PharmaSUG 2018 – Paper SS-02

Preparing to Meet FDA Requirements for Submission of Standardized Data and Documentation

Steve Kirby, Mario Widel, Covance Inc.

ABSTRACT

PDUFA V gave the FDA the authority to require electronic submission of study data in standard format. That authority was confirmed by PDUFA VI. We will provide a practical overview of FDA expectations for submission of standardized study data and associated documentation for statisticians and statistical programmers and share how those expectations can be met.

INTRODUCTION

With CDISC data standards, there is a place for everything and everything should be in its place. That said, by design, CDISC standards are flexible and there are many, many (and dare we say again) many acceptable mapping strategies under several acceptable (per FDA, at a given point in time) versions of CDISC data, documentation and terminology standards. And, lest we forget, coding dictionaries (such as MedDRA and WHO Drug) are updated often, toxicity criteria evolve over time; data from across studies need to be pooled in a comprehensive and scientifically valid way; and planning needs to change based on actual results and company priorities. What follows is an overview of what needs to be submitted and some practical ways to ensure that you will have what the FDA needs and expects available at the time of submission.

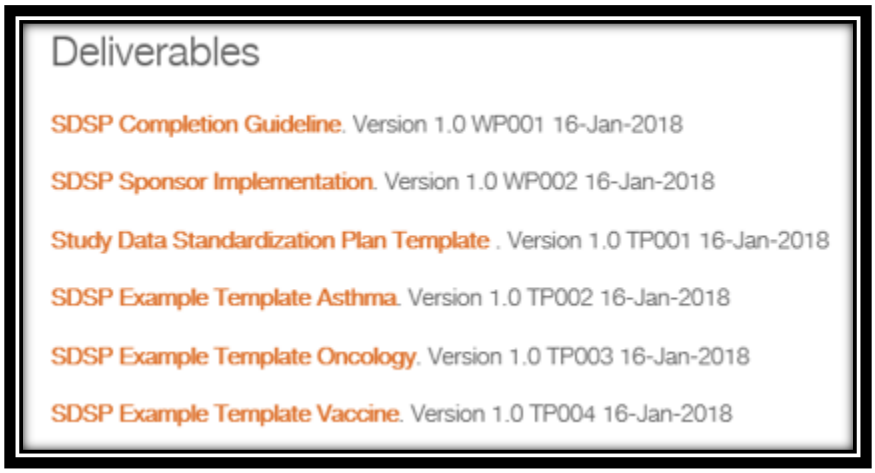
PLANNING ACROSS STUDIES

STUDY DATA STANDARDIZATION PLAN

What versions of SDTM and ADaM standards were followed? What versions will be followed? All those versions are accepted by the FDA, right? Do legacy format data need to be included for traceability? For some other reason? What coding and terminology versions were used and will be used for pooled analysis? And what conformance review program versions were and will be used? And what content will be pooled? And how. And why? And what were those details again? And how do the latest results impact prior planning? And, and, and, dare we say it again, and.

A huge number of interrelated details associated with submission of standardized study data need to be managed over a long period of time. And all the details are important, so important that the FDA expects to collaborate with sponsors to ensure that the scientific objectives are supported by the data and that regulators can efficiently find the information they need to judge whether a drug is safe and effective.

That is a lot of standards related content – and all of it has to be managed and documented and shared with the FDA. Luckily for us, PhUSE has released content designed to support this exact need: the Study Data Submission Plan (SDSP). Table 1 below shows the current completion guidelines, template and available example documents. Table 2 shows the template used to document exchange and terminology standards for individual clinical studies. As with virtually everything in data standards, SDSP content is subject to updates. Please regularly check to make sure you have the latest version. SDSP content was sourced from https://www.phuse.eu/css-deliverables.



