SAS® Macros for Redacting Subject and Clinical Site Codes while Submitting Clinical Trial Data to FDA

Venu Perla, Ph.D.
Independent SAS Programmer, Rockville, MD, USA

SAS Certified Base Programmer for SAS 9
SAS Certified Advanced Programmer for SAS 9
SAS Certified Clinical Trials Programmer Using SAS 9
SAS Certified Statistical Business Analyst Using SAS 9: Regression and Modeling

I. ABSTRACT

One of the final steps in the submission of a clinical study report to the FDA involves the submission of clinical datasets with protected subject and site information. The objective of this paper is to develop SAS macros for redacting subject and clinical site codes in clinical datasets with dummy codes. In this regard, macro 'DUMMY_CODE_MAKER' and macro 'MERGE_MAKER' are created with SAS macro language using DATA step, SQL, SORT and PRINT procedures of SAS. Application of the macros is explained with a model SDTM dataset-Demographics (DM).

II. INTRODUCTION

Submission of a redacted version of the clinical study report (CSR) to the FDA Dockets Management is one of the final steps associated with Investigational New Drug (IND) application in the clinical research. The redacted version should include de-identified datasets that replace patient ID and site ID with some dummy ID [1]. In this context, the objective of this paper is to develop a set of SAS macros that are useful while redacting the patient ID and site ID in SDTM datasets.

CDISC Study Data Tabulation Model (SDTM) domain contains logically related observations with a common topic represented by a single dataset [2]. This paper deals with redaction of only two identifiable variables in the demographics (DM) SDTM dataset viz., 'Unique Subject Identifier (USUBJID)' and 'Study Site Identifier (SITEID).' However, the macros discussed in this paper can be utilized directly or with modifications while redacting similar identifiable variables in other SDTM datasets. Various program elements of SAS, such as DATA step, SQL, SORT and PRINT procedures, and macro language are utilized while developing macros in the paper. All the programs in the paper are created using SAS® University Edition.

III. DEMOGRAPHICS SDTM DATASET AND IDENTIFIABLE INFORMATION

The Demographics (DM) domain includes essential standard variables that describe each subject in a clinical study [2]. A portion of DM dataset from CDISC SDTM Implementation Guide (Version 3.1.2) is utilized here for redacting USUBJID and SITEID. SAS code for creating partial DM dataset is given below with output (**Table 1**). USUBJID and SITEID are identifiable information in the DM Dataset. These traceable variables can be redacted with dummy ID numbers.

```
data DM (label='Partial Demographics Dataset (DM)');
      infile datalines dsd dlm=',' missover;
      input STUDYID: $6. DOMAIN: $2. USUBJID: $11. SITEID: $2.;
      label STUDYID='Study Identifier';
      label DOMAIN='Domain Abbreviation';
      label USUBJID='Unique Subject Identifier';
      label SITEID='Study Site Identifier';
      title 'Partial Demographics Dataset (DM)';
      footnote 'Note: Data is obtained from CDISC SDTM Implementation Guide
               (Version 3.1.2)';
datalines;
ABC123, DM, ABC12301001, 01
ABC123, DM, ABC12301002, 01
ABC123, DM, ABC12301003, 01
ABC123, DM, ABC12301004, 01
ABC123, DM, ABC12302001, 02
;
```

```
run;

proc print data=DM; run;

title;
footnote;
```

Table 1. Partial Demographics Dataset (DM)

Obs	STUDYID	DOMAIN	USUBJID	SITEID
1	ABC123	DM	ABC12301001	01
2	ABC123	DM	ABC12301002	01
3	ABC123	DM	ABC12301003	01
4	ABC123	DM	ABC12301004	01
5	ABC123	DM	ABC12302001	02

Note: Data is obtained from CDISC SDTM Implementation Guide (Version 3.1.2)

