BEST PRACTICES FOR DOCUMENTING DATASET METADATA: DEFINE-XML VERSUS REVIEWER'S GUIDE

2019-APR-05

Revision History

Version	Date	Summary
1.0	05APR2019	Initial Release

Contents

Overview:

This paper aims to define best practices for documenting and describing dataset structures and contents within standard submission deliverables such as Define-XML and the data reviewer's guide (DRG) by addressing common industry challenges and noting some important differences between reviewing agencies and divisions and recommending quality assurance activities.

Scope:

This paper will introduce the standard submission dataset metadata documents which are required or recommended for regulatory submissions of standardized clinical trial data to the US Food and Drug Administration (FDA) and will explore the relationships between these documents, with a focus on the contents of the study-level define.xml files and the clinical DRGs. Similar submission metadata documents may be required for submissions to other agencies, such as PMDA in Japan. Some of the differences in the metadata requirements between these agencies will be discussed. Technical issues and challenges concerning the files (such as style sheets) will not be addressed in this paper.
While the recommendations for the DRGs are based on the PhUSE templates[1], they would be applicable to any sponsor-specific templates as well. Additionally, while this paper focuses on metadata for clinical datasets, some of the recommendations may also apply to nonclinical data.

The suggestions and recommendations within this paper are intended to supplement existing DRG completion guidelines and Define-XML specifications and completion guidelines and to aid clinical trial teams in creating metadata files which have sufficient information and detail to support and facilitate the regulatory review process.

Definitions:

- **Metadata**: information describing datasets and their structures and contents, including the attributes of the variables contained within
- **Define-XML[3]**: model used to describe or transmit metadata for CDISC[2] SDTM, SEND, and ADaM datasets for the purposes of submissions to a regulatory agency, as well as any non-standard (legacy) dataset structure
- **define.xml**: the file specified by the Define-XML model which provides the metadata for the submitted datasets
- **Data reviewer's guide**: a document which describes any special considerations, directions, or conformance issues that may facilitate a regulatory reviewer's use of the submitted data and may help the reviewer understand the relationships between the study report and the data[4]
- **Legacy data**: study data in a non-standardized format, not supported by FDA, and not ever listed in the Data Standards Catalog[4,5]

Acronyms:

- **aCRF**: annotated Case Report Form
- **ADaM**: Analysis Data Model
- **ADRG**: Analysis Data Reviewers Guide
- **ANDA**: Abbreviated New Drug Application
- **ARM**: Analysis Results Metadata
- **BLA**: Biologics License Application
- **CRF**: Case Report Form
- **cSDRG**: Clinical Study Data Reviewer's Guide
- **CSR**: Clinical Study Report
- **CT**: Controlled Terminology
- **DRG**: Data Reviewer's Guide (including cSDRG and ADRG)
- **DSC**: Data Standards Catalog
- **IG**: Implementation Guide
- **IND**: Investigational New Drug Application
- **LDCP**: Legacy Data Conversion Plan and Report
- **NDA**: New Drug Application
- **nSDRG**: Nonclinical Study Data Reviewer's Guide
- **SAP**: Statistical Analysis Plan
- **SDRG**: Study Data Reviewers Guide (clinical and/or nonclinical)
- **SDSP**: Study Data Standardization Plan
- **SDTM**: Study Data Tabulation Model
- **SEND**: Standard for Exchange of Nonclinical Data
- **TCG**: Technical Conformance Guide
- **TLFs**: Tables, Listings, and Figures

Problem Statement:

Regulatory submissions of clinical trial data require creation of a define.xml (data definitions) file and recommend a DRG for the submitted datasets, but current guidance does not provide sufficiently detailed explanations and examples for authors to best complete this documentation to benefit regulatory reviews. Despite the increasing use of standards, there is still much variability in datasets between sponsors, programs, and studies. These supporting metadata files are intended to help reviewers navigate through and understand the datasets, however there is a great degree of subjectivity, judgment, and interpretation as to how certain fields and/or sections should be completed and what level of detail is useful. Additionally, it is not always clear where to document certain items and whether information should be provided in one or both documents. These issues may lead to documents being incomplete or not containing the desired information.

