PharmaSUG 2017 - Paper SS12

Quality Check your CDISC Data Submission Folder Before It Is Too Late!

Bhavin Busa, Vita Data Sciences (a division of Softworld, Inc.), Waltham, MA

ABSTRACT

The standardized clinical study datasets will be required in submissions for clinical and non-clinical studies that start on or after December 17, 2016. FDA has added a technical rejection criteria to the existing eCTD validation criteria to enforce the deadlines. The FDA may refuse to file for NDAs, an electronic submission that does not have study data in conformance to the required standards specified in the FDA Data Standards Catalog. This means that all studies going forward must utilize CDISC SDTM/ADaM standards and should consist of associated submission documents (aCRF.pdf, define.xml/define.pdf, cSDRG.pdf, ADRG.pdf) per the Study Data Technical Conformance Guide (TCG). The submission of these study datasets and documents should be organized into a specific file directory structure per the eCTD requirements.

As Sponsor is preparing for their NDA submission, it is critical for them to verify the content and validity of the dataset folder per the FDA submission requirements, i.e. that the datasets meet the technical specifications per the Study Data TCG and eCTD validation criteria. In addition, it is in their best interest to check whether the datasets that are provided for regulatory publishing is truly the 'final' version.

In this paper, we will provide an overview of a SAS®-based tool to perform a final quality check on your CDISC data submission package ('m5' folder) that incorporate checks per the Study Data TCG and eCTD validation criteria's which are not typically covered by either an existing CDISC datasets compliance tools (e.g. Pinnacle 21®) or by commercially available eCTD publishing software.

INTRODUCTION

The CDISC data standards (SEND, SDTM, ADaM, Define-XML, and more) provides a way to exchange clinical and nonclinical research data across the Sponsor and the regulatory agency in an electronic format. The standardization of submission study datasets greatly facilitates the FDA's ability to explore, process, review and archive submission data more efficiently and effectively. The FDA binding guidance [1] describes the requirements that studies are compliant with the standards outlined in the FDA Data Standards Catalog (DSC).

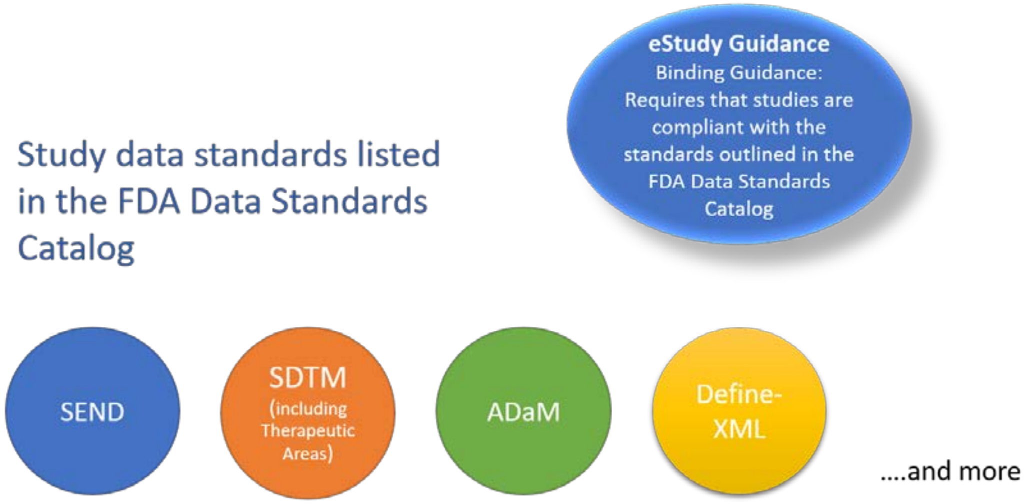

Study data standards listed in the FDA Data Standards Catalog
eStudy Guidance
Binding Guidance: Requires that studies are compliant with the standards outlined in the FDA Data Standards Catalog
SEND
SDTM (including Therapeutic Areas)
ADaM
Define-XML
....and more

Figure 1. Key Standards Outlined in the FDA Data Standards Catalog [2]

The standardized clinical study datasets will be required in submissions for clinical and non-clinical studies that start after December 17, 2016 [3]. FDA has added a technical rejection criteria to the existing eCTD validation criteria to enforce the deadlines. As noted, it will be expected that all the trials conducted after that date must use study data standards that are listed in the FDA DSC [3]. The FDA may Refuse to File (RTF) for NDAs and BLAs or Refuse to Receive (RTR) for ANDAs, an electronic submission that does not have study data in conformance to the required standards specified in the catalog [3].