IV. REDACTION OF SITEID AND USUBJID

A. MACRO 'DUMMY_CODE_MAKER'

The objective, description of macro parameters and SAS code for the macro 'DUMMY_CODE_MAKER,' are explained below. This macro is utilized for redacting SITEID and USUBJID.

```
/*********************************************************************
Macro                : DUMMY_CODE_MAKER.SAS
Objective            : To recode identifiable variable IDs with
                       reproducible random integer numbers.
Author               : Venu Perla
Date                 : January 29, 2017
SAS version          : SAS University Edition
---------------------------------------------------------------------
MACRO PARAMETERS
min                  : Lower integer value to be considered for random
                       numbers.
max                  : Upper integer value to be considered for random
                       numbers.
                       Note: Consider higher digit values (example: 11111)
                       for 'min' and 'max' to avoid generation of duplicate
                       values.
datset               : Input dataset in which identifiable variable is present.
ivar                 : Identifiable variable in 'datset'.
redvar               : Redacted name for 'ivar'. Contains new integer numbers.
redset               : Output dataset with 'redvar' and other required
                       variables.
keep2                : List of variables to be kept in output dataset
                        (blank separated).
redlabel             : Label for redacted variable (redvar).
*********************************************************************/

%macro dummy_code_maker (ivar=, datset=, min=, max=, redvar=, redset=, keep2=,
                         redlabel=);

      %local ivar datset min max redvar redset keep2 totaln redlabel;

      %*sorting identifiable variable without duplicate values;
      proc sort data=&datset out=&datset._ nodupkey;
            by &ivar;
      run;
```

```
        %*A macro variable for total number of observations in the identifiable
         variable;
        proc sql noprint;
               select count (distinct &ivar) into: totaln
               from &datset._;
        quit;
        %put Total number of observations in &ivar=&totaln;

        %*A macro for generating random integers between MIN and MAX values [See
        Reference 3];
        %macro RandBetween(min=, max=);
               (&min + floor((1+&max-&min)*rand("uniform")))
        %mend RandBetween;

        %*Generation of reproducible random integers between MIN and MAX values
        with a fixed seeding value of 123;
        data rd (drop=i);
               call streaminit(123);
               do i = 1 to &totaln;
                      &redvar = %RandBetween(min=&min, max=&max);
               output;
               end;
        run;

        %*Sorting random numbers and eliminating duplicate values. If duplicate
        values are found in output, rerun the program after
        increasing MIN and MAX value digits (example: 111 to 1111);
        proc sort data=rd out=rd nodupkey;
               by &redvar;
        run;

        %*Concatenating DATSET with RANDOM dataset;
        %*Output dataset with redacted REDVAR and required variables;
        data &redset (keep=&keep2);
               merge &datset._ rd;
               label &redvar.="&redlabel";
        run;

        %*Printing output;
        proc print data=&redset;
        run;

%mend dummy_code_maker;
/*************************************************************************/
*Examples for invoking macro DUMMY_CODE_MAKER;
*%dummy_code_maker (ivar=USUBJID, datset=DM, min=111, max=999, redvar=USUBJNM,
                    redset=DM1, keep2=USUBJID USUBJNM, redlabel=Unique Subject
                    Identifier (Redacted));
```

B. MACRO 'MERGE_MAKER'

The objectives, description of macro parameters and SAS code for the macro 'MERGE_MAKER,' are explained below. This macro is utilized for redacting SITEID and USUBJID.

```
/*********************************************************************
Macro               : MERGE_MAKER.SAS
Objective           : 1. To merge two datasets.
                      2. Optionally, to keep &/or drop certain variables in the
                         merged dataset.
Author              : Venu Perla
Date                : January 29, 2017
```

```
SAS version         : SAS University Edition
-----------------------------------------------------------------------
MACRO PARAMETERS
data1               : First dataset to be merged.
data2               : Second dataset to be merged.
byvars              : List of blank separated variables from first and second
                      datasets. Merging is performed on these variables in
                      order of appearance.
if                  : OPTIONAL if statement. Enter numeric values.
                      If 'If' is not applicable, do not use it.
                           1      :      if A;
                           2      :      if B;
                           3      :      if A and B;
                           4      :      if A not equal to B;
                           5      :      if B not equal to A;
                      Note: This list can be expanded.
keep                : OPTIONAL list of blank separated variables to be kept in
                      merged dataset. If not applicable, do not use it.
drop                : OPTIONAL list of blank separated variables to be dropped
                      from merged dataset. If not applicable, do not use it.
outdat1             : Name of merged dataset.
************************************************************************/

%macro merge_maker (data1=, data2=, byvars=, if=, keep=, drop=, outdat1=);

      %local data1 data2 byvars if keep drop outdat1;

      proc sort data=&data1 out=source1;
            by &byvars;
      run;

      proc sort data=&data2 out=source2;
            by &byvars;
      run;

      data &outdat1;
            merge source1(in=A) source2(in=B);
            by &byvars;
                 %if &if ne and &if=1 %then %do;
                                      %bquote(if A;);
                                      %end;
            %else %if &if ne and &if=2 %then %do;
                                      %bquote(if B;);
                                      %end;
            %else %if &if ne and &if=3 %then %do;
                                      %bquote(if A and B;);
                                      %end;
            %else %if &if ne and &if=4 %then %do;
                                      %bquote(if A ne B;);
                                      %end;
            %else %if &if ne and &if=5 %then %do;
                                      %bquote(if B ne A;);
                                      %end;
            %else                     %do;
                                      %nrstr(%*if;);
                                      %end;
      run;

      data &outdat1;
            set &outdat1;
            %if &keep ne  %then %do;
                          %bquote(keep &keep;);
```

```
                              %end;
                              %else %do;
                              %nrstr(%*keep;);
                              %end;
                  %if &drop ne %then %do;
                              %bquote (drop &drop;);
                              %end;
                              %else %do;
                              %nrstr(%*drop;);
                              %end;
         run;

%mend merge_maker;

*Examples for invoking macro MERGE_MAKER;
*%merge_maker (data1=DM, data2=SF, byvars=SITEID USUBJID, outdat1=OUT1);
*%merge_maker (data1=DM, data2=SF, byvars=SITEID USUBJID, if=1, outdat1=OUT1);
*%merge_maker (data1=DM, data2=SF, byvars=SITEID USUBJID, if=1, keep=SITEID
               STUDYID USUBJID SUBJID, outdat1=OUT1);
*%merge_maker (data1=DM, data2=SF, byvars=SITEID USUBJID, if=1, drop=SITEID
               STUDYID USUBJID SUBJID, outdat1=OUT1);
/*************************************************************************/
```