Background:

Studies starting on or after December 17, 2016 are required to submit standardized study data using FDA-supported data standards located in the FDA Data Standards Catalog[5] for NDA, ANDA, and certain BLA submissions, along with a study data definition file in the Define-XML standard[5,6]. Additionally, inclusion of DRGs is recommended[4].

The Define-XML specification[3] (originally known as the Case Report Tabulation Data Definition Specification, or CRT-DDS) was initially released by CDISC (version 1.0) in 2005 and had significant updates (as version 2.0) in 2013-2014. Version 2.0 is preferred by the FDA[4] as version 1.0 support was phased out in March 2018[5]. DRG templates, completion guidelines, and example documents were released by a PhUSE working group in 2013 for SDTM and 2014 for ADaM with updates in 2015 and 2018[1]. Although the PhUSE templates are not required to be used for a submission, they have been reviewed by the FDA and are referenced in the Technical Conformance Guide[4].

FDA presentations over the last several years[7,8,9,10,11,12] and the Technical Conformance Guide have stressed the importance of both documents and have encouraged sponsors to submit comprehensive metadata to help facilitate their reviews.

Introduction:

Key documents for describing the dataset metadata include the define.xml file and the DRGs:

The define.xml (also known as the data definitions file or CRT-DDS) is a required file that describes the structures and contents (datasets, variables, values, and controlled terminologies) of the data in a submission package. There is one file for each set or type of datasets: nonclinical tabulation data (SEND), clinical tabulation data (SDTM), or analysis data (ADaM) (note: legacy data structures may be described in a define.pdf file). The file should be as detailed as possible, particularly for derived variables, and should be written such that someone unfamiliar with the study can locate data values of interest and understand the source of the value or how it was computed or assigned. Companies should aim to use the most recent supported version (see the DSC[5]) as this will ensure that the desired features and functionality are included. A corresponding XSL style sheet should be provided with an appropriate style sheet reference in the XML. A PDF version can be provided for easy printing and should be provided if the XML version cannot be printed[4]. The define.xml file utilizes hyperlinks and bookmarks to allow navigation between related sections of the document or to other documents and is the key reference for documenting the source (CRF, external data, derived, etc.) for the values within each variable.

Analysis Results Metadata[13] (ARM) is a recommended extension to the analysis (ADaM) define.xml file. It is used to link the analysis results to documentation and datasets and provide additional details on the study analyses. It provides traceability for the study results by linking the results to the source analysis datasets (and metadata) along with documentation of the analyses (either a written description or programming code). The sponsor determines which key analyses to include, however the TCG specifically requests programs for all ADaM datasets and for tables and figures that generate primary and secondary efficacy analyses[4].

The DRGs provide additional context beyond the define.xml files, identifying the standards and versions implemented, describing the trial design, and noting any special considerations for the datasets and how they relate to or impact the study results. These documents are recommended to be included with the datasets in a submission package[4]. There is one file for each set or type of datasets: legacy, nonclinical tabulation (SEND), clinical tabulation (SDTM), and analysis data (ADaM). The DRGs complement the define.xml files and can be used to convey information that does not have a place in define.xml or is not easily expressed in that format. The SDRG is focused on the tabulation datasets and data collection/mapping decisions or points of interest. The ADRG is focused on the analysis datasets, providing an overview of the analysis of the study and describing how the Statistical Analysis Plan (SAP) was implemented within the datasets.

If data have been converted from legacy to standardized data formats, it is beneficial to describe the conversion process and results (particularly any challenges encountered and any resulting standards conformance issues) in a legacy data conversion plan and report (LDCP) which is included as an appendix to the cSDRG[1] and/or ADRG (updated ADRG template pending at the time of this writing). The plan describes the legacy data and the process for converting to standard format. The report presents the results of the conversion, and notes any issues identified and their resolution or explanation. It is an important tool for documenting traceability, identifying data points or values which did not fit appropriately into the standards, and explaining differences between the original data/results and standardized data/results. When both legacy data and standardized data are submitted, sponsors may provide a DRG for each set of data, but should minimally provide a DRG with the LDCP appendix with the standardized data.

