PharmaSUG 2013 - Paper DS21-SAS

Some Strategies for Validating Your Data before Submission

Frank Roediger, SAS Institute, Cary, North Carolina
Sandeep Juneja, SAS Institute, Cary, North Carolina

ABSTRACT

Everyone making clinical trial submissions tries for the process to go as smoothly as possible, but there are some places where missteps can cause delays. Avoiding these missteps can greatly streamline the submission review process.

The CDISC SDTM Implementation Guides have hundreds of pages that provide exhaustive (and exhausting) detail about dozens of domains and their variables, but those Implementation Guides don't provide any information about how long any of the variables should be. Careful data design and some utility processes can take the guesswork out of assigning lengths to character variables in submission data sets..

V5 transport files are the approved mechanism for transmitting clinical trial data within a submission. The usual way for creating them is with PROC COPY and the XPORT Libname Engine. But now there are some special-purpose macros that can be downloaded from SAS that can create both V5 transport files (for clinical trial submissions) as well as V8 and V9 transport files (for non-submission purposes).

The define.xml contains information about what a reviewer can expect to find in a submission. Because discrepancies between the define.xml and the submission data sets can cause delays in the FDA review, it is very important to make sure that the define.xml truly reflects the submission data's metadata. The SAS Clinical Standards Toolkit (CST) provides a way to reconcile the define.xml with submission data metadata so that any discrepancies can be resolved before a submission is made.

INTRODUCTION

Preparing clinical trial data for submission involves a succession of tasks, each of which has its own special set of challenges. We have organized these tasks into a common set of eight steps (see Figure 1) and have identified some techniques that we have used to help the steps flow smoothly. We also outline some diagnostic utilities to verify that clinical trial data is in the appropriate state at specific steps, and provide links to coded examples of those utilities.

Our presentation focuses on the following issues:

- managing the length of character variables (Steps 1-3)
- creating a warehouse of clinical trial data and extracting analysis marts from it (Steps 2-3)
- converting clinical trial data sets into V5 transport files (Step 4)
- verifying that the define.xml accurately documents the data in the V5 transport files (Steps 5-8)

CHARACTER FIELD LENGTHS

The CDISC Implementation Guides provide a clear inventory of all the fields that can be included in the standard version of all the SDTM domains. Along with the inventory are specifications for the attributes for all the fields – except for the length of character fields. Because of the submission requirement that data sets be delivered as SAS Version 5 transport files, the maximum size for character fields is 200 bytes. It is tempting to assign all these character fields the maximum length of 200 bytes. However, the FDA, in *Study Data Specifications*, has clearly indicated that "the allotted character column length/size for each column should be the maximum length used." Consequently, creating every character field with a length of 200 bytes is not an option. We need to determine how big to make each submission character field so that it can accommodate all the values that it is to contain.