Deliverables

SDSP Completion Guideline. Version 1.0 WP001 16-Jan-2018

SDSP Sponsor Implementation. Version 1.0 WP002 16-Jan-2018

Study Data Standardization Plan Template . Version 1.0 TP001 16-Jan-2018

SDSP Example Template Asthma. Version 1.0 TP002 16-Jan-2018

SDSP Example Template Oncology. Version 1.0 TP003 16-Jan-2018

SDSP Example Template Vaccine. Version 1.0 TP004 16-Jan-2018

Table 1. SDSP - List of Supported Content

4.2 Clinical

Study Identifier	Brief Title	Study Design	Study Status	Study Start Date	Exchange Standards	Terminology Standards
<Phase <x>> <Interventional/Observational/Expanded Access> Studies - <indication or Healthy Subjects or Healthy Volunteers>						
If value is unknown, specify TBD		Please See Completion Guidelines for more information If values are unknown, leave blank or specify TBD	COMPLETED ONGOING PLANNED	ccyy-mm-dd <(forecasted Informed Consent)> TBD	ANALYSIS LEGACY ADaM v<version>/ ADaM IG <version> ADaM define.xml <version> ADaM vTBD/ ADaM IG TBD ADaM define.xml TBD TABULATIONS LEGACY SDTM v<version>/ SDTM IG <version> SDTM define.xml <version> SDTM vTBD/ SDTM IG TBD	Sponsor Defined Terminology CDISC SDTM Terminology <date> <TBD> MedDRA (Adverse Events/ Medical History/ <other>) Initial <version> Final <version> Initial TBD Final TBD WHO-DD (Medications) <version> <TBD> LOINC (Lab Test Term) <version> <TBD>

Table 2. SDSP - Content for Individual Studies

Since all these decisions need to be managed one way or another, why not use the PhUSE SDSP content as a framework for collecting information for completed studies and to document future plans? Building this work into your data standards efforts will help to align your work with what you need to submit. It can also highlight any gaps in planning and help broadcast core submission data and documentation needs to the study team and upper management. We will dig into some of the most important practical considerations for tabulations and analysis content in the tabulations and analysis sections below.

And there is more! Safety and efficacy information across studies will need to be pooled in a comprehensive and scientifically reasonable way. The scope of content that will be pooled for a given submission needs to be determined by the sponsor in coordination with the FDA. The details within that scope need to be managed by the sponsor. And, as with individual study data, PhUSE has a template designed to help sponsors plan and document the details of their pooling strategy. Table 3 below contains example content for pooled studies from the PhUSE SDSP template.

4.3 Pooled Studies

Data Pool Identifier	Data Pool (List of	Pool Status	Pool Description	Exchange Standards	Terminology Standards
		CURRENT	ISS <any additional information, such as certain domains>	ANALYSIS LEGACY	Sponsor Defined Terminology
		PLANNED	ISE <any additional information, such as certain domains>	ADaM v<version>/ ADaM IG <version>	CDISC SDTM Terminology <date>
				ADaM define.xml <version>	<TBD>
				ADaM vTBD/ ADaM IG TBD	MedDRA (Adverse Events/ Medical History/ <other>) Initial <version> Final <version>
				ADaM define.xml TBD	Initial TBD Final TBD
				TABULATIONS LEGACY	WHO-DD (Medications)
				SDTM v<version>/ SDTM IG <version>	<version> <TBD>
				SDTM define.xml <version>	LOINC (Lab Test Term) <version> <TBD>
				SDTM vTBD/ SDTM IG TBD	SNOMED CT (Indication) <version> <TBD>
				SDTM define.xml TBD	

Table 3. SDSP – Content for Pooled Studies

Using this template as part of submission planning will not only make it so you are ready to effectively communicate your plans with FDA reviewers; it will also highlight the submission data work that needs to

be accomplished. With the end goals in mind, you can effectively plan to complete the work in time and easily communicate what needs to be done with your internal submission leaders. There is typically a need for a number of additional internal documents to precisely specify how the deliverables listed in the SDSP will be generated.

Using safety pooling as an example, below are a few key operational considerations that need to be clearly planned.

- What studies need to be included?
- What content needs to be pooled?
- Will SDTM be submitted as pooled domains?
- What SDTM version and CT version will be followed for domains that will be pooled?
- What dictionary versions will be used across studies and what studies need to be up-versioned?
    - Common MedDRA version
    - Common WHO-DD version
    - Sponsor dictionaries/Queries
- What other content needs to be made consistent across pooled studies?
    - CTCAE?
    - LOINC?
    - Other content?