This means that all studies going forward must utilize CDISC SDTM and ADaM standards for their tabulation and analysis datasets respectively and should consist of data definition file (define.xml) to describe the metadata of the submitted electronic datasets along with associated submission documents such as annotated CRF (aCRF.pdf), Clinical Study Data Reviewers Guide (cSDRG.pdf), and Analysis Data Reviewers Guide (ADRG.pdf) per the Study Data TCG [4].

As Sponsor is preparing for their NDA submission, it is critical for them to verify the content and validity of the dataset folder per the FDA submission requirements, i.e. that the datasets meet the technical specifications per the Study Data TCG and eCTD validation criteria. In addition, it is in their best interest to check whether the datasets that are provided for regulatory publishing is truly the ‘final’ version.

In the sections below, we have provided an overview of the recent FDA technical rejection process to enforce study data standards along with the summary of key items from the technical conformance guide. We have provided details about why there is a need to perform a final quality check on your CDISC data submission package (‘m5’ folder) that incorporate checks per the Study Data TCG and eCTD validation criteria’s which are not typically covered by either an existing CDISC datasets compliance tools (e.g. Pinnacle 21 Validator) or by commercially available eCTD publishing software.

TECHNICAL REJECTION PROCESS TO ENFORCE STUDY DATA STANDARDS

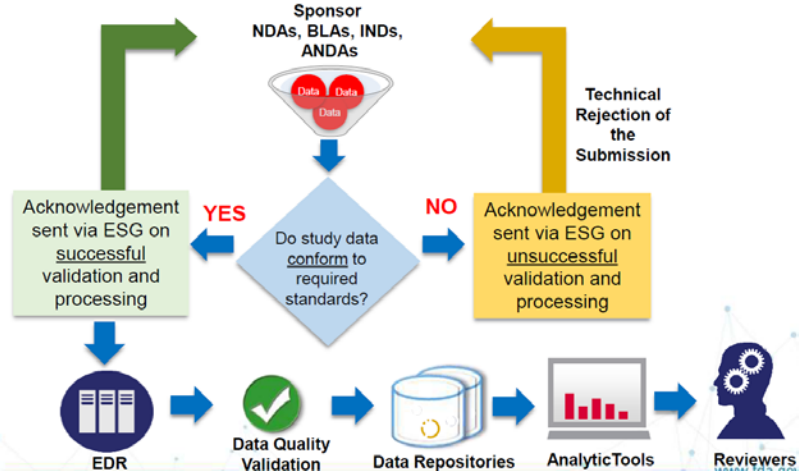

Sponsor
NDAs, BLAs, INDs, ANDAs
Data
Data
Data
Do study data conform to required standards?
YES
NO
Acknowledgement sent via ESG on successful validation and processing
Acknowledgement sent via ESG on unsuccessful validation and processing
Technical Rejection of the Submission
EDR
Data Quality Validation
Data Repositories
AnalyticTools
Reviewers

Figure 2. Study Data Standards Validation & Conformance via a Technical Rejection Step [5]

The above schematic presented by the FDA in the CDER SBIA Webinar Series [5] demonstrates an extra Technical Rejection step which they will run on the Sponsor submitted data as part of their NDA, BLAs, INDs, or ANDAs. This step is to ensure study data conform to required standards, i.e. at the very least the study (clinical and non-clinical) that started after Dec 17, 2016 is using one of the listed data standards for tabulation and analysis datasets from the Standards Catalog.

In addition to the dataset standards, it is also expected that the study data definition file included in the package and the analysis program files also meets the exchange standard as set forth in the data standards catalog. There is also expectation that the Sponsor follows standard terminology code sets listed in the catalog.

