PO09
Data management strategies

Abbas S. Tavakoli, DrPH,MPH,ME; Kirby Jackson, MS, PhD(c); Linda Moneyham, DNS, RN, FAAN; Kenneth Phillips, PhD, RN; Carolyn Murdaugh, PhD, RN, FAAN; & Gene Meding, MA *

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. As a consequence, data management has become a specialty requiring specific skills and knowledge. Data management involves preparatory, data entry and analysis/dissemination stages, each of which is equally important to study outcomes. Data management is a critical and essential component of research whose complexity is often not recognized by an inexperienced investigator. The purpose of this poster is to describe the process and strategies of successful data management as applied to a longitudinal data set.

*. University of South Carolina, College of Nursing.
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, as 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, either of which can result in problems. Although information about data management issues is available (e.g., Burns & Grove, 2001; Davidson, 1996; Hott & Budin, 1999; MacMurray, 1990; Polit, Beck, & Hungler, 2001; Roberts, Anthony, Madigan, & Chen, 1997; Youngblut, J. M., Loveland-Cherry, C. J., & Horan, M. 1990), little is written about the more practical aspects of the process.

Purpose

The purpose of this poster is to describe the process and strategies of successful data management as applied to a longitudinal data set.

Background

- The longitudinal data set was generated in a study which tested a peer-based social support intervention designed for a population of rural women with HIV disease.
- Design: Two Group with three times repeated.
- The 280 study participants were recruited from 10 community-based HIV/AIDS service organization serving rural areas of the southeastern United States.
- Intervention group participants received a total of 12 face-to-face peer-counseling sessions over a period of six months.
- The data set was large and complex because of the study design, with intervention group participants nested by study site and peer counselor, and multiple intervention points for each experimental group participant.
- A comprehensive data management plan was designed to organize the data handling processes in order to assure data integrity and security.

Data Management Process

- Preparatory Stage
- Data Organization
- Data Analysis & dissemination
Preparatory Stage

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

Date Organization Stage

This stage includes data entry, conversion, editing, cleaning; data manipulation; merging; backup; preliminary data analysis; documentation.

Analysis and Dissemination Stage

This is the final stage which includes data analyses for baseline; link data over time; data access procedures; manuscript preparation; documentation; and archiving data.

SAS® software is an integrated software package for data management, analysis, and reporting. Descriptive and inferential statistics programs can be written using SAS. SAS should be selected to analyze the data for two important reasons: 1) SAS is a very flexible package that can accommodate a very large number of variables. 2) SAS allows control over statistical modeling algorithms (See examples of SAS program for this project).

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 is not managed properly.

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 other steps 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.
Table 1. Examples of Variable Names

<table>
<thead>
<tr>
<th>Variable Description</th>
<th>Variable Name Base</th>
<th>Variable Name Time2</th>
<th>Variable Name Time3</th>
</tr>
</thead>
<tbody>
<tr>
<td>What county do you live?</td>
<td>County</td>
<td>Bcounty</td>
<td>Ccounty</td>
</tr>
<tr>
<td>Do you have a paying job?</td>
<td>Payjob</td>
<td>Bpayjob</td>
<td>Cpayjob</td>
</tr>
<tr>
<td>Have you been told you have AIDS?</td>
<td>Taid</td>
<td>Btaid</td>
<td>Ctaid</td>
</tr>
<tr>
<td>Coping Scale (54 items)</td>
<td>Cope1-Cope54</td>
<td>Bcope1-Bcope54</td>
<td>Ccope1-Cope54</td>
</tr>
<tr>
<td>Social Support Scales (19 items)</td>
<td>Ss1-Ss19</td>
<td>Bss1-Bss19</td>
<td>Css1-Css19</td>
</tr>
<tr>
<td>Depression Scales (20 items)</td>
<td>Cesd1-Csed20</td>
<td>Bcesd1-Bced20</td>
<td>Ccesd1-Ccesd20</td>
</tr>
</tbody>
</table>

