Managing Data for a Long-Term Clinical Research Study With SAS/ASSIST® Software

Stanley Cron, Center for AIDS Research, Baylor College of Medicine, Houston, Texas

ABSTRACT

Data from a long-term clinical research study are usually maintained with a database software package. The Data Management module in SAS/ASSIST software can be used to enter and manage data, thereby eliminating the need to transfer data sets into SAS® software for analysis and report writing. This poster will demonstrate the use of SAS/ASSIST to manage data from a 5 year prospective study of oral outcomes in HIV infected children. Topics to be covered include creating a database, designing formats, and match merging data sets.

INTRODUCTION

Data management for a long-term clinical research study typically involves the use of a database software package. Several data sets are usually created to form a relational database, e.g., one data set for patient demographics, one for laboratory values, etc. The data are then transferred to a statistical package, such as SAS software, for analysis. This process can become rather cumbersome when data sets have to be periodically combined for reports and statistical analysis. The Data Management module in SAS/ASSIST provides a simple system for entering and managing data while eliminating the task of repeatedly transferring numerous data sets into SAS software for analysis.

DATABASE CREATION

The relational database for this study of oral outcomes in pediatric HIV infected patients was created with the Create/Import menu in the Data Management module. One data set was created for each data collection form. The following selections were made from within the Create/Import menu:

1.) Enter data interactively...
2.) Enter data one record at a time...
3.) Data set: (Name, Permanent, Libref)

The name, type, length, and label for each variable were defined in the Define a New SAS Data Set window which appears after step 3. By creating a common variable for subject study number, records for a particular patient could be identified from different data sets. In order to allow for the entry of dates, the MMDDYY6. informat was used to read dates into the data set while the MMDDYY8. format was used to display the dates.

CREATING FORMATS

Due to the fact that almost all of the character variables in this study were coded as numeric, user-defined formats were created to accurately display the different values for variables such as sex and race. Following the selection of the Design Formats menu, the name, type, and value for each format were defined within the Create Formats window. The formats were assigned to the appropriate variables in each data set by using the Utilities menu in the Data Management module. After accessing the data set contents, the formats were placed with their respective variables.

MATCH MERGING DATA SETS

Periodic reports and analysis for this study require that data sets be combined. For example, demographic information from one data set may need to be analyzed with laboratory data from another. As stated earlier, a common variable for subject study number was created to link data from the same patient across different data sets. This common variable allows for the match merging of data sets.

In order to match merge, the data sets had to first be sorted by the common variable (study number). This was accomplished by using the Sort a Data Set window in the Sort menu. The data sets were then merged with the Match Merge (First method) window in the Combine menu. By selecting the data sets to be combined and the common variable used to match subject data, a new data set was created for analysis and report writing.

CONCLUSION

The use of the Data Management module in SAS/ASSIST software can simplify the task of managing clinical research data, particularly in those situations where one person is entering and analyzing the data. Data sets can easily be created and maintained without having to transfer data into SAS software for analysis. As such, SAS/ASSIST software could be considered a viable alternative to database software packages.

ACKNOWLEDGEMENTS

This poster was produced in collaboration with the Baylor Oral Manifestations of Pediatric HIV Infection Study Group in the Department of Pediatrics at Baylor College of Medicine.

Stanley Cron
Center for AIDS Research, Baylor College of Medicine
6621 Fannin, Suite 710 MC 3-2316
Houston, Texas 77030
scron@bcm.tmc.edu

indices for Pearson's correlation coefficients and Cohen's (1977) index of effect sizes, d. McDaniel's program also allowed for user-defined measurement error corrections as suggested by Hunter et al. (1982). In a more recent study (Huffcutt, Arthur, & Bennet, 1993), SAS/BASE procedures (SAS Institute Inc, 1989) was utilized to compute the mean effect size and its variance. This program also allowed for corrections due to sampling errors.

This paper introduces simple SAS code which computes meta-analytic statistics according to the Rosenthal-Rubin (Rosenthal, 1991) method. The program accepts various statistical criteria (e.g., F, t, r, χ^2, z), their p values, and the sample sizes used in testing each hypothesis. The program computes the following meta-analytic indicators: (a) Mean Effect Size, unweighted and weighted by sample size (eq. 3), (b) Combined Significance Level (eq. 4), (c) χ^2 for Diffuse Comparison of Effect Sizes (eq. 5), and (d) χ^2 for Diffuse Comparison of Significance Levels (eq. 6). The following formulæ are used in the computations, which can also be found in Mullen (1989), and Rosenthal (1991):

$$\overline{Z}_f = \frac{\Sigma w_j z_j}{\Sigma w_j} \tag{3}$$

$$\overline{Z} = \frac{\Sigma w_j z_j}{\sqrt{\Sigma w_j^2}} \tag{4}$$

$$\chi^2_{k-1} = \Sigma(n_j - 3)(z_j - \overline{Z}_f)^2 \tag{5}$$

$$\chi^2_{k-1} = \Sigma(z_j - \overline{Z})^2 \tag{6}$$

Σw_j represents the sum of individual weights per hypothesis. Often, sample size is used as a suitable weight, thus allowing studies with larger samples to "contribute" more to the computation of the mean effect size and the combined significance level. The χ^2 criteria for diffuse comparisons have $k-1$ degrees of freedom, where k represents the number of hypothesis tests included in the meta-analysis. The SAS code is as follows:

```
data sugi96;
option ls=75 ps=60;
infile 'studies';
input study hyp stat $ score n df p;
```