Note: There is a clear expectation set in this document and reiterated by the FDA reviewers at various forum (e.g. CDER SBIA webinar series) and meetings (e.g. PhUSE US CSS) that the submission of standardized data using any standard not listed in the catalog, a Sponsor should discuss this with the Agency in advance. The FDA Data Standards Catalog v4.5.1 (08-31-2016) which was current at the time of writing this paper (April 2017) is provided in the reference section below [2].

HOW DOES THE FDA CHECK FOR CONFORMANCE TO DATA STANDARDS?

In order to check whether your study started after December 17,2016, the FDA is requiring each study to have Trial Summary (TS) dataset [3, 5]. All TDM datasets should be included in the submissions and Trial Summary (TS) dataset will be used to determine the time of study start. The expectation is that the TS domain must be present for all the studies included in the submission (i.e. new or legacy). The following parameter must be present in TS domain for both clinical and non-clinical studies:

- Clinical (SDTM and legacy): TSPARMCD = SSTDTC and TSVAL= “yyyy-mm-dd” (ISO8601)
- Non-clinical (SEND and legacy): TSPARMCD = STSTDTC and TSVAL= “yyyy-mm-dd” (ISO8601)

Note: A Trial Summary dataset (ts.xpt) must be presented for each study even if the study started prior to December 17, 2016. Non-clinical legacy data submitted in PDF format should be submitted with a TS dataset.

The technical rejection criteria to be used to assess conformance are being added to the existing eCTD validation criteria to enforce the deadlines. These checks are:

- HIGH severity: Demographic dataset (DM) and the define.xml must be submitted in Module 4 for nonclinical data; DM dataset, the Subject level analysis dataset (ADSL) and define.xml must be submitted in Module 5 for clinical data [eCTD check number: 1736]
- HIGH severity: Trial Summary (TS) dataset must be presented for each study in Module 4 or 5 [eCTD check number: 1734]
- MEDIUM severity: Correct STF file-tags must be used for all standardized datasets (data-tabulations-dataset-sdtm, analysis-dataset-adam, and data-tabulations-dataset-send) [eCTD check number: 1735]
- MEDIUM severity: For each study, no more than one dataset of the same type should be submitted as new [eCTD check number: 1737]

Although the checks are not yet effective at the time of writing this paper (April 2017), the FDA will give the industry 30 days’ notice on the eCTD website prior to the criteria becoming effective. It is prudent that the Sponsor is ready to have this implemented for their study immediately and not wait for this check to be effective to avoid any last-minute update to the dataset submission package.

TECHNICAL CONFORMANCE GUIDE (KEY DOCUMENT TO GET IT RIGHT!)

The word cloud below speaks to the various key high-level sections from the TCG. The most current version of the TCG (March 2017) is 45 pages long [4] and it provides specifications, recommendations, and general considerations on how to submit standardized study data using FDA-supported data standards located in the FDA Data Standards Catalog. The FDA plans to publish updated version of a TCG in March and October of each calendar year.

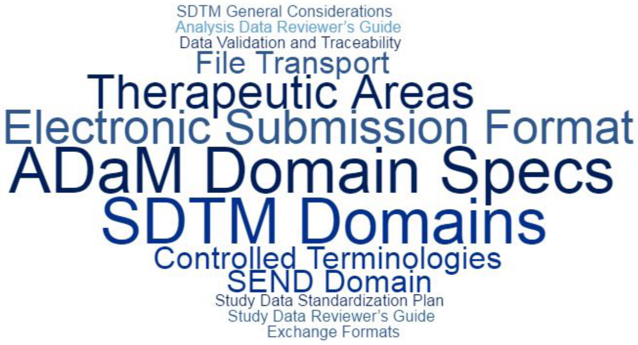

SDTM General Considerations
Analysis Data Reviewer's Guide
Data Validation and Traceability
File Transport
Therapeutic Areas
Electronic Submission Format
ADaM Domain Specs
SDTM Domains
Controlled Terminologies
SEND Domain
Study Data Standardization Plan
Study Data Reviewer's Guide
Exchange Formats

Figure 3: Word cloud of key sections from the TCG

It is critical that the Sponsor (specifically Statistical Programmer responsible for generating study datasets) is familiar with this guidance and more importantly understand/implement the minimum requirements set forth in this document. The key aspects from the TCG is provided in the table below which will become our basis to justify for an additional quality check of the data submission folder before submitting it to the FDA.