Both the define.xml and the DRGs provide reviewers with important information about the submitted datasets. Since define.xml is machine readable, the file can be used by tools to help translate, visualize, summarize, and verify the data. The structure is very strict and must be followed to be compliant and compatible with established tools. By contrast, the information shared in the DRG is provided in a human readable format. The document is flexible in terms of its structure and the manner in which information is presented, allowing for visuals such as diagrams, tables, and figures and lengthier text which can help illustrate and describe more complex data flows or computational algorithms. It also provides a summary of data conformance findings.

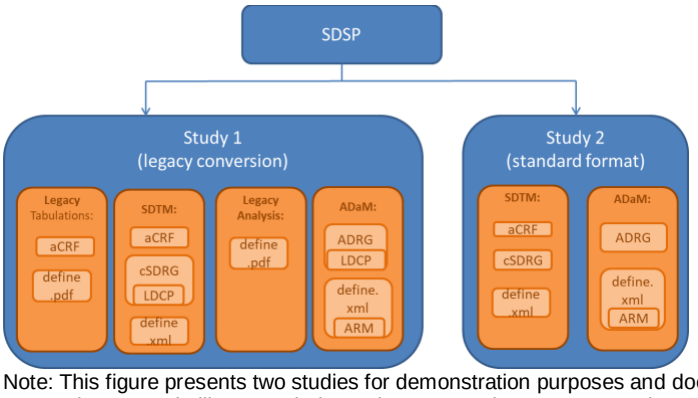
The define.xml and DRGs are highlighted in this paper, however additional documents are referenced in the TCG[4] which are required or recommended to support and describe the clinical trial data included in a regulatory submission, including the Study Data Standardization Plan (SDSP)[1] and the annotated Case Report Form (aCRF) (for clinical trials). Each study has its own set of documents, with exception of the SDSP, which is a program-level document (Figure 1[14]), listing all studies (clinical and nonclinical) with the data standards, versions utilized, and documenting any legacy data formats and any non-conformance to data standards (note that waivers must still be requested). The SDSP is an important tool for communicating standardization plans (including legacy data conversions or any up-versioning) to the FDA to ensure mutual understanding of the submission plans. Annotated CRFs should be provided along with the clinical trial tabulation datasets in a regulatory submission. This is a blank set of unique case report forms with annotations alongside each data collection field identifying the location of the data (dataset and variable name) in the SDTM or legacy tabulation datasets.

Some key information is intentionally duplicated between define.xml and the DRG so the reviewer can have easy access to it without having to shuffle between different files. Similarly, there is some intentional duplication of information in the SDRG and ADRG from the protocol and SAP. To make

navigation easier, when information is spread across documents, the define.xml may link to the DRG to provide additional explanation or information that does not fit into the define.xml structure. For clinical trials, the tabulations define.xml (or define.pdf for legacy data) also contains hyperlinks to the annotated CRFs for variables or values which were collected in the CRFs. Furthermore, ARM specifically links the Clinical Study Report (CSR), SAP, analysis datasets, and programming logic.

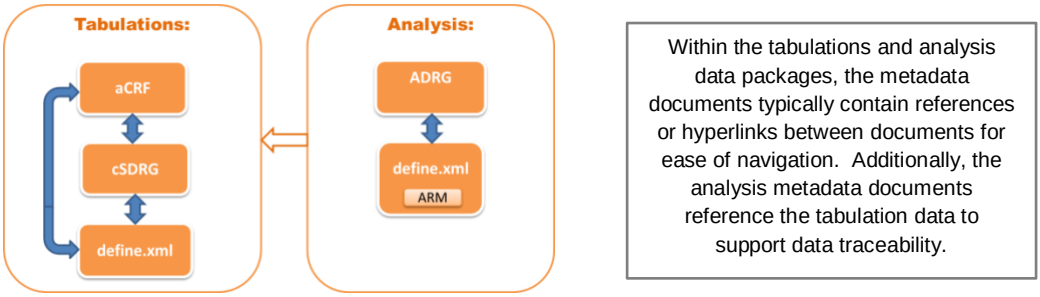
Within a study, the documents are highly related and each document will often hyperlink to or reference some of the other documents (Figure 2)[14]. Additionally, well-documented traceability will allow a reviewer to go from the study results to the analysis data to the tabulation data to the data collection sources, thus the analysis documents should relate back to the tabulations documents.