Pooling data for ISS and ISE is a major challenge. Documenting the target goals as needed for FDA review is a useful way to thoroughly establish what needs to be completed, get buy in from all the stakeholders (including the FDA), and effectively plan to close out the work.

Now that we have an overview of what you need to submit, it is time to look at key SDTM and ADaM deliverables. Just as you need to plan to have all the content needed for review available in the right format, you need to make sure all that the content complies with applicable rules.

Our focus will be on CDISC data and documentation deliverables. And the main focus of submission data and documentation activities should be on CDISC format data as well. But we do want to point out that there are a few cases where legacy format data needs to be documented and submitted. One key case is where the original data summarization was done using a legacy data input. In those cases, the legacy tabulations and analysis data should be submitted to retain traceability to the original report even if the data are later mapped to SDTM and ADaM for the convenience of reviewers and/or to support integrated summaries. And what if the source data to not cleanly map to SDTM? Requirements for that case are not fully established but the raw data may need to be submitted to ensure there is a clear path from SDTM to source data as collected. When legacy data need to be submitted, those data must be in xpt format and supported by documentation – much as what is needed for CDISC data.

SUBMISSION OF TABULATIONS DATA PACKAGES

The key clinical tabulations data items in a submission package are the SDTM annotated CRF (acrf.pdf, formerly known as the blankcrf.pdf), the Clinical Study Data Reviewer's Guide (csdrg.pdf, formerly known as the sdrg.pdf); the data definitions document (define.xml 2.0 formerly known as the define.xml 1.0 formerly the define.pdf) and, you guessed it, the tabulations datasets in xpt format. So make sure you have all those items in hand, that they follow all applicable guidelines and are appropriately reviewed. In case you need it, some more details are below. And please keep in mind that submission requirements change over time; make sure to stay up to date. It was not so long ago that all NDAs were supported by a big truck or two full of paper reports.

SDTM DOMAINS IN XPT FORMAT

Without data, you are just another person with an opinion. And the data used to support FDA submission must not only support your scientific conclusions and have a clear relation to study conduct; it also needs to be fit for FDA use and review.

xpt file requirements

One part of fit for use is that the datasets must be in an approved format – otherwise they cannot be used and reviewed. Currently data must be submitted in SAS v5 xpt format. (Version 5 xpt files are vendor neutral – they can be created and used without using a SAS product). Some format limitations apply to SAS v5 files. To comply with the xpt format requirements:

- Dataset and variable names must be <= 8 characters
- Dataset and variable labels must be <= 40 characters
- Dataset and variable names and labels should only include American Standard Code for Information Interchange (ASCII) text codes
- Dataset and variable names should not contain punctuation, dashes, spaces, or other non-alphanumeric symbols or special characters
- Character variables must be <=200 characters in length.
- See, generally, the Study Data Technical Conformance Guide (current version as of this paper is v 4.0, Oct. 2017)

And that's not all. The xpt datasets need to be no (or at least not much) bigger than necessary. To meet that requirement:

- Character variables in main domains need to be trimmed to the minimum length needed across datasets;
- Character variables in supplemental domains need to be trimmed to the minimum length needed within each dataset.

A common approach to handle the requirement that datasets be no bigger than needed is to initially generate domains using the maximum acceptable length for character variables and then use standard code:

- Trim data to the minimum length needed,
- Convert to xpt format, and
- Verify that dataset attributes and content were not changed (but for variable length).

Establishing a process covering all these details (if not already in place) is an easy way to streamline future work. Note that these xpt file requirements apply to all submitted data (including legacy format data).

And one more thing. Each xpt file needs to be < 5GB (as of when this paper is written – the allowed size increases over time). If a dataset (as listed in the domains section of the define.xml) is >= 5 GB, it should be split to meet the size requirements following the guidelines in the SDTM IG. Split datasets should be placed in a subdirectory labeled "split" and a clear explanation regarding how these datasets were split needs to be presented within the csdrg.pdf. The larger non-split datasets should be submitted as well. A separate define.xml does not need to be submitted based on the split datasets.