TCG Section	**Key requirements (Note: This is not an extensive list)**
Planning and Providing Standardized Data – SDSP	Study Data Standardization Plan (SDSP) at the pre-IND/IND stage. The cover letter accompanying a study data submission should describe the extent to which the latest version of the SDSP was executed.
Planning and Providing Standardized Data – SDRG	Inclusion of the Study Data Reviewers Guide (SDRG) for nonclinical studies (nSDRG) and SDRG for clinical studies (cSDRG) with the study data in Module 4 and 5, respectively, in the Electronic Common Technical Document (eCTD).
Planning and Providing Standardized Data – ADRG	Inclusion of the Analysis Data Reviewers Guide (ADRG) for clinical studies (ADRG) with the study data in Module 5, in the eCTD.
Exchange format – file types	File type included in the submission should be restricted to XML (define files), PDF (study documents), and XPT (datasets). Other file types could be included such as ASCII (analysis programs) and XSL (define style sheet).
Exchange format – SAS transport format	XPORT files should be created by the COPY Procedure in SAS. All SAS XPORT transport files should use .xpt as the file extension. There should be one dataset per XPORT file and the files should not be compressed.
Exchange format – Dataset size	Datasets greater than 5 gigabytes (GB) in size should be split into smaller datasets no larger than 5 GB. The split datasets should be placed in a separate sub-directory labeled “split”.
Exchange format – Dataset requirements	Dataset column length is set to a maximum length of the variable used. The length of variables should be less than or equal to 8. The length of variable labels should be less than or equal to 40. The length of dataset label should be less than or equal to 40. The variable and dataset names should not contain punctuation, dashes, spaces, or other non-alphanumeric symbols. The variable and dataset names should not contain special characters.
Study Data Submission Format - Nonclinical dataset (tabulations)	Nonclinical study that started after Dec 17, 2016 is using CDISC SEND data standards per the Standards Catalog. Supported version: o Model: 1.2; IG 3.0

TCG Section	**Key requirements (Note: This is not an extensive list)**
Study Data Submission Format – Clinical dataset (tabulations)	Clinical study that started after Dec 17, 2016 is using CDISC SDTM data standards per the Standards Catalog. Supported version: o Model: 1.1; IG 3.1.1 - SUPPORT ENDED 01/28/2015 o Model: 1.2; IG 3.1.2 o Model: 1.2, IG 3.1.2 amendment 1 o Model: 1.3, IG 3.1.3 o Model: 1.4, IG 3.2
Study Data Submission Format – Clinical analysis dataset	Clinical study that started after Dec 17, 2016 is using CDISC ADaM data standards per the Standards Catalog. Supported version: o Model: 2.1; IG 1.0
Study Data Submission Format – Trial Design Domains	All TDM datasets should be included in the submissions and Trial Summary (TS) dataset will be used to determine the time of study start. This will apply for both clinical (SDTM) and nonclinical (SEND) studies. A Trial Summary dataset (ts.xpt) must be presented for each study even if the study started prior to December 17, 2016.
Study Data Submission Format – Data Definition File	Inclusion of a define.xml for all study data submission format (SEND, SDTM, and ADaM). A printable define.pdf for define version 1.0 should be included. Supported version: o Standard: 1.0 - SUPPORT ENDING 03/15/2018 o Standard: 2.0
Study Data Submission Format – Annotated Case Report Form (aCRF)	Inclusion of aCRF for clinical tabulations datasets (legacy and SDTM compliant). Annotated and bookmarked per the SDTM Metadata Submission Guidelines (MSG).
Terminology – Controlled Terms	Sponsors should use the terminologies and code lists in the CDISC Controlled Terminology, which can be found at the NCI (National Cancer Institute) Enterprise Vocabulary Services.
Terminology – Dictionary	Adverse events coded using MedDRA Dictionary version 8 or later. Medication coded using WHO Drug Dictionary latest version – required after 03/15/2018 for NDA, ANDAs and certain BLAs.
Terminology – Other Standards	Utilize terminology standards such as: FDA Unique Ingredient Identifier, Pharmacological Class, and National Drug File (NDF), SNOMED CT. Laboratory test terminology using LOINC - required after 03/15/2018 for NDA, ANDAs and certain BLAs

