PharmaSUG 2016 - Paper PO05

A Data Preparation Primer: Getting Your Data Ready for Submission

Janet Stuelpner, SAS Institute, New Canaan, CT

ABSTRACT

Does your data conform to industry standards? Is it in the format that is necessary for submission to the regulatory authorities? Beginning next year, the FDA will require that submission data adhere to CDISC industry standards. All data will need to be in SDTM and ADaM standard format. Above and beyond that, there will be requirements to submit your data in very specific formats. What do you need to do to make sure that your data is in the correct format? The first thing to do is create procedures to know what to do and where to start. This presentation will point out some of the requirements and make suggestions as to the things that you need to do to have a better chance to have your submission accepted and not sent back.

INTRODUCTION

Data isn't data until it can be changed into valuable information. There are many tools on the market where you can check your data for compliance to CDISC standards (SEND, SDTM, ADaM). However, is that enough? The tools that are available will check a whole bunch of things like: domain details, data values, consistency across domains, date fields, column attributes, column values, controlled terminology. These are all good, but it does miss a number of things. Is the field that character values are in too long? Are there fields that are all missing values? Are the death records summarized and consistent across all domains? Is the order of the variables in the datasets consistent with the domain templates provided by CDISC? We are going to explore some factors that could be questioned by regulatory authorities that can easily be reviewed by you before you create your transport files for submission.

ORDER OF THE VARIABLES

The order of the variables in any of the CDISC datasets is determined in the CDISC domain templates. To make data review easier, the order of the variables are specified in the domains template definitions. The order of variables in the define.xml must reflect the order of variables in the dataset. The order of variables in the CDISC domain models has been chosen to facilitate the review of the models and application of the models. Variables for the three general observation classes must be ordered with Identifiers first, followed by the Topic, Qualifier, and Timing variables.

CHECKING THE DATA VALUES

Of course, we want the data values to be correct. We perform edit checks to clean the data and send queries back to the investigators. The data must be correct and of good quality. High quality data is the key to making better decisions. There are many different items to check before it is sent off in the submission. Some of these include: identifying missing values, correct coding, identifying duplicate data, using standard units, applying controlled terminology correctly, summarizing death information and making sure that information is consistent across all domains.

CHECKING THE VARIABLE LENGTH

Data can get quite large. If character variables are set to a very large length and the data isn't big enough for those fields, it takes up a great deal of storage space. Data managers and programmers will create a field that is 200 characters long to make sure that all of the data will fit in the variable. But, at the end of the study, what if the fields (e.g., AETERM, AEBODSYS, CETERM, CEDECOD, CMDECOD, DSTERM, EXDOSTXT, MHTERM, MHDECOD) contain much shorter values than were originally expected? Should you include all of that data? The best practice would be to trim the fields. If you search for the longest data value that are contained within the field and then reset the length so that it is no longer than that value, you will be saving a great deal of space.

V	A	S	O	S	P	A	S	M											
R	A	S	H																
P	N	E	U	M	O	N	I	A											
F	E	V	E	R															
S	E	P	S	I	S														

As you can see in this example, the field AETERM is set up to contain 20 characters. The largest data value is 9. So, the length of the field should be 9 and not 20. If this is done for every character field, it would be a much smaller amount of space that the submission would contain

CORE VARIABLE

According to the Study Data Tabulation Model Implementation Guide: Human Clinical Trials Version 3.2 the concept of a core variable is used both as a measure of compliance, and to provide general guidance to sponsors. Three categories of variables are specified in the "Core" column in the domain models:

- **Required:** This type of variable is any variable that is basic to the identification of a data record (i.e., essential key variables and a topic variable) or is necessary to make the record meaningful. Required variables must always be included in the dataset and cannot be null for any record.

- **Expected:** This variable is any variable necessary to make a record useful in the context of a specific domain. Expected variables may contain some null values, but in most cases will not contain null values for every record. When no data has been collected for an expected variable, however, a null column must still be included in the dataset, and a comment must be included in the define.xml to state that data was not collected.