Controlled terminology

Another part of fit for use is following CDISC controlled terminology (CT) where applicable; and with SDTM, controlled terminology is broadly applicable. First things first. At a minimum it is important to pick and document a specific version of controlled terminology for each set of SDTM that will be submitted

(whether for an individual study or for pooled domains). This is an obvious point; but all too often the applicable version is not established at the time the eCRF is designed (or at least before mapping starts). Accurately reviewing SDTM data is impossible without knowing the specific CT target list. If a need to up-version to a more recent version of CT surfaces after mapping starts, a comparison of the two documented CT versions can inform how the update should be completed.

To the extent practically possible, the same CT version should be used in as many studies as possible. As discussed further in the conformance checks section below, when an applicable CT choice exists it must be used. While many terminology lists can be extended if needed, sponsor additions should not be used unless there is no other choice.

ACRF.PDF

Everyone likes pictures, and the SDTM annotated CRF provides a pictorial representation of how CRF data (and where reasonable, associated collection documents) are mapped to SDTM. The acrf.pdf contains text boxes that associate clinical data collection fields with corresponding SDTM variables or values. When data are collected but not submitted, the associated acrf.pdf content should be annotated with the text “NOT SUBMITTED.” When more than one domain is on a single CRF page, different background colors should be used for the annotation boxes. The acrf.pdf should be bookmarked by form or domain and visit. Establishing a local process to define how the acrf.pdf is generated is recommended. Without that process it can be challenging (if not impossible) to have consistent annotations across a submission.

The acrf.pdf has a close association with the define.xml. When the define.xml specifies an origin of CRF page xx for a variable or value, the physical pdf page xx of the acrf.pdf needs to have an annotation for that variable or value. Consistency checks between the define.xml origins and the acrf.pdf annotations can (and should) be supported programmatically.

DEFINE.XML

All SDTM data must be supported by a define.xml. The define.xml presents the SDTM metadata in a user-friendly format that can be programmatically reviewed for conformance and consistency with the study data. The document displays the metadata at increasing levels of specificity. Table 4 below contains a pictorial representation of the information layers.