TCG Section	**Key requirements (Note: This is not an extensive list)**
Electronic Submission Format - eCTD File Directory Structure	Study datasets and their supportive files should be organized into a specific file directory structure when submitted in the eCTD format (‘m5’ datasets folder). e.g.: o Directory structure per the Study Dataset and File Folder Structure as defined in the eCTD TCG o Folders are named per the nomenclature required for eCTD o File name does not exceed maximum length (64 characters) per eCTD specification o For naming files, one should use lower case characters and avoid special characters such as hyphen, underscore, punctuation, spaces and non-alphanumeric variables.
Study Data Validation – Conformance to Standards	The datasets submitted conformance to the published standards (e.g. CDISC SDTM, CDISC ADaM, Controlled Terminologies, CDISC Define.xml, MedDRA, WHO Drug).
Study Data Validation – Technical Rejection Criteria	FDA eCTD Technical Rejection Criteria for Study Data that assess conformance to the standards listed in the FDA Data Standards Catalog (see above section for details).
Study Data Validation – FDA Business Rules	All business rules should be followed where applicable. The business rules supersede previously published validation rule sets for both clinical and nonclinical data. Per the current version 1.1 (March 2017), there are 54 unique business rules (n=9 for nonclinical datasets, n=13 for clinical datasets, and n=32 for both clinical and nonclinical datasets).
Study Data Validation – FDA Validator Rules	The business rules are accompanied with validator rules which provide detail regarding FDA's assessment of study data for purposes of review and analysis. Per the current version 1.1 (March 2017), there are 115 unique validator rules.

QUALITY CHECK YOUR CDISC DATA SUBMISSION

FDA Binding Guidance is already in effect! CDISC data standards is a must for every new drug submission. The Sponsor needs to be prepared for CDISC data submission which meets the minimum set requirements per the FDA Data Standards Catalog, Technical Rejection Criteria’s, Study Data Technical Conformance Guide, Conformance to Data Standards, FDA Business Rules, FDA Validator Rules, and eCTD Validation Criteria’s.

The requirements could be overwhelming if they are not well understood and are not incorporated early in the clinical trial process. The Sponsor should be prepared to deliver a ‘submission-ready’ datasets from the get go. However, due to the lengthy drug development process (average 6 to 11 years just in the clinical stage), there will be multiple instances where a Sponsor will have to go back to their older/completed study to ensure it meets the needs and requirements per the current FDA expectations and standards.

Yes, there are tools that are available to check for the compliance of the datasets (e.g. Pinnacle 21) on a study level. In addition, there are tools to check against the eCTD specifications. However, one has to look at their data submission package in a more holistic way and incorporate quality step that checks for some of the critical items before a Sponsor submits their New Drug Application (NDA) to the FDA or PMDA. In the next sections, we intend to provide few examples of the checks that could be applied on a global level (i.e. across all studies which are part of the NDA) and also suggest ideas for how one could implement those checks using SAS-based tool.