C. REDACTION OF USUBJID WITH USUBJNM

Two macros discussed above ('DUMMY_CODE_MAKER' and 'MERGE_MAKER') are executed below for redacting USUBJID with USUBJNM (**Table 2** and **3**). If duplicate random integer values are generated, USUBJNM will produce missing values. Under such circumstances, rerun the macro after increasing (or decreasing) MIN and MAX values in the macro.

```
*Redacting USUBJID with random integer numbers and creating USUBJNM;
%dummy_code_maker (ivar=USUBJID, datset=DM, min=111, max=999, redvar=USUBJNM,
                   redset=DM1, keep2=USUBJID USUBJNM, redlabel=Unique Subject
                   Identifier (Redacted));
```

Table 2. Dataset DM1 with USUBJID and USUBJNM

Obs	USUBJID	USUBJNM
1	ABC12301001	142
2	ABC12301002	180
3	ABC12301003	403
4	ABC12301004	455
5	ABC12302001	628

```
*Merging DM and DM1;
*Dropping USUBJID and retaining USUBJNM;
%merge_maker (data1=DM, data2=DM1, byvars=USUBJID, if=1, drop=USUBJID,
              outdat1=DM2);

proc print data=DM2;
       title "USUBJID Redacted to USUBJNM";
run;
title;
```

Table 3. USUBJID Redacted to USUBJNM

Obs	STUDYID	DOMAIN	SITEID	USUBJNM
1	ABC123	DM	01	142
2	ABC123	DM	01	180
3	ABC123	DM	01	403
4	ABC123	DM	01	455
5	ABC123	DM	02	628

D. REDACTION OF SITEID WITH SITENM

Two macros discussed above ('DUMMY_CODE_MAKER' and 'MERGE_MAKER') are executed below for redacting SITEID with SITENM (**Table 4**). If duplicate random integer values are generated, SITENM will produce missing values. Under such circumstances, rerun the macro after increasing (or decreasing) MIN and MAX values in the macro. Final redacted partial demographics dataset (R_DM) with USUBJNM and SITENM is shown in **Table 5** and **6**. Furthermore, contents of R_DM dataset are exhibited in **Table 7**.

```
*Redacting SITEID with random integer numbers and creating SITENM;
%dummy_code_maker (ivar=SITEID, datset=DM2, min=1111, max=9999,
                   redvar=SITENM, redset=DM3, keep2=SITEID SITENM,
                   redlabel=Study Site Identifier (Redacted));
```

Table 4. Dataset DM3 with SITEID and SITENM

Obs	SITEID	SITENM
1	01	1427
2	02	6281

```
*Merging DM2 and DM3;
*Dropping SITEID and retaining SITENM;
%merge_maker (data1=DM2, data2=DM3, byvars=SITEID, if=1, drop=SITEID,
              outdat1=R_DM);

proc print data=R_DM;
       title 'Redacted Partial Demographics Dataset(R_DM) without Labels';
run;

proc print data=R_DM label;
       title 'Redacted Partial Demographics Dataset(R_DM) with Labels';
run;

proc contents data=R_DM;
       title 'Contents of Redacted Partial Demographics Dataset(R_DM)';
run;
```

Table 5. Redacted Partial Demographics Dataset (R_DM) without Labels

Obs	STUDYID	DOMAIN	USUBJNM	SITENM
1	ABC123	DM	142	1427
2	ABC123	DM	180	1427
3	ABC123	DM	403	1427
4	ABC123	DM	455	1427
5	ABC123	DM	628	6281