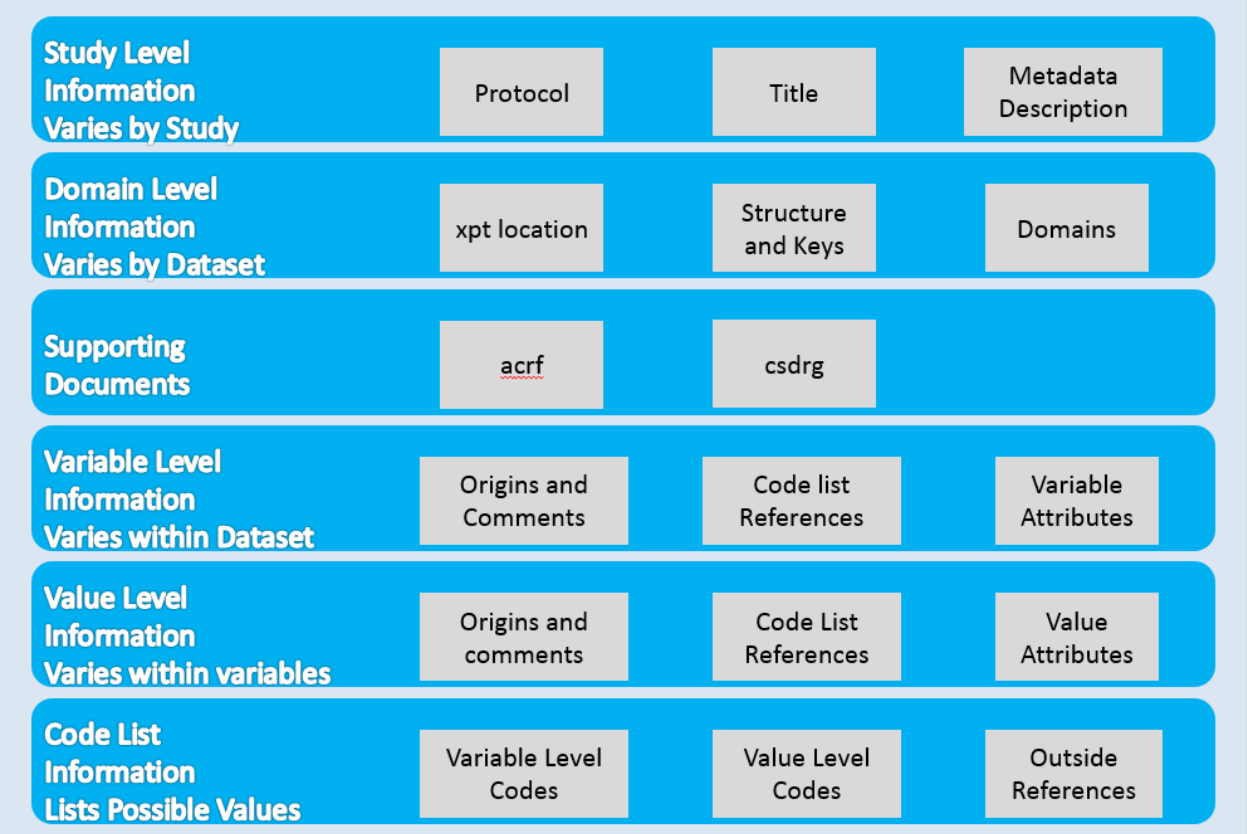


Table 4. Define.xml Information Layers

The define.xml shows the origins and derivations and (where applicable) acceptable values (code lists) for each variable and value in the SDTM data, as well as providing a list of domains and key variables. And importantly, the define.xml also provides hyperlinks to the acrf.pdf, the csdrg.pdf and the xpt files.

While define.xml 1.0 is allowed for studies starting prior to 2018, we recommend moving to define.xml 2.0 (whether you have to or not) as soon as practically possible. Define.xml 1.0 contains less information than define.xml 2.0. And unlike the define.xml 2.0, the define.xml 1.0 cannot be printed and must be accompanied by a define.pdf. Creating a pdf version of the define.xml is often burdensome and adds another potential source of error.

CONFORMANCE CHECKS

Saying you are following a specific version of SDTM (and CT) is not enough, you need to prove it with programmatic conformance review. Similarly, sponsors are required to show that the define.xml (by itself) is well formed and that (when data and define are reviewed together) the define.xml is consistent with the data.

Conformance review is typically done through the use of a Pinnacle 21 (P21) application. Building in conformance review early (and often) during generation of SDTM data and documentation reduces the chance that changes to the SDTM domains will be needed after downstream users have relied on the data. And time spent building an issue management process and supporting it with helper applications and metadata has led to a notable decrease in conformance review time spent by our teams.

Just to say it out loud (as there are often many practical reasons to go a slightly different direction): All avoidable issues should be avoided. As not all issues are avoidable, sponsors are required to explain remaining conformance issues in the csdrg.pdf. If you repeatedly find yourself faced with a large number of unavoidable issues, the data as collected is likely inconsistent with SDTM standards and it is probably time to update your standard collection documents (and maybe even your protocol template) for consistency with SDTM.

CSDRG.PDF

We like to think of the csdrg.pdf as an executive overview of the tabulations data. When well formed, it will quickly orient a data consumer to the referenced data. Working on the csdrg.pdf while mapping is ongoing can save time and lead to a more complete result. For example, documenting mapping choices that benefit from further explanation is most easily done when the choices are made. And being mindful of how the data need to be summarized in the csdrg.pdf can usefully guide how sponsors document content ranging from specifying applicable dictionaries, SDTM and CT versions to how conformance issues are managed and explained.

As with the SDSP, the PhUSE organization has stepped up and developed a template and associated completion guidelines for the csdrg.pdf (see http://www.phusewiki.org/wiki/index.php?title=Study_Data_Reviewer%27s_Guide) for the specifics, and monitor to make sure you are aware if an updated version is released). Following this template is the safest way to ensure that your csdrg.pdf contains all the content data consumers (most notably the FDA) need.