Table 2. Program Name Examples

<table>
<thead>
<tr>
<th>Program names</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rwhp1.sd2</td>
<td>SAS permanent data set for baseline</td>
</tr>
<tr>
<td>Rwhp2.sd2</td>
<td>SAS permanent data set for time 2</td>
</tr>
<tr>
<td>Rwhpprm1.sas</td>
<td>SAS program to create permanent data set for baseline</td>
</tr>
<tr>
<td>Rwhpprm2.sas</td>
<td>SAS program to create permanent data set for time 2</td>
</tr>
<tr>
<td>Rwhpdis1.sas</td>
<td>SAS program to analyze data for baseline</td>
</tr>
<tr>
<td>Rwhpdis2.sas</td>
<td>SAS program to analyze data for time 2</td>
</tr>
<tr>
<td>Rwhp1.log</td>
<td>SAS Logs for baseline</td>
</tr>
<tr>
<td>Rwhp2.log</td>
<td>SAS logs for time 2</td>
</tr>
<tr>
<td>Rwhpout1.lst</td>
<td>SAS output for baseline</td>
</tr>
<tr>
<td>Rwhpout2.lst</td>
<td>SAS output for time 2</td>
</tr>
</tbody>
</table>
libname library 'C:\abbast\moneyham\rwhp1\';
libname rwhp1 'c:\abbast\moneyham\rwhp1\';
**** this program written for southern women health survey ****;
**** for Dr moneyham ****;
**** filename : rwhp1prm.sas ****;
**** author : abbas tavakoli ****;
**** last modification : 03/22/06 ****;

PROC FORMAT library=rwhp1.formats;

VALUE STATEF 1 = 'SOUTH CAROLINA'
             2 = 'GEORGIA'
             3 = 'ALABAMA'
             7 = 'do not know'
             8 = 'refused'
             9 = 'does not apply'
;

VALUE LCITYF 1 = 'NO'
              2 = 'YES'
              7 = 'do not know'
              8 = 'refused'
              9 = 'does not apply'
;

VALUE groupf 1 = 'control'
             2 = 'intervention'
             7 = 'do not know'
             8 = 'refused'
             9 = 'does not apply'
;

data one;
  set rwhp1.rwhp1;

data rwhp1.rwhpls;
  set one;

LABEL
  DATE = 'DATE'
  PID  = 'PARTICIPANT ID#'
  IID  = 'INTERVIEWER ID#'
  IN   = 'INTERVIEW NUMBER'
  DOB  = 'DATE OF BIRTH'
  STATE = 'LIVING STATE'

FORMAT state statef. lcity payjob hiv1-hiv8 welfare comp unemp wic sup ssi ssd mphmon abuspg mast1-mast25 dast1-dast20 lcifty.

Run;
**PROGRAM2. Part of program RWHP1DIS.SAS**

```sas
option nodate nocenter yearcutoff=1910;;

libname library 'c:\abbast\moneyham\rwhp1\';
libname rwhp1 'c:\abbast\moneyham\rwhp1\';

**** this program written for southern women health survey ****;
**** for Dr moneyham ****;
**** filename : rwhp1dis.sas ****;
**** author : abbas tavakoli ****;
**** last modification : 03/22/06 ****;

proc format;
  value tmastgf 0='no abuse'
         1='abuse';

  value martgf 1 = 'SINGLE/sep/div/wid'
              2 = 'MARRIED/living/other'
;
  value ylearngf 1='2001'
                  2='1996-2000'
                  3='1991-1995'
                  4='>1990';
.
.
.

data one;
set rwhp1.rwhp1s;

data two;
  set one;

  **** SETTING THE 9, 99, 999 AS MISSING VALUE **;
ARRAY items race marital living whol-who8
       sex1-sex8 hiv1-hiv8 educ relig payjob
       hourswk job mearner self -- hother ivuse -- tolddep
       aanti -- attempt freq1-freq31 both1-both31 phys1-phys11
       employ timel-time6 ssl-ss19 ssqw -- ssqus cope1-cope54
       cesd1-cesd24 wherecr -- mpot adms relsee whygo
       misdoc -- medcare reasl-reasl9 mislast schmed strcll-sctr14
       sasl-sasl1 nalcoh -- abuspg mast1-mast25 dast1-dast20
       dstr1-dstr14 dsas1-dsas11 ;
  do over items;
    if items=9 then items=.;
  end;

  **** reversing code for the items ****;
array itema time2 time5 time6 cesd4 cesd8 cesdl2 cesdl6 sasl2 sasl4 sasl6
       dsasl2 dsasl4 dsasl6;
  do over itema;
    itema = 5 - itema;
  end;
```

