PharmaSUG China 2015 - Paper 27

SDTM Electronic Submissions to FDA: Guidelines and Best Practices

Christina Chang, PAREXEL International, Taipei, Taiwan
Kyle Chang, PAREXEL International, Taipei, Taiwan

ABSTRACT

Electronic data submission is the future of clinical trials. United States Food and Drug Administration (FDA) released several submission guidance documents since last year. The guidance of "Study Data Technical Conformance Guide" provides specifications, recommendations, and general considerations on how to submit standardized study data using FDA-supported data standards. It was developed in an effort to combine the existing Common Issues, Study Data Specifications and Traceability Guidance documents, as well as Validation Rules, in order to offer one technical document that coordinates all these sources for the industry. This will reduce the likelihood of the FDA requesting data to be represented in a manner that contradicts CDISC rules. It also provides technical recommendations to sponsors for the submission of study data and related information in a standardized electronic format.

This paper elaborates on the following fundamental and core components to be considered for FDA submissions: Study Data Submission Format, Terminology, Electronic Submission Format, Data Validation and Traceability.

INTRODUCTION

In a regulated industry such as pharmaceutical and biotechnology industry, FDA have several submission guidance documents of electronic data submission for study data tabulation model (SDTM), Analysis Data Model (ADaM) data, standard for exchange of nonclinical data (SEND). An electronic data submission followed FDA standard requirements can help the reviewers to navigate submission documents and datasets, and then understand the relationship between submission report and datasets. In the industry, every effort is made by sponsors to reduce the review time of data submission. Generating electronic data submission which applied FDA requirement may ease the review, hence may reduce the review time. This paper will focus on the electronic submission for SDTM and will take examples.

STUDY DATA SUBMISSION FORMAT

Clinical Data Interchange Standards Consortium (CDISC) is a nonprofit standards development organization (SDO) that has been working to develop global data standards for clinical and nonclinical research. Study Data Tabulation Model (SDTM) defines a standard structure for human clinical study data tabulations and for nonclinical study data tabulations that are to be submitted as part of a product application to a regulatory authority such as FDA.

SDTM GENERAL CONSIDERATIONS

The Study Data Tabulation Model Implementation Guide (SDTMIG) should be followed. Here, we highlight noteworthy aspects when preparation submission. Variables in the SDTM dataset classifies as required, expected, or permissible. The length of variable names, descriptive labels, and dataset labels should not exceed the maximum permissible number of characters described below. Variable and dataset names should not contain punctuation, dashes, spaces, other non-alphanumeric symbols, or special characters. Variable and dataset labels can include punctuation characters, but still should not contain special characters. This is to avoid possible incompatibility with SAS V5 Transport files.

Table 1. Maximum Length of Variables and Dataset Elements

Element	Maximum Length in Characters
Variable Name	8
Variable Descriptive Label	40
Dataset Label	40

The value of following variables should be no more than the maximum characters in length which also defined in FDA SDTM validation rules v1.0.

Table 2. Maximum Length of Variables & FDA Rules

FDA Rule ID	SDTM Variable	Maximum Length in Characters
FDAC057	--TEST	40
FDAC059	--PARM	40

FDA Rule ID	SDTM Variable	Maximum Length in Characters
FDAC060	--PARMCD	8
FDAC067	ARMCD	20
FDAC070	ETCD	8
FDAC198	ACTARMCD	20

Other than the basic limitation above, the length of the variable should be set to the maximum length of the variable used across all datasets in the study. Datasets should be split into smaller datasets no larger than 1 gigabyte (gb). The SDTMIG also requires dates and times of day to be stored according to the international standard ISO 8601.

The following are examples of some of the Permissible and Expected variables in SDTM and SEND that should be included, if available:

Baseline flags (--BLFL): Baseline flags should be submitted or derived in all finding domain, such as LB or EG domain.

Epoch (EPOCH): As part of the design of a trial, the planned period of subjects' participation in the trial is divided into Epochs. Each Epoch is a period of time that serves a purpose in the trial as a whole.

Date variable and study day (--DTC, --STDTC, --ENDTC, --DY, --STDY and --ENDY): When the date/time of collection is reported in any domain, the date/time should go into the --DTC field (e.g., EGDTC for Date/Time of ECG). Whenever --DTC, --STDTC or –ENDTC are included, the matching Study Day variables (--DY, --STDY, or --ENDY, respectively) should be included. For example, in most Findings domains, --DTC is Expected, which means that --DY should also be included.

