A SAS Based System for Calculating and Tracking Data Processing Error Rate Estimates

Virginia Ozer and Ellen Gilkerson
Syntex Research

Abstract

An interactive SAS® based system was developed by the data compliance group in the Biostatistics and Clinical Information Processing (BIOCLIP) department to calculate and track data processing error rates for clinical trial data. This system permits tracking and reporting of error rate information across project, study, or form type. In addition to saving time and improving accuracy as compared to previous hand-calculation methods, the system provides detailed feedback on pre- and post-QA error rates to data processing management.

Background

The BIOCLIP department is responsible for the data processing and statistical analysis of clinical trial data at Synrex Research. The department consists of a data processing group that is responsible for data entry, a small data compliance group that reviews the data for errors, and a biostatistics group that analyzes the data and reports conclusions based on the data.

Case report forms are passed from the clinical trial monitors to the data processing group, where the forms are logged in and the data entered into the IDMS data base. The initial data entry function includes online edit, format, and range checks. Periodically, a batch discrepancy report repeats the online checks and in addition performs simple logical cross-checks. After the discrepancies have been reviewed and corrected, the study is released to the data compliance group.

The data compliance group extracts the study data from the data base into SAS files, scans listing of all variables and performs additional checks on the data using SAS. After making corrections, a random sample of each type of form is selected, stratified by investigator. The sample size is usually 20% for variables critical to the analysis (also called 'required fields') and 10% for other variables. After all the data in the samples are compared with the case report forms, the estimated residual error rates are calculated for both critical and other variables.

The previous method

Error rate calculations were performed by hand. The calculations were laborious and prone to error and results consisted of four numbers: the required fields sample size and error rate estimate and the other fields sample size and error rate estimate. No useful information about the kinds of errors found was derived, nor was the information stored in a way that allowed summarization across studies or by other variables of interest.

The system

The system consists of an online component in SAS/FSP® run via CLISTs (execute files of TSO commands), an archive SAS file of error counts and derived error rate estimates, a partitioned data set of SAS code, a format library, automatic report generation, and report-generating capability in batch mode.

Input error counts

A transaction file is created and edited via SAS/FSP using two screens, one for study level parameters which are entered only once for each clinical trial and the other for entry of error counts and evaluative comments specific to each type of case report form.

Calculate error rate estimates

Error rate estimates are calculated for the transaction file at the form level. Summary estimates are calculated for critical and other variables. This is done via a module of simple SAS code which is accessed from a partitioned data set using %INCLUDE.

Maintain archive files

Several generations of the archive file are maintained in a SAS library to provide additional security for the system. Automated back-up of the archive file at the beginning of the session protects against problems resulting from interrupted sessions. The transaction file is appended to the archive SAS file at the end of each session.

Display summary error estimates

Summary error rate estimates for critical and other fields are displayed on the screen through use of PUT statements. This provides immediate access to the information needed for the study report that is written by the analyst.

Generate summary reports

Simple detailed hardcopy reports providing error rate counts and estimates at both the form and study levels are automatically printed at the end of the online session nsing PROC PRINTTO. Formats for the variables printed are stored in a format library. The reports are reviewed by the data compliance section manager to determine whether additional checking is necessary before the study is released for analysis.

Correct existing observations

If corrections to the study and/or form information are necessary, they are implemented through a second CLIST execution. Study and form information are retrieved from the archive file, corrected using SAS/FSP, and returned to the archive file. Calculations are automatically repeated and a new summary report is produced.

Other reports

Management reports developed thus far include reports generated monthly, printed in memo format, and sent to managers in the data processing department. The memo includes error counts, pre- and post-QA error rate estimates for each form , and any evaluative comments entered into the archive. Response has been positive due to the specific nature of the feedback and the informal style of presentation.

Additional reports can easily be developed using SAS PROCs, for example, reports summarizing the number of studies reviewed over a period of time, total number of errors found, or differences in error rates on different types of forms.

Summary

There several clear advantages of this system over the previons hand-calcnlation method. Originally developed to save time and increase accuracy, the system provides more detailed and quantitative information than was feasible with hand-calculation methods. is now available. The system was easy to set np and maintain. Reports can be reformatted and functions added or modified by any moderately skilled SAS programmer. The use of CLISTs and automatic backups has provided a safe system that is transparent to the analyst users. Althongh SAS/FSP has a several dollar setup charge when run nnder TSO, this cost is considered moderate, as npdates are usually made less than once weekly by each user. The primary resource created by the system, the SAS archive of error counts, estimated error rates, comments, and categorization variables, will be increasingly important to management via both regular summary reports and ad hoc reports that can be generated as needed to aid decision-making.

