SESUG Paper 71-2025

Optimizing Clinical Research Efficiency through SDTM Dataset Splitting Using SAS: Insights from New SDTM Programmer

Maddie Hays and Tiffany Bainter, Rho, Inc.

ABSTRACT

Splitting SDTM (Study Data Tabulation Model) domains into multiple datasets can serve as a valuable tool, streamlining the process of creating and reviewing large, complex data domains. In SDTM, a general observation class domain may be split into separate datasets. In practice, we commonly split the Findings About (FA) and QS (Questionnaires) domains. Common reasons for dataset splitting include better representation of complex data, improved traceability from Case Report Forms (CRF) to SDTM specifications to datasets, and simplified programming and validation. Domains can be split using several methods, including purpose, category, or time points. Creating compliant and consistent split domains can be supported by programming habits in SAS to organize and check your data. This paper will cover practical tips and FAQs such as, ensuring consistent variable naming across split datasets, using unique identifiers (e.g., STUDYID, USUBJID,-SEQ) to link split datasets, updating the Define.XML file, CDISC compliance queries, CRF annotation and SAS code examples.

INTRODUCTION

SDTM domains are often large and complex, making them difficult to manage and review. Splitting these domains into smaller, more manageable datasets can improve clarity, traceability, and compliance. This paper discusses the motivations behind dataset splitting and provides practical guidance for implementation.

REASONS TO SPLIT

The most common reasons for dataset splitting include better representation of complex data, improved traceability from Case Report Forms (CRF) to SDTM specifications to datasets, simplified programming and validation, and enhanced data quality. Splitting can also aid in strategic segmentation for targeted analyses and improved compliance with FDA regulations. If a dataset is over 5GB, the FDA Study Data Technical Conformance guide states the SDTM dataset should be split.

SPLITTING METHODS

There are a few main methods of splitting domains: by purpose, category, or time points. For example, a QS (Questionnaires) dataset can be split by purpose when a study collects multiple different questionnaires. In a study with a large LB (Laboratory Results) dataset, it could be split by category (e.g., Hematology, Chemistry, Urinalysis) or by longitudinal time points (e.g., Baseline, Week 12, Week 24).

DEFINE-XML AND REVIEWER'S GUIDE DOCUMENTATION

Reasons for splitting datasets can be categorized into two groups: those due to file size and those due to unique dataset definitions. The way in which the split should be documented depends on the type of split. If the split is due to file size, no further documentation is needed in the Define-XML. In the Reviewer's Guide, the split should be explained in the dataset subsections. The following is an example taken from the SDTM MSG v2.0.

> **3.4.4 LB – Laboratory Test Results**
>
> An unsplit LB is included in the "sdtm" folder and represented in the Define-XML document as a single table. Because of the excessive file size, LB is then split by category, LBCAT, into LBCH ("Chemistry"), LBHE ("Hematology") and LBUR ("Urinalysis" and "Other"). The split datasets have been placed into the "split" sub-folder.

If the split is due to unique dataset definitions, each split dataset should be documented in the Define-XML to ensure clarity and regulatory compliance. In the Reviewer's Guide, a separate subcategory should be created for each of the split datasets. Below is an example of this type of split, as referenced in SDTM Metadata Submission Guidelines (MSG) v2.0.

> **3.4.6 QS – Questionnaires-QSPH**
>
> QS was split by sponsor decision. QSPH contains the Patient Health Questionnaire-9 data.
>
> **3.4.7 QS – Questionnaires-QSSL**
>
> QS was split by sponsor decision. QSSL contains the Satisfaction with Life Survey data.

PRACTICAL TIPS AND FAQS

To ensure consistency across split datasets, use consistent variable naming and lengths so split datasets can easily be concatenated together.

The DOMAIN value should remain the same across datasets (e.g., DOMAIN='QS' in QSCS and QSHA). Additionally, Dataset labels should also remain the same as these datasets are considered as one domain. This can be set using a standardized macro to assign metadata or using the LABEL= option in a data step or PROC SORT:

```
proc sort data=final out=sdtm.qscs(label="Questionnaires");
 run;
proc sort data=final out=sdtm.qsha(label="Questionnaires");
 run;
```

Variables requiring domain prefixes must use DOMAIN as the prefix (e.g., QSTESTCD).

