PharmaSUG 2015 - Paper QT02

# ISPLIT Macro: to split large SAS datasets

Hany Aboutaleb, Biogen, Cambridge, MA

## ABSTRACT

In July 18, 2012, the FDA (CBER) issued a guidance imposing certain requirements on electronic submissions. Under this guidance, an analysis data set must be split if the dataset is greater than 1 GB.
As a result, it is important for programmers to have a convenient and reliable tool to perform this conversion. At Biogen Idec we have developed such a tool in the form of a utility macro. This macro can be used to efficiently and effectively split data sets when they exceed the FDA-imposed size limitation.

## INTRODUCTION

The FDA issued a guidance imposing certain requirements on electronic submissions (Study Data Specifications, July 18, 2012, version 2.0) http://www.fda.gov/downloads/forindustry/datastandards/studydatastandards/ucm312964.pdf. Under this guidance the maximum size of an individual dataset is dependent on many factors; In general, datasets greater than 1 GB in size should be split into smaller datasets, each no larger than 1GB in size. Datasets divided to meet the maximum size restrictions should contain the same variable presentation so they can be easily concatenated.

Datasets which are divided should be clearly named to aid the reviewer in reconstructing the original dataset, e.g., xxx1, xxx2, xxx3, etc. The files that have been divided and need to be concatenated should be noted in the data definition document. This documentation should identify the range of subject numbers (or other criteria used for division) in the label for each of the divided datasets. For further information on file size limitations for files submitted to CBER, contact eData@fda.hhs.gov, for files submitted to CBER, contact CBER.CDISC@fda.hhs.gov. Before splitting into smaller datasets, the large dataset should have all the variables resized to their maximum length prior to splitting. Split data should be noted in the data definition document clearly identifying the method used for the dataset splitting. CDISC Implementation Guidelines (Study Data Tabulation Model Metadata Submission Guidelines - page 24) makes clear that, sponsors may “split” domains. Splitting a domain means defining the domain in terms of sub-components. Split domains are not specific to “Findings” domains; however Findings may lend themselves to being split because they can become quite large. The rules associated with splitting domains are in SDTMIG Section 4.1.1.7. In the Sample submission from Section 4.1.1.7, the QS domain was split into 3 datasets using the questionnaire name in QSCAT. This QS domain has been split for illustrative purposes only. The intention is to show how to split domains, and does not speak to any rationale for splitting domains. The domain variable value for all split domains is QS, however, the dataset names are unique and prefixed with QS. The annotated CRF refers to the domain name (QS) as opposed to the dataset name (QSCG, QSCS, or QSMM). If the decision had been made to split a submission dataset, it is recommended that the sponsor communicate with their review division regarding exactly what needs to be included in the submission, i.e. the un-split datasets and the split datasets.

## SPLITTING DOMAINS

Sponsors may choose to split a domain of topically related information into physically separate datasets. In such cases, one of two approaches should be implemented:

1. For a domain based on a general observation class, splitting should be according to values in --CAT (which must not be null).
2. The Findings about the (FA) domain (321HSection 6.4) can be split either by --CAT values (per the bullet above) or relative to the parent domain of the value in --OBJ. For example, FACM would store Findings about CM records. See 322HSection 6.4.2 for more details.

The following rules must be adhered to when splitting a domain into separate datasets to ensure they can be appended back into one domain dataset:

1. The value of the DOMAIN must be consistent across the separate datasets, as it would have been, had not been split (e.g., QS, FA).
2. All variables that require a domain prefix (e.g., --TESTCD, --LOC) must use the value of DOMAIN as the prefix value (e.g., QS, FA).
3. --SEQ must be unique within USUBJID for all records across all the split datasets. If there are 1000 records for a USUBJID across the separate datasets, all 1000 records need unique values for --SEQ.
4. When relationship datasets (e.g., SUPPxx, FAxx, CO, RELREC) relate back to split parent domains, IDVAR should generally be --SEQ. When IDVAR is a value other than --SEQ (e.g.

--GRPID, --REFID, --SPID), care should be used to ensure that the parent records across the split datasets have unique values for the variable specified in IDVAR, so that related children records do not accidentally join back to incorrect parent records.