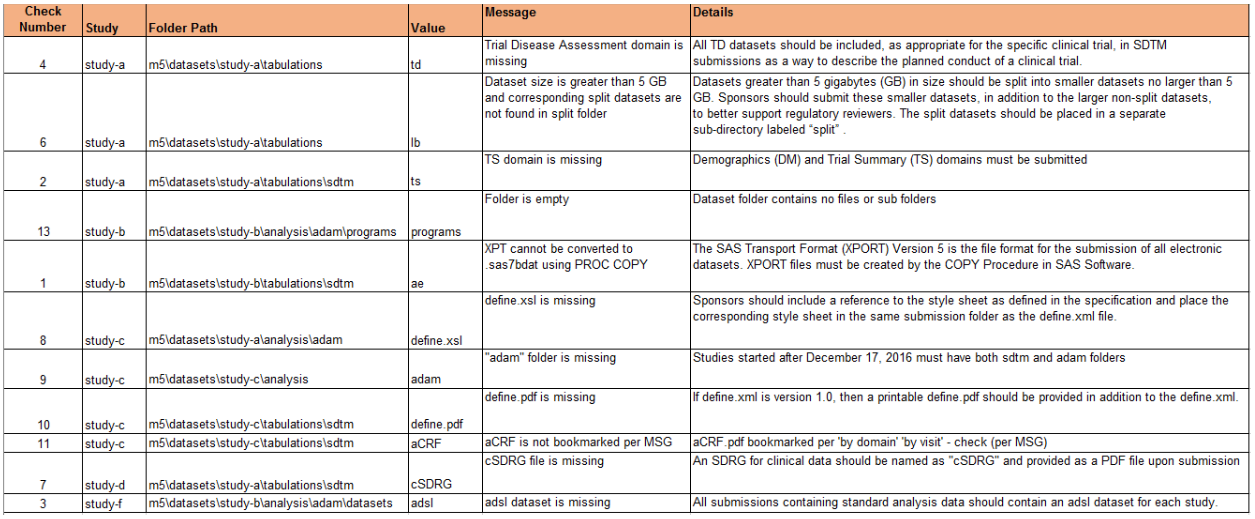
EXAMPLES OF THE CHECKS

Check Number	Details	Message
1	The SAS Transport Format (XPORT) Version 5 is the file format for the submission of all electronic datasets. XPORT files must be created by the COPY Procedure in SAS Software	XPT is not able to convert to SAS using PROC COPY
2	Demographics (DM) and Trial Summary (TS) domains must be submitted	TS domain is missing
3	All submissions containing standard analysis data should contain an ADSL file for each clinical study	ADSL dataset is missing
4	All TD datasets should be included, as appropriate for the specific clinical trial, in SDTM submissions as a way to describe the planned conduct of a clinical trial.	Trial Disease Assessment domain is missing
5	For each study, no more than one dataset of the same name should be submitted as new	Same name dataset is present in more than one folder within a study.
6	Datasets greater than 5 gigabytes (GB) in size should be split into smaller datasets no larger than 5 GB. Sponsors should submit these smaller datasets, in addition to the larger non-split datasets, to better support regulatory reviewers. The split datasets should be placed in a separate sub-directory labeled “split”.	Dataset size is greater than 5 GB and corresponding split datasets are not found in split folder
7	An SDRG for clinical data should be named as "cSDRG" and provided as a PDF file upon submission	cSDRG file is missing
8	Sponsors should include a reference to the style sheet as defined in the specification and place the corresponding style sheet in the same submission folder as the define.xml file.	define.xsl is missing
9	Studies started after December 17, 2016 must have both sdtm and adam folders	"adam" folder is missing

Check Number	Details	Message
10	If define.xml is version 1.0, then a printable define.pdf should be provided in addition to the define.xml.	define.pdf is missing
11	aCRF.pdf bookmarked per 'by domain' 'by visit' - check (per MSG)	aCRF is not bookmarked per MSG
12	Study datasets and their supportive files should be organized into a specific file directory structure when submitted in the eCTD format	The study dataset folder is not per the eCTD File Directory Structure
13	Dataset folder contains no files or sub folders	Folder is empty
14	File contains invalid file extension (allowed file extensions: .pdf, .xpt, .xml, .xsl, and .txt)	Invalid file type found
15	The datasets included in the submission package must be the most current version available for the study	Datasets included are out of date and do not match the latest version

EXAMPLE OF THE QUALITY REPORT VIA SAS-BASED TOOL

The above listed checks were incorporated using a sophisticated suite of macros in SAS. The tool is still under development and we intend to share more information about it during the presentation at PharmaSUG 2017 and at future public events. However, as an example, the below figure provides a snapshot of the quality report that gets generated as a result of running our SAS-based tool.