Figure 1: Overview of Study Document Organization



Note: This figure presents two studies for demonstration purposes and does not represent all possible study scenarios. The legacy conversion scenario illustrates independent conversion to SDTM and ADaM, as described in Table 4 of the TCG[4].

Figure 2: Relationships between Documents



Integrated data is not directly addressed in most guidance and specifications but may be submitted following a similar approach. If integrated analysis datasets are submitted, then an analysis define.xml file and ADRG should be provided to describe the metadata. The ADRG should provide appropriate explanations to ensure a reviewer can understand traceability from individual study level data and results. Likewise, if integrated tabulation datasets are submitted, then a tabulation define.xml and cSDRG should be provided. The SDSP should outline integration plans and is usually attached to a pre-submission briefing document or IND.

Summary of Key Information:

The table below identifies standard key information to include in the dataset metadata files and notes which files to include it in. Additional information, recommendations, and some informative examples are summarized below the table.

KEY: Y = yes (included), N = no (not included), O = optionally included

Topic	Description	define.xml*	cSDRG	ADRG
Protocol Information	Summary of protocol number, title, and design.	Y	Y	Y
Trial Design Data	Information pertaining to trial design datasets.	Y	Y	O
Acronyms	List of sponsor-specific acronyms.	N	Y	Y
Standards Versions	Summary of standards used for data and metadata.	Y	Y	Y
Dictionary Versions	Summary of external dictionaries used.	Y	Y	Y
Annotated CRFs	Summary of annotation conventions.	N**	Y	N
Study Data Overview	A high level summary of the submitted data.	N	Y	Y
Traceability/Data Flow	Information which describes the various data sources and how the datasets were created.	Y	Y	Y
Dataset Metadata	Descriptions of dataset contents and structures.	Y	Y	Y
Variable Metadata	Descriptions of variables in each dataset.	Y	O	O
Value-Level Metadata	Description of particular data subsets within vertically structured datasets.	Y	O	O
Codelists/Controlled Terminology	Description of possible values within a variable or data subset.	Y	O	O
Computational Algorithms	Descriptions of derivations.	Y	O	O
Comments	Notes and explanations.	Y	O	O
Analysis Results Metadata	Information on analysis datasets, programs, and results.	Y	N	O
Analysis Considerations	Summary of analysis rules and definitions.	Y	N	Y
Conformance / Compliance Findings	Summary of standards compliance checks and issues.	N	Y	Y
Legacy data conversion	Describes conversion from legacy to standard format.	N	O	O

* Note that there are separate define.xml files provided to describe each dataset package (i.e. one for SDTM and one for ADaM). Information pertaining to data collection and the tabulation data should reside in the SDTM define file and information pertaining to analyses or study results should reside in the ADaM define file.

** Note that the tabulation define.xml contains links to the annotated CRFs.

Protocol information is included in both the define.xml and DRGs. The define.xml includes the Study Name (a short external name assigned to the study; this may be the protocol number), the Study Description (a text description of the contents of the study from the protocol; this may be the protocol title), and Protocol Name (the sponsor's internal name assigned to the trial; this may also be the protocol number), whereas the DRGs include the protocol number and title and a summary of the different versions of the protocol, with the intention to highlight any changes that may have affected data collection or mapping, interpretation, or analysis. The DRGs also allow for graphics or text to describe the study design. It is recommended to include a flow diagram showing the high-level flow of treatments/phases from the protocol. This can often be copied directly from the protocol.

If there were protocol amendments during the trial, it is helpful to summarize any changes which impacted the datasets and not just list the amendment number and date for each version. Some of the changes that should be considered and addressed (note: this is not an exhaustive list):

- If inclusion/exclusion criteria were added, removed, or modified, how was this addressed in the IE and TI domains?
- If there were changes to study design, how did they impact the TS domain?
- If an assessment was added or removed and may not be present for all subjects/timepoints, please note this along with any conditions for its presence or absence.
- If there was a change in dosage or to a dosing regimen, note how this was implemented and is reflected in the data, as well as any impact to treatment group assignments or analyses.