Table 6. Redacted Partial Demographics Dataset (R_DM) with Labels

Obs	Study Identifier	Domain Abbreviation	Unique Subject Identifier (Redacted)	Study Site Identifier (Redacted)
1	ABC123	DM	142	1427
2	ABC123	DM	180	1427
3	ABC123	DM	403	1427
4	ABC123	DM	455	1427
5	ABC123	DM	628	6281

Table 7. Contents of Redacted Partial Demographics Dataset (R_DM)

Alphabetic List of Variables and Attributes				
#	Variable	Type	Len	Label
2	DOMAIN	Char	2	Domain Abbreviation
4	SITENM	Num	8	Study Site Identifier (Redacted)
1	STUDYID	Char	6	Study Identifier
3	USUBJNM	Num	8	Unique Subject Identifier (Redacted)

V. CONCLUSION

Identifiable information in partial demographics (DM) dataset is redacted with two macros in this paper. The two macros ('DUMMY_CODE_MAKER' and 'MERGE_MAKER') are employed to redact identifiable variables (USUBJID and SITEID) with dummy random ID numbers. In conclusion, these two macros can be utilized directly or with modification while redacting similar variables in other SDTM datasets for FDA submission.

REFERENCES

[1] Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Accessed on January 31, 2017. Available at https://www.gpo.gov/fdsys/pkg/CFR-2002-title45-vol1/pdf/CFR-2002-title45-vol1-sec164-514.pdf
[2] CDISC SDTM Implementation Guide (Version 3.1.2). Available at https://www.cdisc.org/
[3] Wicklin, Rick. 2015. How to Generate Random Integers in SAS®. Accessed on January 31, 2017. Available at http://blogs.sas.com/content/iml/2015/10/05/random-integers-sas.html

ACKNOWLEDGMENTS

I would like to thank the organizers for giving me an opportunity to present this paper at the Philadelphia Area SAS Users Group (PhilaSUG) Winter 2017 Meeting on April 19, 2017 at the PRA Health Sciences, 721 Arbor Way, Blue Bell PA 19422. I would also like to thank Rob Howard, CEO of Veridical Solutions & Adjunct Faculty at the University of California-San Diego; and Justina M. Flavin, Adjunct Faculty at the University of California-San Diego for their suggestions.

RECOMMENDED READING

- Carpenter, Art. 2004. Carpenter's Complete Guide to the SAS® Macro Language, Second Edition, SAS® Institute Inc., Cary, NC, USA.
- Gupta, Sunil. 2016. Sharpening Your Advanced SAS® Skills. CRC Press, Boca Raton, FL, USA.
- Lafler, Kirk Paul. 2013. PROC SQL: Beyond the Basics Using SAS®, Second Edition, SAS® Institute Inc., Cary, NC, USA
- Li, Arthur. 2013. *Handbook of SAS® DATA Step Programming*. CRC Press, Boca Raton, FL, USA.

- *SAS® 9.4 Product Documentation,* SAS Institute Inc., Cary, NC, USA. Available at http://support.sas.com/documentation/94/index.html
- *SAS/STAT® 9.3 User's Guide,* SAS Institute Inc., Cary, NC, USA. Available at http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#intro_toc.htm
- *SAS® 9.2 Macro Language: Reference*, SAS Institute Inc., Cary, NC, USA. Available at http://support.sas.com/documentation/cdl/en/mcrolref/61885/HTML/default/viewer.htm#titlepage.htm
- *SAS® 9.3 SQL Procedure User's Guide,* SAS Institute Inc., Cary, NC, USA. Available at http://support.sas.com/documentation/cdl/en/sqlproc/63043/HTML/default/viewer.htm#titlepage.htm

AUTHOR BIOGRAPHY

Venu Perla, Ph.D. is a SAS Certified Advanced Programmer, Clinical Trials Programmer and Statistical Business Analyst for SAS 9. Dr. Perla is also a biomedical researcher with about 14 years of research and teaching experience in an academic environment. He served the Purdue University, Oregon Health & Science University, Colorado State University, West Virginia State University, Kerala Agricultural University (India), and Mangalayatan University (India) at different capacities. Dr. Perla has published 15 scientific papers and 2 book chapters, obtained 1 international patent on orthopaedic implant device, gave 10 talks and presented 18 posters at national and international scientific conferences in his professional career. Dr. Perla was invited to serve as an editorial board member for several national and international scientific journals. He was trained in clinical trials and clinical data management. Currently, he is actively employing SAS® programming techniques in clinical research.

CONTACT INFORMATION

Phone (Cell): 304-545-5705
E-mail: venuperla@yahoo.com
Web: https://www.linkedin.com/pub/venu-perla/2a/700/468