6
**** creating scales and subscales ****;

\[
tf_{freq} = \text{sum (of freq1-freq31)};
\]
\[
tboth = \text{sum (of both1-both31)};
\]
\[
tphys = \text{sum (of phys1-phys11)};
\]

**** Creating new variables from original ****;

if \[0 < t_{mast} < 5\] then \(t_{mastg} = 0\);
   else if \[5 \leq t_{mast} < 51\] then \(t_{mastg} = 1\);

if \[0 < t_{dast} < 6\] then \(t_{dastg} = 0\);
   else if \[6 \leq t_{dast} < 41\] then \(t_{dastg} = 1\);

if \(t_{mastg} = 1\) or \(t_{dastg} = 1\) then \(\text{abuse} = 1\);
   else if \(t_{mastg} = 0\) and \(t_{dastg} = 0\) then \(\text{abuse} = 0\);
   else if \(t_{mastg} = .\) and \(t_{dastg} = 0\) then \(\text{abuse} = 0\);

label

\[
tf_{freq} = \text{'total /hivsymfreq '}
\]
\[
tboth = \text{'total / hivsymboth '}
\]
\[
tphys = \text{'total / fsq activity of daily living'}
\]

format abuse t_{mastg} t_{dastg} t_{mastgf}.

**** descriptive statistics ****;

\text{proc f freq data=one;}
   \text{tables id -- dsas11;}
   \text{title ' frequency tables ';}
   \text{title2 'rural women''s health project';}
   \text{title3 'moneyham/ questionnaire 1 '};

\text{proc means data=two;}
   \text{var ages nchild - dsas1 tfreq -- tdsas ;}
   \text{title ' means ';}
   \text{title2 'rural women''s health project';}
   \text{title3 'moneyham/ questionnaire 1 '};
PROGRAM 3. Part of program Checking Data

```sas
proc format;
  value $missf '' = 'missing'
    other='nonmissing';

**** Macro that takes lower and upper limits for numeric variables and*
**** an id variable to print out an exception report tot the output
Windows ***;
%macro range (dsn, var, low, high, idvar);
    data _null_;
    set &dsn ;
    file print;
    if (&var lt &low and &var ne .) or &var gt &high then
    put "&idvar:" &idvar @15 "variable: &var"
    @35 "value:" &var
    @50 "out-of-range";

run;
%mend range;

%range (one, copel, 0, 3, id);

**** checking data: counting missing values ***;
proc freq;
  tables _character_ / nocum missing;
  format _character_ $missf.;
  title 'frequency tables /missing';
  title2 'rural women\'s health project';
  title3 'moneyham/ questionnaire 1';
proc means nmiss;
  title 'means /missing';
  title2 'rural women\'s health project';
  title3 'moneyham/ questionnaire 1';
run;
```