DATA DEFINITION FILE

A data definition file, formally called Case Report Tabulation Data Definitions (CRT DD), is necessary to facilitate the review of the study data submitted to a regulatory authority. The sponsor needs to provide complete details in this file, especially for the derived variables and make certain that the code list and origin for each variable are clearly and easily accessible from the define file.

The define file should be submitted in XML format, i.e., a properly functioning define.xml. Creating define.xml is difficult especially if you don't have any knowledge about XML at the beginning. However, there are several and great papers presented in PharmaSUG, which using the SAS based solution for define.xml. The in-house SAS based solution is more flexible rather than doing it manually.

In addition to the define.xml, a printable define.pdf should be provided if the define.xml cannot be printed. Creating define.pdf can use the attached XSL file to render the xml file to pdf via Apache FOP, a free open source software. The file can convert compliant define.xml file define.pdf. The define.pdf looks identical to define.xml (when viewed using XSL stylesheet from CDISC) and includes the internal/external links & bookmarks.

ANNOTATED CASE REPORT FORM

An annotated CRF should reflect the data that are expected to be submitted within the SDTM. Annotated CRF should include and annotate unique forms. For annotated in the entire CRF, only the first occurrence should be annotated. Annotated CRF should include bookmark. There are two ways of bookmark (dual bookmarking): bookmarks by time-points and bookmarks by CRF topics or forms. Table of content (TOC) is not required for annotated CRF, but to improved navigation for reviewers, the document must have a TOC if the document is 10 pages or more.

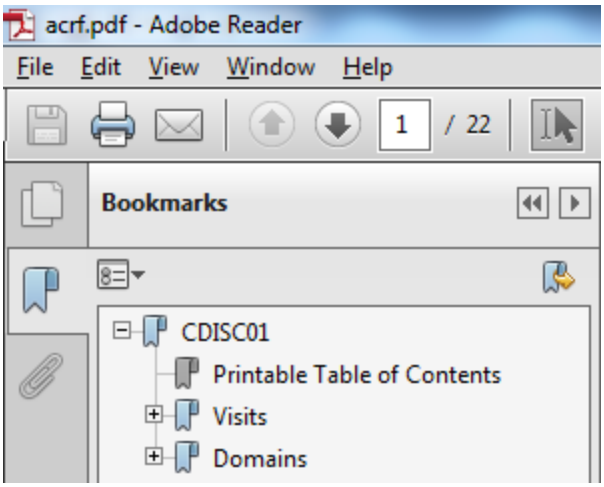

Figure 1. Dual bookmarked SDTM aCRF

STUDY DATA REVIEWER'S GUIDE

The Study Data Reviewer's Guide (SDRG) provides information and directions for FDA reviewers. The SDRG has four main sections and two optional appendices - Introduction, Protocol Description, Subject Data Descriptions, Data Conformance Summary, Appendix I: Inclusion/Exclusion Criteria, and Appendix II: Conformance Issues Details. The SDRG purposefully duplicates information found in other submission documentation (e.g. the protocol, clinical study report, define.xml, etc.) in order to provide FDA Reviewers with a single point of orientation to the SDTM datasets.

