactly. In the RWHP, preliminary data analysis was important to document that the planned procedure was working without any major problems.

5.2 Baseline Data Analysis. Baseline data analysis includes both descriptive and inferential statistics. In our study descriptive statistics were reported for each data collection stage (e.g., time 1, time2, and time3). At this stage an individual with statistical expertise provided consultation and supervised the analysis.

5.3 Linking Longitudinal Data. Since the RWHP was longitudinal with multiple data collection stages, files for each data collection were merged to enable data analyses of effects and patterns across time. General linear model analyses in SAS (GLM and MIXED procedures) were used to exam the effects of: 1) time, 2) treatment, and 3) time by treatment interaction. A SAS macro was used to define parts or collections of SAS statements which could be carried out repeatedly for symbolic names.

5.4 Data Access Procedures. Limiting access to the data is a necessary part of the data management process. In the RWHP, investigators submitted a written request for data analysis that included the purpose of the analysis, variables and relationships to be examined, type of analysis requested, and plans for dissemination. The requests were discussed in team meetings and approved by all of the investigators. In the RWHP, all analyses were supervised by both the data manager and statistician, and any publication was a joint effort of the research team.

5.5 Manuscript Preparation. All research team members need to play a role in planning, developing, and submitting manuscripts. Senior team members can mentor less experienced members in planning analyses and writing reports of findings. Multiple reviews are essential to the development of quality manuscripts and should include reviews by individuals external to the research team. In the RWHP all manuscripts undergo several reviews prior to submission to the appropriate journal for publication.

5.6 Documentation. Documentation is also critical during the data analysis and dissemination stage. All data analysis activities must be documented to create an analysis history. A written summary of each analysis is useful for preventing unnecessary duplication of analyses. Documentation of data analysis is also useful to the process of writing reports for funding agencies.

5.7 Archiving Data. Archived data includes all raw data, the database stored in SPSS datasets, the datasets stored in SAS, all SAS analysis programs, all documentation, and all final standard operational procedures. In the archived data the link between individual and data sets remains separate. Hard copies of raw data and zip discs are secured in a locked storage area used solely for storing archival materials. All data files are also archived in a secure area of the server of the College. A copy of archived data is kept open for ongoing analysis.

6. DISCUSSION

Data management is a critical and essential component of research. The complexity of this process is often not recognized by an inexperienced investigator. A comprehensive data management plan is needed from the onset of the research to organize the data collection and handling processes. It is important to use a clear sequence of data management and procedural steps. Otherwise, a vital step in data preparation may be missed, either at the macro (e.g., data collection procedures) or micro level (e.g., statistical programming). The most sophisticated data analysis programs are useless if the data are not managed properly.
The emphasis on a structured process for data collection and management may seem extreme. However, our experience over many projects indicates that multiple problems may occur when a detailed process is not followed.

In the RWHP, a detailed data management plan was developed prior to data collection. In addition, the data manager and biostatistician were actively involved at all stages of the research, beginning with the grant writing stage. The proposed three staged data management plan outlined in Figure 1, if followed, will produce high quality data organized in an efficient structure that will facilitate data analysis.

7. CONCLUSION

A carefully thought out plan for data management will prevent many of the major data problems that occur in research. Some of the steps may seem self evident and unnecessary and others may be missing. However, a detailed protocol across all stages of a research project will assist in obtaining accurate and complete data to answer the research questions.

Acknowledgment

The data set referenced in this paper was generated for research supported by a grant from The National Institute of Nursing Research, National Institutes of Health (Grant # 1 R01 NR04956).

BIBLIOGRAPHY

Aster, R. and Seidman, R. *Professional SAS Programming Secrets.* Windcrest/McGraw-Hill, New York, New York (1991).

Burlew, M. *SAS Macro Programming Made Easy*. SAS Institute, Cary, North Carolina, (1998).

Burns, N. and Grove, S. K. *The Practice of Nursing Research. Conduct, Critique, & Utilization* (fourth edition). Saunders, Philadelphia, Pennsylvania, (2001).

Cody, R. *Cody's Data Cleaning Techniques Using SAS Software*. SAS Institute, Cary, North Carolina, (1999).

Conceptual Software. Dataflux Corporation, Cary, North Carolina, (1999).

Davidson, F. Principles of Statistical Data Handling. Sage Publications, Thousand Oaks, California, (1996).

Delwiche, L. and Slughter, S. *The Llittle SAS Book*. SAS Institute, Cary, North Carolina, (1996).

DiIorio, F. *SAS Application Programming*. PWS-Kent Publishing Company, SAS Institute, Cary, North Carolina, (1991).

Elliott, R. *Learning SAS in the Computer Lab*. SAS Institute, Cary, North Carolina, (2000).

Hott, J. R. and Budin, W. C. *Notter's Essentials of Nursing Research.* Springer, New York, New York, (1999).

MacMurray, A. R. An Introduction to Data Management. Basic Computer Concepts for Nursing Staff Development Educators. *Journal of Nursing Staff Development, 6,* 168-172, (1990).

Moneyham, L. *A Peer Counseling Intervention for Rural Women with HIV/AIDS.* University of South Carolina, Columbia, South Carolina. R01 NR04956, (1999).

Polit, D. F., Beck, C. T., and Hungler, B. P. Essentials of Nursing Research. Methods, Appraisal, and Uilization (fifth edition). Lippincott, Philadelphia, Pennsylvania, (2001).

Roberts, B. L., Anthony, M. K., Madigan, E. A., and Chen, A. Data Management: Cleaning and Checking. *Nursing Research, 46,* 350-352, (1997).

SAS Institute. *SAS/STAT User's Guide Version 6*. (fourth edition.) (vols. 1-2) SAS Institute, Cary, North Carolina, (1990).

SAS Institute *Combining and Modifying SAS Data Sets*. SAS Institute, Cary, North Carolina, (1995).

Spector, P. *SAS Programming for Research and Social Scientists*. Sage, Newbury Park, California, (1993).

SPSS *SPSS Reference Guide.* Chicago, Illinois, SPSS, (1990).

SPSS. *SPSS Base System User's Guide.* Chicago, Illinois: SPSS, (1994).

SPSS. *SPSS 12.0 Base User's Guide.* Chicago, Illinois, SPSS, (2004).

Youngblut, J. M., Loveland-Cherry, C. J., and Horan, M. Data Management Issues in Longitudinal Research. *Nursing Research, 39,* 188-189, (1990).