PROGRAM 4. Part of program RWHPMERGEOU.T.SAS

```sas
option nodate nocenter yearcutoff=1910;;
libname library 'c:\abbast\moneyham\rwhp1\';
libname rwhp1 'c:\abbast\moneyham\rwhp1\';

**** this program written for southern women health survey ****;
**** for Dr moneyham ****;
**** filename : rwhpmergeout.sas ****;
**** author : abbas tavakoli ****;
**** last modification : 06/22/06 ****;
data one;
set rwhp1.rwhpls;
```
data two;
  set one;
.
proc sort;
  by id;
.
libname library 'c:\abbast\moneyham\rwhp2\';
libname rwhp2 'c:\abbast\moneyham\rwhp2\';

proc format;
  value  btmastgf 0='no abuse'
           1='abuse';
data three;
  set rwhp2.rwhp2;

data four;
  set three;
  id = bid;
.
proc sort;
  by id;
.
libname library 'c:\abbast\moneyham\rwhp3\';
libname rwhp3 'c:\abbast\moneyham\rwhp3\';

data five;
  set rwhp3.rwhp3;

data six;
  set five;
  id = cid;
.
proc sort;
  by id;
.
libname rwhp1f    'c:\abbast\moneyham\rwhp1f';
libname rwhp2f    'c:\abbast\moneyham\rwhp2f';
libname rwhp3f    'c:\abbast\moneyham\rwhp3f';
options fmtsearch=(rwhp1f rwhp2f rwhp3f);
data final;
  merge two (in=a) four (in=b) six (in=c);
  by id;
libname rwhpmrg 'c:\abbast\moneyham\merge';

  data rwhpmrg.rwhpmrg;
    set final;
run;
**PROGRAM5. Part of program RWHPMERGEDIS.SAS**

```sas
option nodate nocenter yearcutoff=1910;;

**** this program written for southern women health survey ****;
**** for Dr moneyham ****;
**** filename : rwphmergedis.sas ****;
**** author : abbas tavakoli ****;
**** last modification : 06/22/06 ****;

libname rwhpmrg 'c:\abbast\moneyham\merge';
libname rwhp1f 'c:\abbast\moneyham\rwhp1';
libname rwhp2f 'c:\abbast\moneyham\rwhp2';
libname rwhp3f 'c:\abbast\moneyham\rwhp3';

options fmtsearch=(rwhp1f rwhp2f rwhp3f);

proc format;
  value cesdgf 1=' less 16'
             2='16 +';
  value cesdgfb 1=' less 16'
                 2='16 -25'
                 3='26-35'
                 4='36+';

data one;
  set rwhpmrg.rwhpmrg;
  by id;
  if last.id;
  dtfreq = btfreq - tfreq;
  dtboth = btboth - tboth;

SAS Procedure . . . (see program5).

**Program6. Example of SAS Procedure**

```sas
**** descriptive statistics ****;
proc freq data=one;
  tables id -- dsas11;
  title ' frequency tables ';
  title2 'rural women's health project';
  title3 ' moneyham/ questionnaire 1 ';

proc means data=two;
  var ages nchild child18 hourwk income actual freq1 -- phys11;
  title ' means ';
  title2 'rural women's health project';
```
title3 'moneyham/ questionnaire 1 ';

**** Inferential Statistics ****;
%macro corr (q);

proc corr nocorr alpha nomiss data=two;
var &q;
;
title ' Reliability coefficient';
title2 'rural women's health project';
title3 'moneyham/ questionnaire 1 ';
%mend corr;

%corr (freq1-freq31);

proc corr data=two;
var tfreq -- tcopeis tcesd --;
title ' correlation coefficient';
title2 'rural women's health project';
title3 'moneyham/ questionnaire 1 ';

%macro glm (d,i);
proc glm data=two;
class &i;
model &d = &i ;
lsmeans &i / pdiff cl adjust=tukey ;
;
title ' glm';
title2 ' baseline';
title3 'rural women's health project';
title4 'moneyham/ questionnaire 1 ';
%mend glm;

%glm (tcesd tlslp tlsln tlslo tlslf treas tsctrl tsas,state);

*******************************************************;
**** mix model                                      ***;
******
***************************************************
%macro mix (a,b,c,d,e);
proc mixed noclprint noitprint;
class id time &a;
model &b = &c / s ;
repeated time / type=&d sub=id(group) ;
title ' mixed model ' &e ;
title2 'rural women's health project';
title3 'Dr Moneyham/ merge file ';

%mend mix;

%mix (group site,cesd,group time group*time site*time ,cs,reduced model/cs);