Two key sections of the csdrg.pdf are the issues summary and the table of Inclusion/Exclusion criteria. Both are easily completed with a bit of upfront planning.

The issues summary documents sponsor explanations for unavoidable issues. It can easily be a vexing experience to decide on a case-by-case basis if an issue can or cannot be resolved; and if it cannot be resolved, it is often challenging to decide on the most appropriate explanation. Having a standard set of explanations on hand for unavoidable issues can streamline the process.

A full set of the Inclusion/Exclusion criteria is needed for the csdrg.pdf. This set needs to contain the full text of each criterion as stated in the protocol. The TI domain is a likely resource for this content. But as IETEST must be < = 200 characters and not contain non-ascii or special characters, TI.IETEST typically does not contain the full text of each criterion as stated in the protocol. Having a TI specification process that retains the full text as a support column can help ensure shortened versions in IETEST are consistent with the original protocol content and can directly support the Inclusion/Exclusion table in the csdrg.pdf.

SUBMISSION OF ANALYSIS DATA PACKAGES

ADAM DATASETS IN XPT FORMAT

The requirements related to xpt file format for ADaM datasets are the same as for SDTM.

CONFORMANCE CHECKS

Similar to SDTM, avoidable issues should be avoided and unavoidable issues need to be explained in the adrg.pdf. And also as with SDTM, time spent planning to manage ADaM conformance issues is typically time well spent. Machine verifiable ADaM compliance issues are fewer in number and usually more easily managed than in SDTM. Unlike with SDTM, ADaM conformance is not subject to how the data are collected, so most issues are based in mapping choices and can be resolved by updating the modeling approach. All that said, many ADaM issues cannot be programmatically detected and require manual review.

ADRG.PDF

An ADaM reviewer's guide is required; and just as with SDTM, PhUSE has a template, completion guidelines and examples. That content can be found on the PhUSE wiki at http://www.phusewiki.org/wiki/index.php?title=Analysis_Data_Reviewers_Guide_(ADRG)_finalized

While the ADRG is similar in format to the SDRG, ADRG content is focused on the relation of the data to objectives and to special circumstances related to analyses.

DEFINE.XML

The define.xml (2.0) files for ADaM and SDTM are structurally similar, however since in the ADaM define the focus is on analysis datasets obtained from SDTM there is no requirement to reference the acrf.pdf.

To better describe analyses, the ADaM define.xml may include an optional (per FDA, required by the PMDA) Analysis Results Metadata (ARM) section where analysis results provenance is explained in detail. Table 5 below shows an example of ARM content.

Display	**Table 14-3.01 Primary Endpoint Analysis: ADAS-Cog - Summary at Week 24 - LOCF (Efficacy Population)**
Analysis Result	**Dose response analysis for ADAS-Cog changes from baseline**
Analysis Parameter(s)	PARAMCD = "ACTOT" (Adas-Cog(11) Subscore)
Analysis Variable(s)	CHG (Change from Baseline)
Analysis Reason	SPECIFIED IN SAP
Analysis Purpose	PRIMARY OUTCOME MEASURE
Data References (incl. Selection Criteria)	ADQSADAS [PARAMCD = "ACTOT" and AVISIT = "Week 24" and EFFFL = "Y" and ANL01FL = "Y"]
Documentation	Linear model analysis of CHG for dose response; using randomized dose (0 for placebo; 54 for low dose; 81 for high dose) and site group in model. Used PROC GLM in SAS to produce p-value (from Type III SS for treatment dose). SAP Section 10.1.1

Programming Statements	[SAS version 9.2]
	proc glm data = ADQSADAS;
	where EFFFL='Y' and ANL01FL='Y' and AVISIT='Week 24' and PARAMCD="ACTOT";
	class SITEGR1;
	model CHG = TRTPN SITEGR1;
	run;

Table 5. ARM – Example of Analysis Results Metadata.