## THE %ISPLIT MACRO

The %isplit macro, will handle SAS datasets when they exceed the maximum size allowed of 1 GB, with some flexibility, in general, data sets up to 1.25 GB would not need to be split, the macro will check the target datasets for maximum size, and will issues. A detailed report, illustrated below, formulates a suitable Data step that will process a large SAS dataset and create a number of smaller datasets having comparable cardinality. The macro will split the data according to the pervious rules of split based on categorical variables – first, by --CAT, and then, if necessary, by --SCAT. Otherwise, the user can specify with the parameter splitvar how to split the data into a desired number of smaller data sets, for example, %isplit(SPLITVAR=FACM). The user chooses the name for the split data with parameter prefix, and can name the split folder with parameter SPLITLIB the default is split, and can change the FDA-imposed parameter SPLITSIZE.

## MACRO INPUT PARAMETERS

| | |
|---|---|
| DS: | Name of domain that need to be split |
| LIBIN: | Name of the libname that has dataset to be processed |
| PREFIX: | Prefix name of the split data |
| SPLITVAR: | Split variable used to split the data |
| SPLITLIB: | Name of the split data sets output folder (default=split) |
| SPLITSIZE: | FDA-imposed size limitation (default=1.25) |
| REPORT: | Generate report for the split data (default=Yes) |
| DEBUG: | Debug the macro (YES/NO) (default: NO) |
| CLEANUP: | Clean up temporarily datasets (YES/NO) (default=YES) |
| VERSION: | Version control for future use in case of new release to the macro (default=1) |

## SAMPLE CALL:

```
libname crtdir '/biostats/drug/study/test/splitdata';
%isplit(ds=crtdir.adlb,prefix=adlb,splitsize=1,debug=Y);
```

## CONCLUSION

The %isplit macro is a useful tool for splitting large data sets, as per FDA guidance, into more manageable datasets, and, it offers a good lesson in using Macro Language. Domain splitting is the necessary action to take when datasets are too big, the process flow for optimizing a dataset size, in a data submission, involves reducing the dataset size as much as possible before splitting.  This means that the first step is to reduce the size of variables.  Once the domains is at its minimum size through this approach, then examine the size of each data set and split according to the FDA guidance, using the “FDA/PhUSE CSS Data Quality Working Group Best Practices” document.

## ACKNOWLEDGMENTS

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

The author would like to thank my manger Matthew Wien for his valuable comments to this paper.

### References

**[1]** ***SAS® 9.2*** **Macro Language: Reference. Cary, NC: SAS Institute Inc.**
**[2]** ***SAS® 9.2*** **Language Reference: Concepts. Cary, NC: SAS Institute Inc.**
**[3]** ***SAS® 9.2 Language Reference: Dictionary*****. Cary, NC: SAS Institute Inc.**
**[4] “*FDA/PhUSE CSS Data Quality Working Group has initiated a Best Practices document”***
http://www.phusewiki.org/wiki/index.php?title=Data_Sizing_Best_Practices_Recommendation .
**[5] “**Study Data Specifications, July 18, 2012, version 2.0”
http://www.fda.gov/downloads/forindustry/datastandards/studydatastandards/ucm312964.pdf

# ISPLIT Macro: to split large SAS datasets

## CONTACT INFORMATION

Your comments and questions are valued and encouraged.
Contact the author at:

**Hany Aboutaleb**
Biogen
14 Cambridge Center
Cambridge MA 02142
Works Phone: (617) 914-7125
Fax: (617) 679-3280
Email: hany.aboutaleb@biogen.com
LinkedIn: Hany Aboutaleb

## ISPLIT Macro: to split large SAS datasets

### Source Code Sample

```
%*--------------------------------------------------------------------
Macro DsnSize demonstrate how the size of a data set is estimated with
the library information extracted from the PROC CONTENTS procedure.
The input SAS data set is represented by the parameter ds.
The macro %DsnSize outputs a global macro variable, called ds_mb.
That holds the size of the input SAS data set in Gigabytes (GB).
---------------------------------------------------------------------;

%macro DsnSize(ds= /* one or two level input SAS data set */);
 %global ds_mb ds_gb ds_;
  %if %index(&ds,%str(.)) ne 0
     %then %let ds_ = %substr(&ds,%eval(%index(&ds,%str(.))+1));
    %else %let ds_ = &ds;

ods output "Library Members"=LibInfo;
ods listing close;
  proc contents data=&ds memtype=DATA;
  run;
ods output close;
ods listing;
  data _null_;
   set LibInfo;
    if name eq "%upcase(&ds_)" then do;
    call symputx('ds_gb', round(FileSize/1073741824,0.01));
     call symputx('ds_mb', round(FileSize/1048576,0.01));
  end;
run;
%mend DsnSize;
```