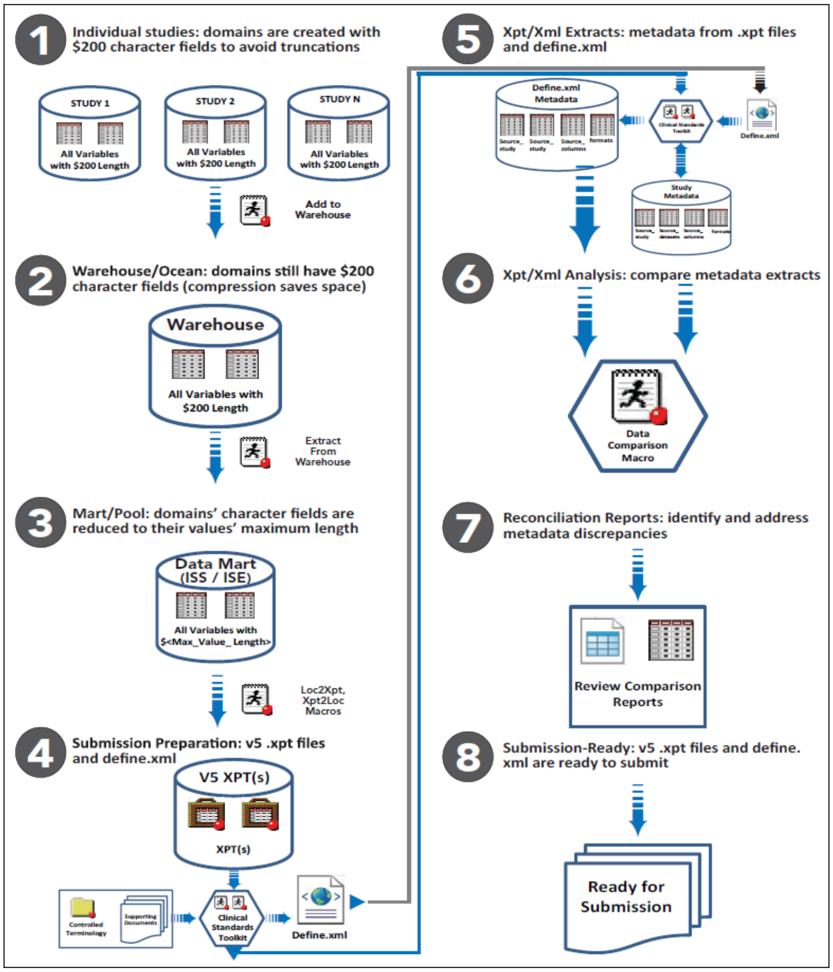
1
Individual studies: domains are created with $200 character fields to avoid truncations
STUDY 1
All Variables with $200 Length
STUDY 2
All Variables with $200 Length
STUDY N
All Variables with $200 Length
Add to Warehouse
2
Warehouse/Ocean: domains still have $200 character fields (compression saves space)
Warehouse
All Variables with $200 Length
Extract From Warehouse
3
Mart/Pool: domains' character fields are reduced to their values' maximum length
Data Mart (ISS / ISE)
All Variables with $<Max_Value_Length>
Loc2Xpt, Xpt2Loc Macros
4
Submission Preparation: v5 .xpt files and define.xml
V5 XPT(s)
XPT(s)
Controlled Terminology
Supporting Documents
Clinical Standards Toolkit
Define.xml
5
Xpt/Xml Extracts: metadata from .xpt files and define.xml
Define.xml Metadata
Source_study
Source_study
Source_columns
formats
Clinical Standards Toolkit
Define.xml
Study Metadata
Source_study
Source_datasets
Source_columns
formats
6
Xpt/Xml Analysis: compare metadata extracts
Data Comparison Macro
7
Reconciliation Reports: identify and address metadata discrepancies
Review Comparison Reports
8
Submission-Ready: v5 .xpt files and define. xml are ready to submit
Ready for Submission

Figure 1. Eight Steps for Preparing Clinical Trial Data for Submission

We can minimize the problem of character field length if we remind ourselves that the FDA's stipulation about limiting character variables to "the maximum length used" applies only to the data sets that are bundled as SAS Version 5 transport files and submitted to the FDA, and not to any up-stream versions of the data sets. Framing the issue this way frees us up to use the maximum 200-byte size for the character fields in all versions of the data sets until we prepare the submission data sets, at which time we re-size the character fields so that they meet the FDA's expectation that "the allotted character column length/size for each column should be the maximum length used."