From https://www.cdisc.org/standards/foundational/analysis-data-model-adam/analysis-results-metadata-arm-v10-define-xml-v20

CONCLUSION

Planning for submission to the FDA typically involves many studies that were conducted over many years and a large team dedicated to the effort. And don't forget the FDA reviewers, whose needs must be met by the submission content. Having a clear and detailed understanding of what content needs to be submitted is a necessary first step – and much better to plan before work starts than to revise content

considered complete before submission planning occurred. Documenting submission data targets in the SDSP makes it so planning for submission also supports a key submission deliverable.

Once documentation of the submission data targets are in place, executing on the plan is reduced to ensuring that the study data and documentation conforms to the applicable CDISC standard. Planning for conformance and proving that conformance with programmatic (and as needed) manual review will nicely close out the effort and make it so your team is justifiably confident that the data submitted will be accepted and efficiently reviewed.

REFERENCES

FDA *Study Data Standardization Plan Checklist Recommendations (SDTM, ADaM and CDASH).*

https://www.fda.gov/downloads/BiologicsBloodVaccines/DevelopmentApprovalProcess/UCM535589.docx

FDA *Providing Regulatory Submissions in Electronic Format — Certain Human Pharmaceutical Product Applications and Related Submissions Using the eCTD Specifications.* https://www.fda.gov/downloads/Drugs/*GuidanceComplianceRegulatoryInformation*/Guidances/UCM333969.pdf

FDA, *Study Technical Conformance Guide.* V 4.0, October 2017 *https://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM384744.pdf*

CDISC, *Study Data Tabulation Model (SDTM).* https://www.cdisc.org/standards/foundational/sdtm

CDISC, *Study Data Tabulation Model Implementation Guide (SDTMIG).*

https://www.cdisc.org/standards/foundational/sdtmig

CDISC, *ADaM Data Model*. https://www.cdisc.org/standards/foundational/adam

CDISC, *ADaM Implementation Guide*. https://www.cdisc.org/standards/foundational/adam

CDISC. *Define-XML Specification*, Version 2.0. March 5, 2013. http://www.cdisc.org/define-xml.

CDISC Case Report Tabulation Data Definition Specification (define.xml), Version 1.0, February 9, 2005. https://www.cdisc.org/system/files/all/standard_category/application/pdf/crt_ddspecification1_0_0.pdf

Jansen, Lex. *Analysis Results Metadata for Define-XML v2,* PhUSE http://www.phusewiki.org/docs/Denmark%202015%20SDE%20Presentations/Analysis%20Results%20Metadata%20.pdf

Randall, Amber and Coar, William. *"Strategic Considerations for CDISC Implementation".* PharmaSUG 2016. https://www.lexjansen.com/pharmasug/2016/SS/PharmaSUG-2016-SS03.pdf

Pang, Hang. *"New Features in Define-XML V2.0 and Its Impact on SDTM/ADaM Specifications". ".* PharmaSUG 2016. https://pharmasug.org/proceedings/2016/SS/PharmaSUG-2016-SS06.pdf

Dey, Mei and Peers, Diane. *"Delivering a quality CDISC compliant accelerated submission using an outsourced model".* PharmaSUG 2016. https://www.pharmasug.org/proceedings/2016/PO/PharmaSUG-2016-PO11.pdf

Busa, Bhavin. *"Quality Check your CDISC Data Submission Folder Before It Is Too Late!"* PharmaSUG 2017. https://www.pharmasug.org/proceedings/2017/SS/PharmaSUG-2017-SS12.pdf

Crockett, Eric. "*Supporting the CDISC Validation Life-Cycle with Microsoft Excel VBA".* PharmaSUG2017. *http://pharmasug.org/proceedings/2017/AD/PharmaSUG-2017-AD23.pdf*

ACKNOWLEDGMENTS

Thanks to Covance and Terek Peterson for supporting this paper. And thanks to the FDA for supporting CDISC data standards.

CONTACT INFORMATION

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

Steve Kirby
Covance, Inc.
Steven.Kirby@Chiltern.com

Mario Widel
Covance, Inc.
Mario.Widel@Chiltern.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.