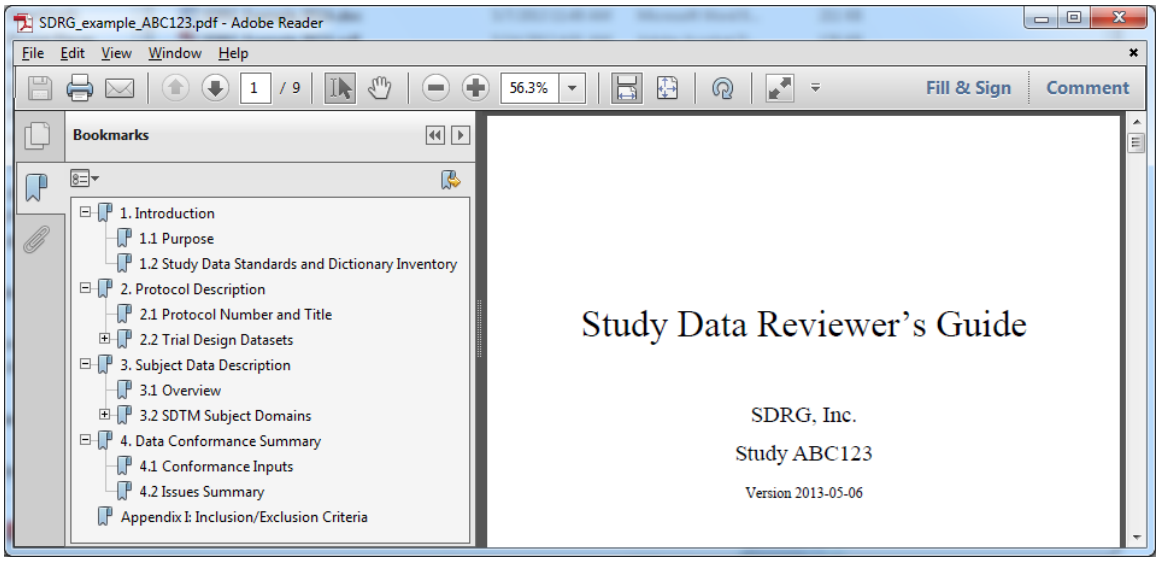

Figure 2. Sample Study Data Reviewer's Guide

TERMINOLOGY

A major problem is the wide variety of terms used to express similar or identical concepts. Such inconsistency makes it nearly impossible to integrate, aggregate, and manage even modest-sized datasets from various sources to answer clinical and research questions.

For example, when submitting datasets containing clinical laboratory data, the variability in the possible representation of unit characters can be equally as limiting with respect to standardization. Interchangeable use of Greek letter symbols with short codes is one source of inconsistency in test unit. The unit 'micro' can be represented as 'µ', 'u' or 'mc'. Inconsistent use of capitalization in units is another cause of inconsistency and possible error.

Code	Codelist Code	Codelist Extensible (Yes/No)	Codelist Name	CDISC Submission Value	CDISC Synonym(s)	CDISC Definition	NCI Preferred Term
C67396	C71620		Unit	ug/kg	Microgram per Kilogram; mcg/kg; ng/g; pg/mg; ug/kg	A unit of a mass fraction expressed as a number of micrograms of substance per kilogram of mixture. The unit is also used as a dose calculation unit.(NCI)	Microgram per Kilogram
C73729	C71620		Unit	ug/kg/day	Microgram per Kilogram per Day	A dose calculation unit expressed in microgram(s) per kilogram per period of time equal to twenty-four hours. (NCI)	Microgram per Kilogram per Day
C73730	C71620		Unit	ug/kg/h	Microgram per Kilogram per Hour	A dose calculation unit expressed in microgram(s) per kilogram per period of time equal to sixty minutes. (NCI)	Microgram per Kilogram per Hour
C71210	C71620		Unit	ug/kg/min	Gamma per Kilogram per Minute; Microgram per Kilogram per Minute; gamma/kg/min; mcg/kg/min	A dose calculation unit equal to one millionth of a gram of a preparation per one kilogram of body mass administered per unit of time equal to one minute.(NCI)	Microgram per Kilogram per Minute
C67306	C71620		Unit	ug/L	Microgram per Liter; Milligram per Cubic Meter; Nanogram per Milliliter; mcg/L; mg/m3; ng/mL; ug/L	A unit of concentration or mass density equal to one nanogram of substance per milliliter of solution or one microgram of substance per liter of solution.	Microgram per Liter

Figure 3. CDISC Controlled Terminology for Units

Controlled terminology standards are an important component of study data standardization. The analysis of study data is greatly facilitated by the use of controlled terms for clinical or scientific concepts that have standard, predefined meanings and representations. It's also useful when consistently applied across studies to facilitate integrated analyses. Sponsors should specify the terminologies and versions used in the study in the SDRG and define.xml.

External Dictionaries

Reference Name	External Dictionary	Dictionary Version
Adverse Event Dictionary (CL.AEDICT_F)	MEDDRA	17.1
Drug Dictionary (CL.DRUGDICT_F)	WHODRUG	201403
ISO3166 (CL.ISO3166)	ISO3166	

Figure 4. External Dictionaries in define.xml

ELECTRONIC SUBMISSION STRUCTURE

Study datasets and their supportive files should be organized into a specific file directory structure when submitted in the eCTD format. The submitted data can be classified into four types: 1) analysis datasets, 2) data tabulations, 3) miscellaneous datasets, and 4) subject profiles. The specification for organizing datasets and their associated files in folders within the submission is summarized in the following figure.