One restriction must be noted with regards to the degrees of freedom per test. Only χ^2 with 1 degree of freedom and F tests with 1 df in the numerator are allowed. Rosenthal (1994) notes that often an "omnibus F" (with more than 1 numerator df) will be reported as though it was a contrast between two samples.

```
/*
 transform initial criteria
 to meta-analytic criteria.
 Use z for significance level and
 r (and fisher's z) for  effect size
*/

select (stat);
 when ('t')
  do;
   z=sqrt(df*(log(1+(score**2/df))))*sqrt(1-(1/(2*df)));
   r=sqrt(score**2 /(score**2 + df));
   zf=.5*(log((1+r)/(1-r)));
  end;
 when ('f')
  do;
   z=sqrt(df*(log(1+(score/df))))*sqrt(1-(1/(2*df)));
   r=sqrt(score/(score+df));
   zf=.5*(log((1+r)/(1-r)));
  end;
 when ('x2')
  do;
   z=sqrt(score);
   r=sqrt(score/n);
   zf=.5*(log((1+r)/(1-r)));
  end;
 when ('z')
  do;
   z=score;
   r=sqrt(score**2/n);
   zf=(1/2)*(log((1+r)/(1-r)));
  end;
 when ('r')
  do;
   t=(score*sqrt(n-2))/sqrt(1-score**2);
   z=sqrt(df*(log(1+(t**2/df))))*sqrt(1-(1/(2*df)));
   r=score;
   zf=.5*(log((1+r)/(1-r)));
  end;
 otherwise
  do;
   z=abs(probit(score));
   r=sqrt(z**2/n);
   zf=(1/2)*(log((1+r)/(1-r)));
  end;
 end;
 label n='sample size'
  hyp='hypothesis retrieved'
  stat='statistic used'
  score='value of statistic'
  df='degrees of freedom'
  p='p-level'
  z='z-score'
  zf='Fisher Z transformation of r'
  r='Pearson r';

 /*
  Calculate:
  1) product of Sample Size n (the weight) and Z-score,
  2) squared sample size (w=n**2), to be used
     in estimating the combined significance level.
  3) Weight for Diffuse Comparison of Effect Sizes.
 */
  nz=n*z;
  w=n**2;
  wzf=n-3;
```

```
run;
```

The combined effect sizes and significance levels are estimated using the MEANS procedure several times. First, unweighted and weighted by sample mean effect sizes are calculated and these values are stored in temporary data sets.

```
/*
 Calculations of Combinations of Effect Sizes and
 Significance Levels.
 Calculations of Diffuse Comparisons of E.Ss and S.Ls
 Use separate PROC MEANS to calculate the various
 meta-analytic indices.

 Step 1. Mean Effect Size Unweighted and
         Weighted By Sample Size
*/

proc means noprint data=metanal;
   var zf;
   output out=combzf1 mean=meanzf1;
   run;

proc means noprint data=metanal;
   var zf;
   weight n;
   output out=combzf2 mean=meanzf2;
   run;
```

In step 2, χ^2 for the diffuse comparison of effect sizes are estimated and its value stored also in a temporary data set using PROC MEANS.

```
/*
 Step 2. Calculate chi^2 for
Diffuse Comparison of Effect Sizes.
 Chi^2 has k-1 degrees of freedom.
*/
proc means css noprint data=metanal;
   var zf;
   weight wzf;
   output out=diffzf css=csszf;
   run;

/*
 Step 3. Combinations and Diffuse Comparisons of S.L
 Calculate sums of N*Z and Squared Weights to be
 used for Combination of S.L,
 chi^2(df=k-1) for Diffuse Comparison of S.L.
*/
proc means noprint data=metanal;
   var nz w;
   output out=sigcomb sum=sumnz sumw;
   run;

proc means noprint data=metanal;
   var z;
   output out=sigdiff css=cssz;
   run;
```

This step merges all temporary data sets into one and estimates the combined significance level using the values output in step 3. The appropriate p values are also estimated and labeled.

```
/*
 Step 4. Final Calculations for
 Combined Significance Level,
 Probability of Significance Level, and
 Probability of chi^2 for Diffuse Comparison
   of Effect Sizes.
*/
data final;
   merge combzf1 combzf2 diffzf sigcomb sigdiff;
   zcomb=sumnz/sqrt(sumw);
   probcomb=1-probnorm(zcomb);
   probz=1-probchi(cssz,_FREQ_-1);
   dfz=_FREQ_-1;
   probzf=1-probchi(csszf,_FREQ_-1);
   dfzf=_FREQ_-1;
   keep meanzf1 meanzf2 zcomb cssz csszf
        dfz dfzf probcomb probz probzf;
   label meanzf1='Mean Magnitude of Effect, Unweighted'
      meanzf2='Mean Mag. of Effect, Weighted by n'
      zcomb='Z, Combination of Significance Levels'
      probcomb='Probability for Z'
      cssz='x2, Diffuse Comparison of Sig. Levels'
      probz='Probability of x2'
      dfz='degrees of Freedom'
      csszf='x2, Diffuse Comparison of Effect Sizes'
      probzf='Probability of x2'
      dfzf='degrees of Freedom';
      run;
```