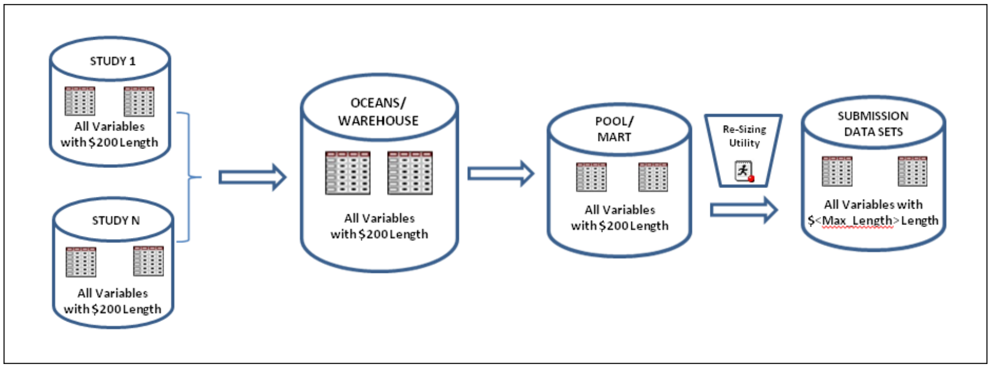
STUDY 1
All Variables with $200 Length
STUDY N
All Variables with $200 Length
OCEANS/ WAREHOUSE
All Variables with $200 Length
POOL/ MART
All Variables with $200 Length
Re-Sizing Utility
SUBMISSION DATA SETS
All Variables with $<Max_Length> Length

Figure 2. Delay Final Sizing of Character Field Lengths until Creating Submission Domains/Data Sets

Figure 2 depicts the position of the Re-Sizing Utility program within the context of an integrated submission:

- the study-level domains are combined in a warehouse
- the rows for studies to be integrated are extracted from the warehouse domains and are stored in a mart
- the mart domains are processed by the Re-Sizing Utility to create the submission data sets

In a single-study submission, the first two steps would be bypassed and the Re-Sizing Utility would create the submission data sets directly from the study domains.

RE-SIZING UTILITY

The SAS Drug Development Forum is a Web page (https://communities.sas.com/community/support-communities/sas-drug-development) where users can get help with pharma programming issues. Although the Forum was set up specifically to help SAS Drug Development users, it also contains information about more general pharma questions as well. A sample utility program that re-sizes character fields is available on the SAS Drug Development Forum.

The utility’s major functional tasks are outlined in the following bullet points:

- Inventory all the character fields in the source domain
- Create a parallel data set with one row for each row in the source domain and with one numeric field for each character field in the source domain (see Table 1, below); populate the numeric field with the length of the value in the corresponding character field
- Determine the maximum value for each of the parallel data set’s numeric fields (this will be the length of the character field in the submission domain)
- Create the submission domain with the new character lengths
- Verify that the submission domain is identical to the source domain, except for the length of character fields

Source Domain		**Parallel Data Set**	
STUDYID	**USUBJID**	**STUDYIDx**	**USUBJIDx**
ABCD	ABCD-1	4	6
ABCD	ABCD-99	4	7
ABCD	ABCD-999	4	8

Table 1. Example of Re-Sizing Utility’s Parallel Data Set

SDTM WAREHOUSES AND MARTS

OCEANS, TRIBUTARIES, AND POOLS

In Figure 3, the study-level domains are migrated into a warehouse structure that is labeled, “Oceans.” This type of warehouse is described more fully in *Creating Analysis Data Marts from SDTM Warehouses*, but is summarized here as a convenient reference.

There are two approaches for integrating study-level domains: directly from the study-level data structures, and indirectly from a warehouse into which the study-level domains have already been aggregated. “Oceans” refers to a warehouse that has been constructed specifically for clinical trials’ SDTM domains (we use the term, “Oceans,” to help distinguish the aggregated form of the domains from the study-level forms (“Tributaries”) and the analysis-level subsets (“Pools”) – see Figure 3; note that there are two “Oceans,” one for active studies and one for closed studies).