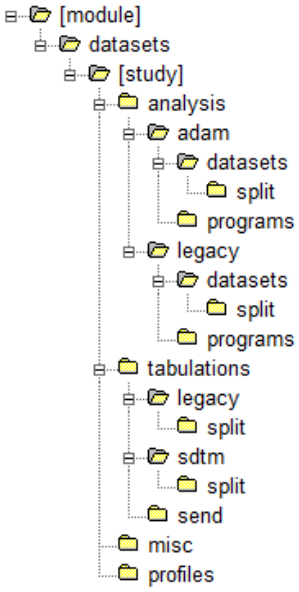

Figure 5. Electronic Submission Folder Structure

The define.xml and supportive style sheet should reside in the same folder along with the submission datasets. The bookmarked and annotated CRF from the study should be saved in a PDF named acrf.pdf and stored in the "sdtm" folder. All unique CRF pages or forms should be annotated to match the SDTM datasets and variables. The reviewers' guide and complex algorithms which provides additional information for the reviewers about the submitted data should be stored in the folder as well.

In addition, datasets greater than 1 gigabyte (gb) in size should be split into smaller datasets no larger than 1 gb. There is a new rule in the Study Data Technical Conformance Guide. Sponsors should submit the smaller split files in the "split" sub-folder in addition to the larger non-split file in the original data folder.

DATA VALIDATION AND TRACEABILITY

STUDY DATA VALIDATION

Data validation is a process to ensure that submitted data are both compliant and intended use. Sometimes serious issues in the submitted data are only evident through manual inspection of the data and may only become evident once the review is well under way.

FDA recognizes two types of validation rules, conformance validation and quality checks. These rules help ensure that the data conform to the standards, while quality checks help ensure the data support meaningful analysis. Last year (13-Nov-2014), FDA published its first official list of validation rules for CDISC SDTM. These long awaited rules cover both conformance and quality requirements, as described in the FDA Study Data Technical Conformance Guide. To bring the validation process forward in the clinical data life cycle, the ultimate purpose of this validation tool is to check that domains are submission-ready; however any versatility in the tool could significantly enhance the efficiency of production of the final domains. A number of approaches can be taken for validating the SDTM data. Other than SAS® PROC CDISC, there are two ways to validate the SDTM data efficiently:

OpenCDISC Community: Fortunately, OpenCDISC have implemented the new FDA validation rules in their validator, OpenCDISC Community 2.0. It upgraded with FDA validation rules and ability to validate against study specific value level metadata.

OpenCDISC ID	Publisher ID	Message	Description	Category	Severity
CT2001	FDAC340	-- value not found in '--' non-extensible codelist	-- (--) variable must be populated with terms from '--' CDISC controlled terminology codelist. New terms cannot be added into non-extensible codelists.	Terminology	Error
CT2002	FDAC341	-- value not found in '--' extensible codelist	-- (--) variable should be populated with terms from '--' CDISC controlled terminology codelist. New terms can be added as long as they are not duplicates, synonyms or subsets of existing standard terms.	Terminology	Warning
CT2003	FDAC342	-- and -- values do not have the same Code in CDISC CT	-- and -- must be populated using terms with the same Codelist Code value in CDISC control terminology. There is one-to-one relationship between -- and -- values defined in CDISC control terminology by Codelist Code value.	Terminology	Error
CT2004	FDAC343	-- value not found in '--' non-extensible codelist	-- (--) variable must be populated with terms from '--' CDISC controlled terminology codelist, when --. New terms cannot be added into non-extensible codelists.	Terminology	Error
CT2005	FDAC344	-- value not found in '--' extensible codelist	-- (--) variable should be populated with terms from '--' CDISC controlled terminology codelist, when --. New terms can be added as long as they are not duplicates, synonyms or subsets of existing standard terms.	Terminology	Warning
CT2006	FDAC345	-- and -- values do not have the same Code in CDISC CT	-- and -- must be populated using terms with the same Codelist Code value in CDISC control terminology. There is one-to-one relationship between -- and -- values defined in CDISC control terminology by Codelist Code value within the same --.	Terminology	Error
SD0001	FDAC014	No records in data source	Domain table should have at least one record	Presence	Error
SD0002	FDAC018	NULL value in -- variable marked as Required	Required variables (where Core attribute is 'Req') cannot be NULL for any records	Presence	Error
SD0003	FDAC038	Invalid ISO 8601 value for -- variable	Value of Dates/Time variables (*DTC) must conform to the ISO 8601 international standard	Format	Error
SD0004	FDAC056	Inconsistent value for DOMAIN	Domain Abbreviation (DOMAIN) variable should be consistent with the name of the dataset	Consistency	Error
SD0005	FDAC044	Duplicate value for --SEQ variable	The value of Sequence Number (--SEQ) variable must be unique for each record within a domain and within each Unique Subject Identifier (USUBJID) or Pool Identifier (POOLID) variables value when they are present in the domain.	Consistency	Error