Acknowledgments

The authors are indebted to A.J.L. Cary at Syntex Research for technical assistance.

For additional information or comments, contact the authors at:

Syntex Research
Mailstop A3-430
P.O. Box 10850
Palo Alto, CA 94303

```
PROC 0 ACCESS(NORMAL)
ALLOCATE F(INSAS) DA('MEN.R7152.ESG.ERRCALC.SASLIB') OLD
ALLOCATE F(FT20F001) SYSOUT(A) DEST(R3)
ALLOCATE F(SASCODE) DA('MEN.R7152.ESG.SOFTWARE.LIB') OLD
ALLOCATE F(SASLIB) DA('MEN.R7152.ESG.FORMATS.LIB') SHR
DATA
%SAS

     DATA ONLINE;
          SET INSAS.DUMMY;
          FORMAT QADATE MMDDYY6.;
          QADATE=DATE();

     DATA NEW;
          SET INSAS.NEW;
     PROC DATASETS DDNAME=INSAS NOLIST;
          AGE NEW PREV1-PREV4;

     PROC FSEDIT DATA=ONLINE SCREEN=INSAS.STUDYSCR OPTION=1;
     PROC FSEDIT DATA=ONLINE SCREEN=INSAS.CALCSCRN OPTION=1;

     %INCLUDE SASCODE(ERRCALC);

     OPTIONS  CENTER DATE CAPS LINESIZE=132 PAGESIZE=50;

     %INCLUDE SASCODE(PRINTERR);

     DATA INSAS.NEW;
          SET NEW ALLFORMS;

     ENDSAS;
ENDDATA
FREE FILE(INSAS)
END
```

Figure 1: CLIST used for executing error calculations and adding new observations to archive file.

```
*** SPECIFY OUTPUT DEVICE;
    PROC PRINTTO UNIT=20 NEW;

*** PRINT STUDY SUMMARY INFORMATION;
    PROC PRINT DOUBLE LABEL SPLIT=* DATA=SUMMARY;
         ID STUDY;
         VAR COMPOUND DTHID QADATE VISMOVE RSIZE RERROR
              OSIZE OERROR;
         TITLE SUMMARY INFORMATION;
         FORMAT QADATE MMDDYY8.;

*** PRINT FORM INFORMATION FOR REQUIRED FIELDS;
    PROC PRINT DOUBLE LABEL SPLIT=* DATA=ALLFORMS;
         ID FNAME;
         VAR FMTYPE FNUM RFN RF100 RFNOT
              RFX RFE RFTOT RSAMP RERR RRESID;
         FORMAT FMTYPE TYPEFMT.;
         TITLE REQUIRED FIELDS;

*** PRINT FORM INFORMATION FOR OTHER FIELDS;
    PROC PRINT DOUBLE LABEL SPLIT=* DATA=ALLFORMS;
         ID FNAME;
         VAR FMTYPE FNUM OFN OF100 OFNOT
              OFX OFE OFTOT OSAMP OERR ORESID;
         FORMAT FMTYPE TYPEFMT.;
         TITLE OTHER FIELDS;
         TITLE ALL FIELDS FOR CURRENT STUDY;
```

Figure 2: Code accessed by the '%INCLUDE SASCODE(PRINTERR)' statement, used to print study and form data.

```
COMMAND ===>          Edit SAS data set: WORK.ONLINE          | screen  1
                                                              | obs     1
------------------------------------------------------------------------------

                    Compound RS:_____

                  Study number: _____

                  Other study ID: _______ (e.g., 003/BZL)

                  Data compliance analyst: ___

                  Clinical data specialist: ___

                  Estimated number of visit moves: ____

               Comment: ____________________________________________
```

Figure 3: Screen for study data, used only once per study.

```
COMMAND ===>          Edit SAS data set:  WORK.ONLINE          | screen  1
                                                              | obs     1
------------------------------------------------------------------------------

    Form name _______  Form Type ___  Description _________
          Total number of forms: ____

          Required/critical fields:
              # fields 100% checked:      _____
              # fields not 100% checked:  _____
              # errors pre-sample:        _____ (include 100% & special cks)
              # forms sampled:            _____
              # errors in sample:         _____

          Other fields:
              # fields 100% checked:      _____
              # fields not 100% checked:  _____
              # errors pre-sample:        _____ (include 100% & special cks)
              # forms sampled:            _____
              # errors in sample:         _____

Comment: ____________________________________________
          Sample size reduced? (y/blank): _
```

Figure 4: Screen for form data, used once for each form.