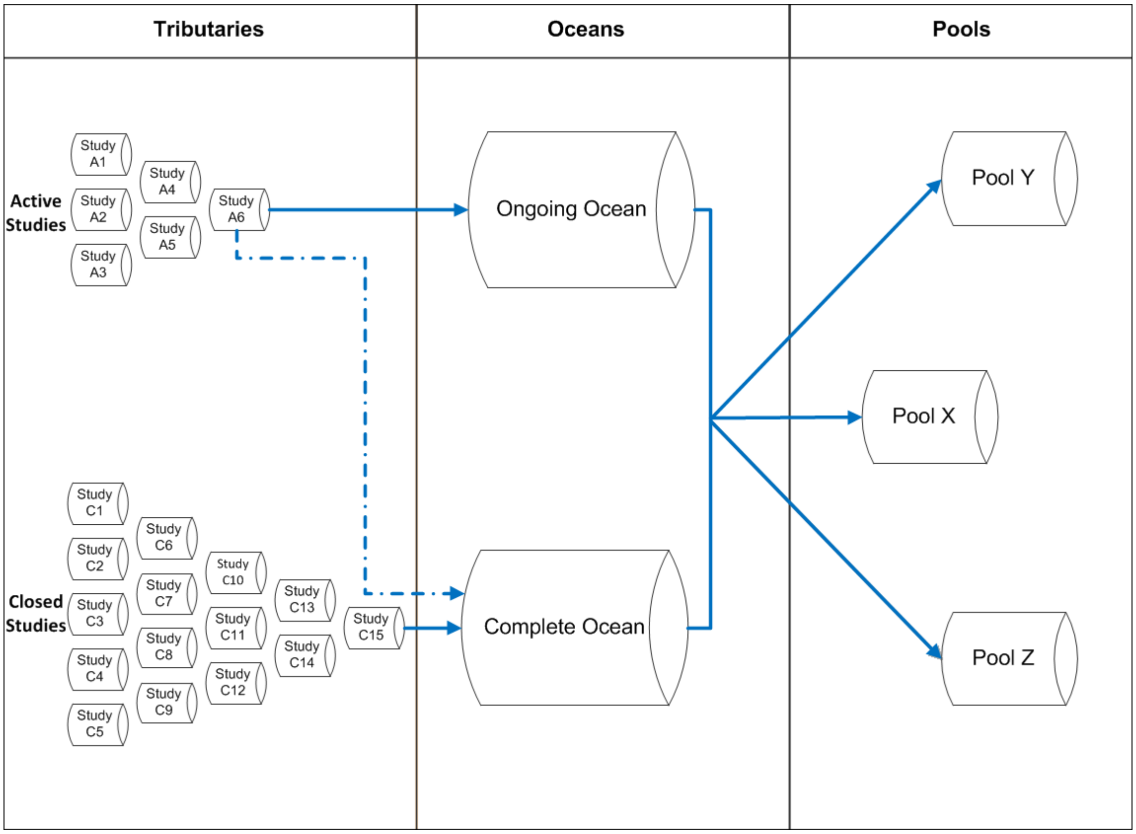

Figure 3. Flow in Warehouse Strategy of Domain Management

The "Oceans" approach to domain management requires more overhead than aggregating domains directly from their studies, but it has several advantages, too.

- Data harmonization: Even if every study uses a common set of domain definitions, they probably will not share the same version of thesauruses (such as MedDRA) and Controlled Terminology, especially when the studies have been conducted over a span of time.
- Data currency: Having all usable data in the "Oceans" can take the guesswork out of whether you are accessing the current version of domains.
- Validation: Even though the individual studies' domains may be validated, an on-demand ad hoc integration of those domains is not.

MAINTAINING THE OCEANS

There are two Oceans: the Ongoing Ocean, which contains the domains from active studies, and the Complete Ocean, which contains the domains from closed studies. The Ongoing Ocean is rebuilt with entirely new content whenever it is updated (this makes sure that the latest version of the domains for active studies is available); the Complete Ocean is updated with only the studies that have closed since the last update (this makes sure that the final version of the domains for closed studies is available). The two Oceans are mutually exclusive; whenever an active study is closed, its domains are deleted from the Ongoing Ocean, and its final domains are added to the Complete Ocean.