In the following steps, the primary statistics (obtained from the individual studies included in the meta-analysis) are printed, followed by the estimated meta-analytic statistics.

```
/*
 Presentation Step 1.
 Print Primary Statistics of Individual Studies
*/
proc print data=metanal label uniform;
 id study;
 var hyp n stat score df p z r zf;
 title  'Meta-Analytic Integration Using SAS';
 title2 'Initial Statistics and Transformations';
run;

/*
 Presentation Step 2.
 Print Meta-Analytic Statistics obtained with SAS
*/
proc print data=final label uniform noobs;
 var meanzf1 meanzf2 zcomb probcomb cssz
     dfz probz csszf dfzf probzf;
 title2 'Combinations and Diffuse Comparisons';
 title3 'of Effect Sizes and Significance Levels';
 footnote; footnote2;
run;
```

Finally, the GCHART and GPLOT procedures (SAS Institute Inc, 1990) are utilized to demonstrate the distribution of effect sizes and to plot the effect sizes against their respective sample sizes. PROC GCHART is ran to determine any unusual patterns in the distribution of effect sizes which, under favorable

conditions, should be close to normal. PROC GPLOT addresses the perennial problem of *publication bias* against studies which report small effect sizes using small sample sizes (and often remain unpublished) (Mullen, 1989; Rosenthal, 1991). If a publication bias does indeed exist in the area under investigation, the lower left quadrant of the scatter-plot should be empty or have very few observations.

```
/*
 Presentation Step 3.
 Chart of Effect Sizes.
 Use PROC CHART if PROC GCHART unsupported.
*/
goptions gunit=pct cback=black htitle=6
      htext=3 ftext=swissb ctext=yellow;
proc gchart data=metanal;
 vbar zf / midpoints=0 to 1 by .1;
 title2 'Frequency Distribution of Effect Sizes';
run;

/*
 Presentation Step 4.
 Plot Effect Sizes against Sample Sizes
 (aka the funnel plot)
 Use PROC PLOT if PROC GPLOT unsupported
*/

proc gplot data=metanal;
 plot n*zf/ haxis = 0 to 1 by .1
            vaxis = 0 to 120 by 20;
 title2 'Plot of Fisher Zf and Sample Size';
run;
```

3 Conclusions

A meta-analytic integration of the literature has significant advantages when compared to a narrative description and review of the same studies. It is objective, precise, and replicable (Mullen, 1989). Because it offers a quantitative response, it can offer definite answers to literature reviews with mixed results.

Using SAS can facilitate the work of the meta-analytic reviewer. The program can accommodate any number of hypothesis tests, and several statistical criteria. In addition unlike early meta-analytic software packages which had limited graphics capabilities, it provides high resolution graphics which the user can customize and incorporate in other applications s/he uses.

References

Cohen, J. (1977). *Statistical power analysis for the behavior sciences* (rev. edition). New York: Academic Press.

Glass, G. (1976). Primary, secondary and meta-analysis of research. *Educational Research, 5*, 3–8.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-Analysis in Social Research.* Newbury Park, CA: Sage Publications.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* New York, Academic Press.

Huffcutt, A. I., Arthur, W., & Bennet, W. (1993). Conducting meta-analysis using PROC MEANS procedure in SAS. *Educational and Psychological Measurement, 53*, 119–131.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies.* Newbury Park, CA: Sage.

Johnson, B. T., Mullen, B., & Salas, E. (1992). Comparison of three major meta-analytic approaches. *Journal of Applied Psychology, 80*, 94–106.

McDaniel, M. A. (1986). Computer programs for calculating meta-analysis statistics. *Educational and Psychological Measurement, 46*, 175–177.

Mullen, B. (1989). *Advanced BASIC Meta-Analysis.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Rosenthal, R. (1991). *Meta-analytic procedures for social research.* Newbury Park, CA: Sage Publications.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper, & L. V. Hedges (Eds.), *The Handbook of Research Synthesis*, pp. 231–244. New York: Russell Sage Foundation.

SAS Institute Inc (1989). *SAS Procedures Guide, Version 6, Third Edition.* Cary, NC: SAS Institute Inc.

SAS Institute Inc (1990). *SAS/GRAPH Software: Reference, Version 6, First Edition.* Cary, NC: SAS Institute Inc.

Author Information:

Ioannis Dimakos, Computing & Media Services, Syracuse University, 120 Hinds Hall, Syracuse, NY 13244-2390. Email address: idimakos@syr.edu