Figure 6. Validation Rules (OpenCDISC & FDA) in OpenCDISC Report

SAS Macro Based Solution: The in-house SAS based solution includes a set of SAS macros that checks each SDTM domain for compliance with the latest SDTM/SDTM IG. Using this method could customize the comparison between the metadata information obtained from the SDTM mapping specification or CDISC SDTM metadata versus the SDTM datasets, especially for sponsor custom domains. In the previous version of OpenCDISC validator, it couldn't validate SDTM datasets against study specific value level metadata.

Sponsors should validate their study data before submission using the published validation rules and either correct any validation errors or explain in the SDRG why certain validation errors could not be corrected. The recommended pre-submission validation step is intended to minimize the presence of validation errors at the time of submission.

STUDY DATA TRACEABILITY

Another important component of a regulatory review is the traceability of the sponsor's results back to the CRF data. It's an understanding of the relationships between the analysis results, analysis datasets, tabulation datasets, and source data. Therefore, establishing traceability is one of the most problematic issues associated with legacy study data converted to standardized data.

Here is a recommendation for data traceability within a sponsor/submission. The --SPID variable (Sponsor-Defined Identifier) is included in all SDTM general observation classes (Findings, Interventions and Events). To have study data traceability, we could add the row number on the data collection form or original source file name to the --SPID variable whenever data are collected on a CRF or electronically submitted.

LBSPID	LBTESTCD	LBTEST
LB-0001-00113-B01-26107-56-BASO	BASOLE	Basophils/Leukocytes
LB-0001-00113-C01-26141-56-BASO	BASOLE	Basophils/Leukocytes
LB-0001-00113-C02-26173-56-BASO	BASOLE	Basophils/Leukocytes
LB-0001-00113-C03-26205-56-BASO	BASOLE	Basophils/Leukocytes

Figure 7. LBSPID for Data Tracebility

CONCLUSION

Generating the electronic data submission followed FDA released submission guidance documents can reduce the review time and may even advance the time for approval. Thus, it is important for both sponsors and CROs' to understand the standard of electronic data submission. This paper includes the authors' experience and real cases of electronic data submission which can provide a comprehensive view for sponsors and users. However, the latest version of submission guidance documents is still updated by FDA frequently. Even some of documents or rules are only recommended in current version, but it may be changed in the latest version. Sponsors and users will need to check the electronic submission package followed the latest version of submission guidance documents before submission.

REFERENCES

U.S. Center for Drug Evaluation and Research (CDER). "CDER Common Data Standards Issues Document (Version 1.1)." December 2011. Available at http://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/FormsSubmissionRequirements/ElectronicSubmissions/UCM254113.pdf

U.S. Food and Drug Administration, "Study Data Technical Conformance Guide v2.1." March 2015. Available at http://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM384744.pdf

Clinical Data Interchange Standards Consortium. "Study Data Tabulation Model Metadata Submission Guidelines (SDTM-MSG)." December 2011. Available at http://www.cdisc.org/content3402

U.S. Food and Drug Administration, "Providing Regulatory Submissions in Electronic Format — Standardized Study Data", December 2014. Available at http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM292334

U.S. Food and Drug Administration, "Providing Regulatory Submissions in Electronic Format — Submissions Under Section 745A(a) of the Federal Food, Drug, and Cosmetic Act", December 2014. Available at http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM384686.pdf

Chang, Kyle. "An implementation of XSL-FO techniques to convert define.pdf from define.xml." PharmaSUG China 2014.

CONTACT INFORMATION

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

Christina Chang
PAREXEL International
22F, Far Glory International Center, No. 200,
Sec. 1, Keelung Road, Taipei, Taiwan 11071, ROC
Christina.Chang@PAREXEL.com
http://www.PAREXEL.com/

Kyle Chang
PAREXEL International
22F, Far Glory International Center, No. 200,
Sec. 1, Keelung Road, Taipei, Taiwan 11071, ROC
Kyle.Chang@PAREXEL.com
http://www.PAREXEL.com/