Both Oceans can be up-versioned so that their domains refer to the latest thesaurus and terminology standards. And both Oceans have inventory tables that keep a record of their domains' "pedigrees."

CREATING POOLS

Over time, the Oceans (particularly the Complete Ocean) will grow quite large. So that you can focus on only a segment of the Ocean (for example, on studies of a specific compound with a specific trial design), there is a process that can copy a subset of the Ocean and store it as a separate Pool. As with the Ocean, each Pool has an inventory of the "pedigree" of its contents. Each Pool's "pedigree" also contains a record of the selection criteria that were used to identify its studies in the Oceans.

SOMETHING NEW FOR CREATING TRANSPORT FILES

Domains, whether they are from a single study or from a Pool, need to be bundled as SAS V5 transport files before they can be submitted to the FDA. Traditionally, the code for creating a SAS V5 transport file is a PROC Copy step that uses the XPORT Libname Engine, like the following example:

```
libname sdtm "<path for SAS library with SDTM domains>";
libname _xpt xport "<path/filename for .xpt>";
proc copy in     =sdtm
          out    =_xpt
          memtype=data;
   select <domain data set name>;
run;
```

There is now another way to create a SAS V5 transport file: the *%loc2xpt* macro. SAS Usage Note 46944 describes how to use *%loc2xpt* and its companion macro, *%xpt2loc* (which creates SAS data sets from transport files). Usage Note 46944 can be located by searching the SAS Tech Support site (support.sas.com) or can be directly accessed from its own URL (http://support.sas.com/kb/46/944.html).

Usage Note 46944 also has a *Downloads* tab with clickable links for downloading the macros and the XPTCOMMN.sas file, which contains code that is used by both macros. Follow the instructions in the Usage Note for incorporating the macros into your own SAS environment.

The following example provides a convenient reference for using the *%loc2xpt* macro:

```
libname sdtm "<path for SAS library with SDTM domains>";
filename xptfile "<path/filename for .xpt>";
%loc2xpt(libref= sdtm
        ,memlist=<domain data set name>
        ,filespec=xptfile
        ,format=V5
        );
```

The two methods appear very similar and require you to provide the same information, so why would you choose to use the macro instead of the PROC?

One reason to opt for the macro is that you would need only one method for creating transport files, independent of the version of the type of transport file you create. Note that the *%loc2xpt* macro has an optional Format= parameter. In the example, the argument for this parameter specifies that we want to create a V5 transport file, but we could also use the “V8” and “V9” arguments for this parameter to specify SAS V8 or V9 transport files (but just for non-submission purposes!). Unless you use the macro, you will have two different methods for creating transport files: PROC Copy for V5 and PROC CPORT for all other versions.

Another use of the *%loc2xpt* macro is to check whether a post-V5 feature has crept into your SDTM data sets (this is a hazard of creating SAS V5 data sets in a much more advanced development environment, like SAS V9.3). Admittedly, many version conflicts are reported by the XPORT Libname Engine with ERROR: messages in the SAS Log (for example, if your data set contains a variable name that is longer than 8 characters, the SAS Log will contain an ERROR: message that the .xpt file has not been saved). However, the XPORT Libname Engine does not identify every version conflict with an ERROR: message (for example, if a variable label is longer than the SAS V5 standard of 40 characters, the SAS Log will contain only a NOTE: message).

V5 FEATURE VALIDATION UTILITY

Probably the safest way to verify that a SAS data set in a V5 transport file is free of any post-V5 features would be to assess the data set with a validation utility that:

- creates V5 and V9 transport files for the data set (using the *%loc2xpt* macro)
- converts each transport file into a SAS data set (using the *%xpt2loc* macro)
- compares the re-constituted SAS data sets to verify that they are identical (any post-V5 feature will not have been represented in the V5 transport file, and therefore will not be present in its re-constituted SAS data set)

The SAS Drug Development Forum (https://communities.sas.com/community/support-communities/sas-drug-development) contains a code sample that implements this utility.