Check Number	Study	Folder Path	Value	Message	Details
4	study-a	m5\datasets\study-a\tabulations	td	Trial Disease Assessment domain is missing	All TD datasets should be included, as appropriate for the specific clinical trial, in SDTM submissions as a way to describe the planned conduct of a clinical trial.
6	study-a	m5\datasets\study-a\tabulations	lb	Dataset size is greater than 5 GB and corresponding split datasets are not found in split folder	Datasets greater than 5 gigabytes (GB) in size should be split into smaller datasets no larger than 5 GB. Sponsors should submit these smaller datasets, in addition to the larger non-split datasets, to better support regulatory reviewers. The split datasets should be placed in a separate sub-directory labeled “split” .
2	study-a	m5\datasets\study-a\tabulations\sdtm	ts	TS domain is missing	Demographics (DM) and Trial Summary (TS) domains must be submitted
13	study-b	m5\datasets\study-b\analysis\adam\programs	programs	Folder is empty	Dataset folder contains no files or sub folders
1	study-b	m5\datasets\study-b\tabulations\sdtm	ae	XPT cannot be converted to .sas7bdat using PROC COPY	The SAS Transport Format (XPORT) Version 5 is the file format for the submission of all electronic datasets. XPORT files must be created by the COPY Procedure in SAS Software.
8	study-c	m5\datasets\study-a\analysis\adam	define.xsl	define.xsl is missing	Sponsors should include a reference to the style sheet as defined in the specification and place the corresponding style sheet in the same submission folder as the define.xml file.
9	study-c	m5\datasets\study-c\analysis	adam	"adam" folder is missing	Studies started after December 17, 2016 must have both sdtm and adam folders
10	study-c	m5\datasets\study-c\tabulations\sdtm	define.pdf	define.pdf is missing	If define.xml is version 1.0, then a printable define.pdf should be provided in addition to the define.xml.
11	study-c	m5\datasets\study-c\tabulations\sdtm	aCRF	aCRF is not bookmarked per MSG	aCRF.pdf bookmarked per 'by domain' 'by visit' - check (per MSG)
7	study-d	m5\datasets\study-a\tabulations\sdtm	cSDRG	cSDRG file is missing	An SDRG for clinical data should be named as "cSDRG" and provided as a PDF file upon submission
3	study-f	m5\datasets\study-b\analysis\adam\datasets	adsl	adsl dataset is missing	All submissions containing standard analysis data should contain an adsl dataset for each study.

Figure 4: Example of the Quality Report via SAS-based Tool

CONCLUSION

FDA Binding Guidance is already in effect! CDISC data standards is a must for every new drug submission. The Sponsor needs to be prepared for CDISC data submission which meets the minimum requirements set forth by the FDA and PMDA. The intention with the checks suggested above is not to replace what a Sponsor does on a study-level. The below checks are to supplement those and to implement them at a global level before submitting your dataset package to the FDA. These checks are meant to be at a dataset level and not necessarily at the variable or content level. The Sponsor can build these checks using a SAS-based tool to ensure your CDISC Data Submission folder meets the quality and requirements before the FDA/PMDA runs their tools and potentially issue RTF or RTR for non-conformance.

REFERENCES

[1] Providing Regulatory Submissions In Electronic Format - Standardized Study Data, December 2014. https://www.fda.gov/downloads/drugs/guidances/ucm292334.pdf

[2] FDA Data Standards Catalog, v4.5, September 3, 2015. https://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm.

[3] FDA Technical Rejection Criteria for Study Data, revised March 02, 2017. https://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/FormsSubmissionRequirements/ElectronicSubmissions/UCM523539.pdf

[4] FDA Study Data Technical Conformance Guide, Technical Specification Document, version March 2017. https://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM384744.pdf

[5] Study Data Standards in eCTD: What You Need to Know About the New Technical Rejection Criteria, October 12, 2016. http://sbiaevents.com/files/eCTD-Study-Data-Standards-Webinar-Oct-2016.pdf

CONTACT INFORMATION

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

Bhavin Busa
Organization: Vita Data Sciences, a division of Softworld, Inc
Address: 281 Winter St., Waltham, MA 02451
Work Phone: 781-373-8455
Email: bbusa@vitadatasciences.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.

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