If there were no amendments, it is advisable to include a statement to that effect rather than simply listing the protocol number. An example from the PhUSE cSDRG package[1] is shown here:

Example:

2.1 Protocol Number and Title

Protocol Number:	ABC123
Protocol Title:	PHASE I, SINGLE DOSE, OPEN-LABEL, DOSE ESCALATION PHARMACOKINETICS STUDY OF NEWDRUG IN HEALTHY SUBJECTS
Protocol Versions:	All subjects participated under the original protocol. There were no amendments.

Trial design data provides a quick overview of the trial design data and supports data warehouses for clinical trial data. It is included in both the define.xml and the DRGs. The define.xml describes the metadata (variables and their attributes) for the trial design datasets and lists the controlled terminology used in those datasets; however the codes may have minimal meaning without any other context. The cSDRG may describe the contents of those datasets and anything of special note concerning the study design or the use of the Trial Design domains in the submission. The ADRG may be used to describe how the trial design affected the structure and design of the analysis datasets and how the treatments and study phases or elements are identified within those datasets. Additionally, trial inclusion/exclusion criteria should be fully documented within the cSDRG to include the full text of all versions of criteria when there may be truncated or abbreviated text used in the TI (Trial Inclusion/Exclusion Criteria) domain or in cases where there were multiple versions and/or changes in criteria during the trial (note: per section 7.4

of SDTMIG v3.2, if criteria text is >200 characters, then the value of IETEST should be abbreviated with meaningful text and if the criteria is amended during the study, then additional IETESTCD values should be added to ti.xpt[15]).

Any **acronyms** that are used within the metadata documents should be listed in the DRGs. It is recommended to include a comprehensive set of acronyms, even if they are available in other documents. Study documents such as the protocol, CRFs, and SAP should be referenced to identify acronyms that are used within the study. The list should also include acronyms used in the metadata, such as within variable labels (particularly for SDTM custom domains and ADaM datasets) or controlled terminology or in the define.xml. This list should include anything that is not commonly known within the industry, is sponsor-specific, or is specific to a therapeutic area or indication. When in doubt, it is advisable to include the acronym. When the same acronym is used across documents, the definition/description should be the same in each document.

Standards versions identify the data standards (for example, CDASH, SDTM or ADaM) and metadata standards (i.e. Define-XML) models used, along with the version number of each model and its respective implementation guide. The models and versions are reported in both the define.xml and the DRGs. This information is needed for generating validation and/or compliance checks and for use with any machine tools. The FDA posts supported versions of each standard in the Data Standards Catalog[5].

Dictionary versions are included in both the define.xml and DRGs. All dictionary version information applied in the study database or during the data mapping should be included, such as the version dates/numbers of WHO Drug and MedDRA. In the define.xml, the dictionary version information is associated with the variables utilizing it. In the DRGs, it should be described in the Study Data Standards and Dictionary Inventory section. As versions may change during the course of the study, it is important to ensure the correct versions are listed within all documents.

The **annotated CRFs** are described in the cSDRG and referenced by the define.xml. The cSDRG can be used to provide additional information for the SDTM annotations such as a description of content organization and bookmarking, general annotation conventions used, or anything of special note - for example, an explanation of any variables which were annotated on the CRF but for which no data were submitted (e.g. if concomitant medications were collected but none were used in the study and thus none were entered in the study database). Legacy tabulation datasets, if submitted, should have a corresponding set of annotated CRFs[4]. Any notes regarding the legacy dataset annotations may also be included in the cSDRG, especially if any of the data could not be mapped to SDTM.

The define.xml file should contain links to the SDTM aCRFs whenever a variable or value has an origin of "CRF". The Origin column in the define.xml should provide a hyperlink to the unique page(s) in the annotated CRF where the source variable or value is annotated. This helps to support data traceability. It is important to annotate for the data package being submitted (e.g. referencing SDTM domains and variables when submitting SDTM data) and not include internal/raw data references. Any data fields which are not submitted should be clearly annotated accordingly[4].