CLINICAL STANDARDS TOOLKIT

The SAS Clinical Standards Toolkit (CST) is a set of macros and metadata data sets that help you carry out the tasks of preparing clinical trial data for submission, including validating data sets against CDISC standards and generating define.xml. If you do not yet have the Toolkit, you can get it from SAS as a supplement to your licensed SAS software for no additional cost.

CRT-DDS/define.xml

Case Report Tabulation Data Definition Specification (CRT-DDS) is the formal name for the define.xml, the inventory that accompanies clinical trial data when it is submitted to a regulatory agency. It provides a complete set of information about the submission data sets and their contents. The Toolkit builds the define.xml from the metadata for the submission data sets. For a more complete description of the Toolkit, please refer to *Introduction to SAS Clinical Standards Toolkit* and *Using the SAS Clinical Standards Toolkit 1.4 for define.xml Creation*, both of which are cited in the References section.

There is an intrinsic limitation with having a clinical trial submission contain an inventory (define.xml) that is separate from the data sets that it describes (see Figure 4). In many ways, it is like having a generated table of contents in a word processing document: if you make a change to the document but do not re-build the table of contents, the page references may no longer be accurate. Similarly, if you modify submission data sets after you have created the define.xml, the inventory will no longer be accurate. A further complication is that the data sets themselves are not even part of the submission – their V5 transport files are. Consequently, any late-occurring change to a submission data set also requires that the corresponding V5 transport file be re-created as well.

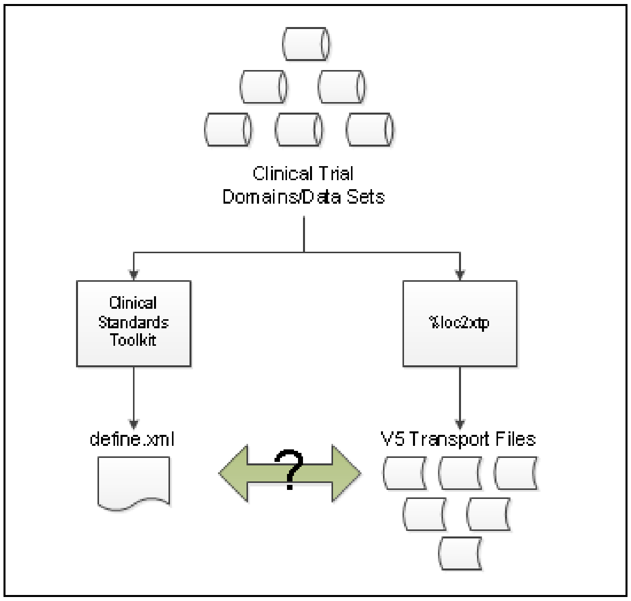

Figure 4. Separate Processes Create define.xml and V5 Transport Files

To verify that the V5 transport files and the define.xml match, we recommend that you execute a reconciliation utility like the one that we describe below.

EXTRACT METADATA FROM V5 TRANSPORT FILES

Follow these steps to create a set of metadata based on the V5 transport files.

1. **Convert the V5 transport files to SAS datasets.** You can code your own conversion utility using the %xpt2loc macro, or adapt existing code like the example that is presented in *Sample 33918: Read and convert multiple transport (XPORT) files to SAS data sets*, available at http://support.sas.com/kb/33/918.html.
2. **Run *create_sourcemetadata.sas*.** This program creates a set of SAS data sets that represent the SDTM metadata:
    - source_tables
    - source_columns
    - source_study
3. **Run *create_crtdds_from_sdtm.sas*.** This program creates the 39 data sets that comprise the SAS representation of the CRT-DDS data model.
4. **Run *create_formatsfromcrtdds.sas*.** The program generates the controlled terminology SAS formats catalog. The output can also be captured as the source_format SAS dataset.

EXTRACT METADATA FROM define.xml

Follow these steps to create a set of metadata based on the define.xml.