The --SEQ variable must be unique within USUBJID across all records. One strategy is to add a multiple of 1000 to --SEQ across different domains. This assumes there will be no more than 1000 records per subject for that domain, if it is greater than adjust the strategy accordingly. The below data step can be used to assign --SEQ:

```
data x2;
 set x1;
  by sortvars;
    if first.USUBJID then QSSEQ = 1000;
     Else QSSEQ + 1;
 run;
```

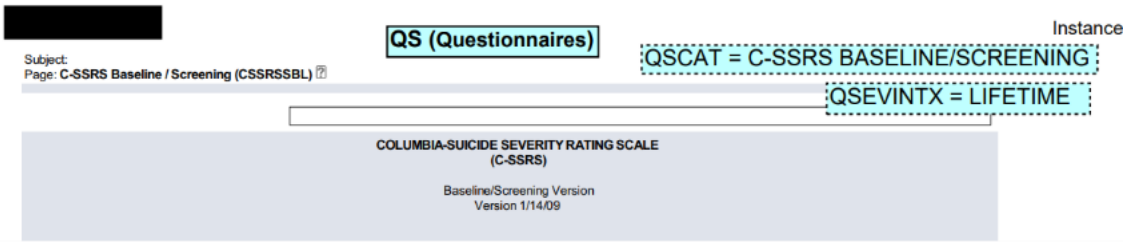
While SDTM character variable length should be limited to the longest length of the variable, all split datasets should be considered when setting the length for variables in datasets and the define.xml. Internal macros should be utilized to ensure variable length is defined by the data once it is final. There are multiple ways in SAS this can be achieved. The LENGTH function can be used to quickly find the length of a character variable. To calculate the longest iteration, follow with a PROC SORT, PROC SQL or PROC MEANS:

```
/*LENGTH Function*/
   data qs1;
    set qs;
    ltest =length(QSTEST);
   run;

   /*PROC SORT*/

   proc sort data=qs1;
     by descending ltest;
   run;

   data qs2;
    set qs1;
     if _N_=1 then output;
   run;

   /*PROC SQL*/
     proc sql;
       create table qs2 as select max(ltest) as maxtest from qs1;
     run;

   /*PROC MEANS*/
   proc means data=qs1 ;
    var ltest;
     output out=qs2 max=maxtest;
   run;
```

Consistency in the metadata between split domains makes it possible to seamlessly combine datasets with a simple data step:

```
   data QS;
     set QSCS QSHA;
   run;
```

When annotating CRFs, use domain names rather than dataset names (e.g., QS instead of QSCS/QSHA). The following is an example of how to annotate a split domain. The content of this page will go into split domain QSCS. The domain is annotated as QS but QSCAT is used to identify which form the data is coming from.



If a file exceeds the size limit (5 GB) and needs to be split, submit the smaller split files in a designated "split" sub-folder, along with the original full-size file in the main data folder. There is no need to include an additional define.xml file within the split sub-folder. Datasets split due to unique definitions should be included in the SDTM datasets directory.

CONCLUSION

Splitting SDTM domains can significantly enhance the efficiency of clinical research data management. By following best practices and documenting appropriately, researchers can ensure clarity, compliance, and streamlined analysis. This paper provides a comprehensive overview of the motivations, methods, and practical considerations for dataset splitting.

REFERENCES

Study Data Technical Conformance Guide - Technical Specifications Document | FDA Center for Drug Evaluation and Research. Study Data Technical Conformance Guide - Technical Specifications Document." FDA. 27MAR2025. Available at https://www.fda.gov/regulatory-information/search-fda-guidance-documents/study-data-technical-conformance-guide-technical-specifications-document

SDTMIG v3.4 | CDISC. "SDTMIG v3.4." Clinical Data Interchange Standards Consortium, Inc. 29NOV2011. Available at https://www.cdisc.org/standards/foundational/sdtmig/sdtmig-v3-4.

SDTM Metadata Submission Guidelines v2.0 | CDISC. "SDTM Metadata Submission Guidelines v2.0" Clinical Data Interchange Standards Consortium, Inc. 30MAR2021. Available at SDTM Metadata Submission Guidelines v2.0 | CDISC.

CONTACT INFORMATION

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

Maddie Hays
Rho, Inc.
maddie_hays@rhoworld.com

Tiffany Bainter
Rho, Inc.
tiffany_bainter@rhoworld.com