- **Permissible:** This core variable should be used in a domain as appropriate when collected or derived. Except where restricted by specific domain assumptions, any SDTM Timing and Identifier variables, and any Qualifier variables from the same general observation class are permissible for use in a domain based on that general observation class. The Sponsor can decide whether a Permissible variable should be included as a column when all values for that variable are null. The sponsor does not have the discretion to exclude permissible variables when they contain data.

Now that you have defined all of the variables in the study, you know that you must include all of the identifiers because they are required. These fields must not contain any null values. As specified earlier, if the values in the data are much shorter than the length of the field as it was defined, you can shorten the length to save space. Another area where you must include the variables are the Expected variables. To make the record meaningful, the variables must be there, but they can have null values. Again, the length of character variables should be checked to see if these can be reduced in size. Permissible variables can be null just like the Expected ones. If a variable contains all null values you have the discretion to keep them or delete them. You need to make sure that if values exist in a variable that is permissible, then those variables must be kept.

CONCLUSION

The data can be reviewed and reviewed and then reviewed again. There can be computer run edit checks and human review of the data. In all of the various methods that we use to make sure that the data we send to the regulatory authorities is correct, we need to make sure that the data is received in the manner that is acceptable for regulatory review. Make sure that you know those regulations. Check to see the format that is necessary at the moment in time that you are preparing your submission. You want to make sure that your drug is approved the first time that you submit. The intent of this paper is to give you some things to think about as you clean, query, analyze and reports on the data in your submissions.

REFERENCES

Bocchicchio, Ben and Roediger, Frank, 2015, "Getting Rid of Bloated Data in FDA Submissions", Proceedings of PharmaSUG 2015 Conference, Orlando, FL

Rosario, 2015, Lilliam, “JumpStarting the Regulatory Review Process: The Review Perspective”, Keynote Address at PharmaSUG 2015 Conference, Orlando, FL

http://www.fda.gov/Drugs/DevelopmentApprovalProcess/FormsSubmissionRequirements/ElectronicSubmissions/ucm248635.htm, Study Data Standards for Submission to CDER

http://www.cdisc.org/fda-guidance-on-standardized-study-data-, FDA Guidance on Standardized Study Data for Electronic Submissions FDA Guidance on Standardized Study Data for Electronic Submissions, February 6, 2014

Gaffney, Alexander. “With PDUFA VI Negotiation Process Fast Approaching, BIO Takes Critical Look at Regulations.” Regulatory Affairs Professional Society. August 6, 2014.
Available at http://www.raps.org/Regulatory-Focus/News/2014/08/06/19967/With-PDUFA-VI-Negotiation-Process-Fast-Approaching-BIO-Takes-Critical-Look-at-Regulations/

CDISC Available at http://www.cdisc.org

OpenCDISC. Available at http://www.opencdisc.org

SAS Drug Development Forum. Available at https://communities.sas.com/community/support-communities/sas-drug-development

U.S. Department of Health and Human Services, Food and Drug Administration, *Study Data Technical Conformance Guide*: *Technical Specifications Document*. December 2014. Available at http://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm

Wilson, Todd Allen. “Inside Health Policy - Industry Looks At 21st Century Cures To Set Stage For PDUFA VI.” Friends of Cancer Research. December 19, 2014. Available at http://www.focr.org/news/inside-health-policy-industry-looks-21st-century-cures-set-stage-pdufa-vi

ACKNOWLEDGMENTS

This paper would not have been written if it hadn’t been for the support given to me by my husband, Robert Stuelpner. Bob diligently read this paper in order to correct obvious errors and keep me on the right track. His criticisms were constructive and his support never ending.

CONTACT INFORMATION

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

Name:	Janet Stuelpner
Enterprise:	SAS Institute
Address:	100 Cary Parkway
City, State ZIP:	Cary, NC 27513
Work Phone:	919-531-9758
E-mail	janet.stuelpner@sas.com:
Web:	www.sas.com

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.