1. **Run *create_sascrtdds_fromxml.sas*.** This program reads the define.xml file and creates the 39 data sets that comprise the SAS representation of the CRT-DDS data model.
2. **Run *create_sourcemetadata.sas*.** This program creates a set of SAS data sets that represent the SDTM metadata:
    - source_tables
    - source_columns
    - source_study
3. **Run *create_formatsfromcrtdds.sas*.** The program generates the controlled terminology SAS formats catalog. The output can also be captured as the source_format SAS dataset.

COMPARE METADATA TABLES

At this point, we have two sets of the four metadata tables (source_study, source_tables, source_columns and source_format): one set that has been extracted from the V5 transport files and one set from the define.xml. We can use PROC Compare to compare each metadata table against the corresponding table in the other set to identify any discrepancies.

The V5 Feature Validation Utility, which is available on the SAS Drug Development Forum, can be adapted to perform the comparisons.

CONCLUSION

Our goal in this presentation has been to describe some of the utilities that we have used to assess clinical trial submission data. If you do not have these types of utilities already in place, we hope that you can use some of our ideas and code samples to create your own utilities. If you already have utilities like these in place, we hope that we have given you some ideas that you can use to evaluate them and either confirm that they the optimal ones for your work practices or give you some points to consider for enhancements.

REFERENCES

- FDA. Study Data Specifications, July 18, 2012. Available at http://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM312964.pdf, 23JAN2013.
- Jansen, Lex. May 2012. “Using the SAS Clinical Standards Toolkit 1.4 for define.xml Creation.” *Proceedings of PharmaSUG 2012*. Available at http://www.lexjansen.com/pharmasug/2012/HW/PharmaSUG-2012-HW02-SAS.pdf, 27MAR2013.
- Mangold, Andreas and Nichole Wächter. October 2010. “Introduction to SAS Clinical Standards Toolkit.” Available at http://www.phusewiki.org/docs/2010/2010%20PAPERS/TU06%20Paper.pdf, 27MAR2013.

- Roediger, Frank. May 2012. "Creating Analysis Data Marts from SDTM Warehouses." *Proceedings of PharmaSUG 2012*. Available at http://www.lexjansen.com/pharmasug/2012/DS/PharmaSUG-2012-DS21-SAS.pdf, 25MAR2013.
- Sample 33918: Read and convert multiple transport (XPORT) files to SAS data sets. Available at http://support.sas.com/kb/33/918.html, 31MAR2013.
- SAS(R) 9.2 Language Reference: Dictionary, Fourth Edition, "VARLENCHK= System Option." Available at http://support.sas.com/documentation/cdl/en/lrdict/64316/HTML/default/viewer.htm#a003269301.htm, 31JAN2013.
- Study Data Tabulation Model Implementation Guide: Human Clinical Trials, Version 3.1.3, Final, July 16, 2012. Available in sdtmigv3.1.3.zip at http://www.cdisc.org/sdtm, 01FEB2013.
- Usage Note 46944: New SAS transport format and tools available. Available at http://support.sas.com/kb/46/944.html, 30MAR2013.

ACKNOWLEDGMENTS

We would like to acknowledge Beth Hutchins, Robin MacBain, Jody Graham, Kim Wilson, Ben Bocchicchio, and Lex Jansen, our colleagues at SAS, for their help and support.

CONTACT INFORMATION

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

Name:	Frank Roediger
Enterprise:	SAS Institute, Inc.
Address:	720 SAS Campus Drive, U3102
City, State ZIP:	Cary, NC 27513
Work Phone:	919 531-0519
E-mail:	frank.roediger@sas.com

Name:	Sandeep Juneja
Enterprise:	SAS Institute, Inc.
Address:	720 SAS Campus Drive, U3117
City, State ZIP:	Cary, NC 27513
Work Phone:	919 531-0541
E-mail:	sandeep.juneja@sas.com