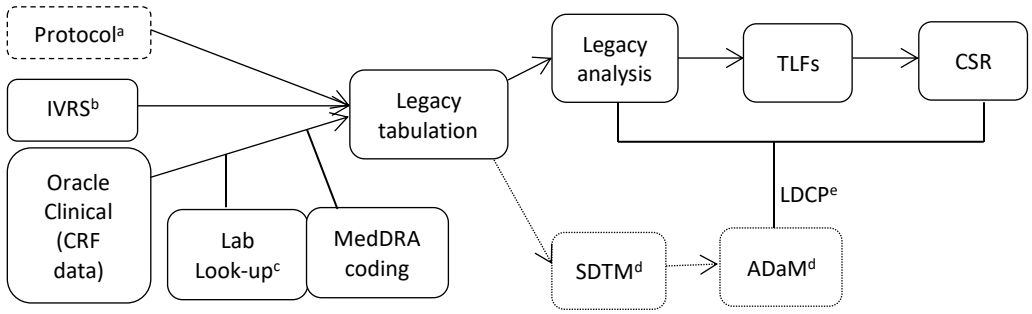
Study data overview information is included in each DRG. The DRG templates provide specific "overview" sections in which to provide a general description of what data has been included in the submission and where to locate specific data points of interest. This provides the reviewer with a summary of what types of data were submitted and not submitted (if any data was omitted or not available, and why), the sources of the data, and if there was any interim data cut used. A description of

how the treatment-emergent events are identified, any general rules for baseline flagging, plus descriptions of all reference start and end dates and how they are used in the various datasets may also be helpful to reviewers[9].

Traceability information is included in both the define.xml and the DRGs. This information is key to being able to reproduce and verify the study results. The define.xml file should identify the source(s) for each variable and value - whether from a CRF, external data file, or that it was assigned or derived. Variables or values with origins of "Derived" or "Assigned" should additionally include information on how they were derived or assigned. Computational algorithms should provide sufficient detail that a reviewer could recreate the values.

Within the DRGs, it is helpful to include one or more data flow diagrams (see examples below), illustrating the various data sources and how they were used to generate the standardized datasets as well as the study results, including any intermediate or reference datasets. The cSDRG template[1] now includes a specific section for this.

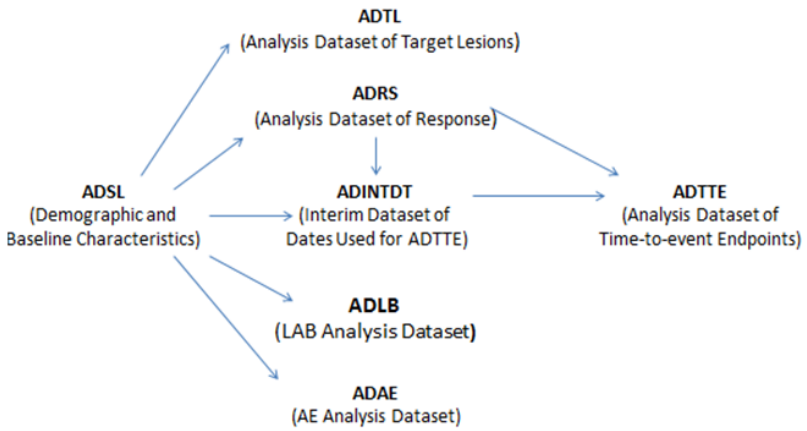
Data Flow Example 1:



a. The protocol is used to populate the Trial Design domains and for dosing information in the EX domain.
b. The IVRS data is used to assign the treatment arms.
c. A lab look-up table is referenced for application of standard units and conversion factors to local lab data from the Oracle Clinical database.
d. The original study analyses and CSR were generated using the legacy analysis datasets. The data was later converted to SDTM and ADaM in sequence.
e. The Legacy Data Conversion Plan and Report describes any differences between the original analysis datasets and study results compared to the ADaM datasets.

It is helpful to include diagrams in the ADRG's Analysis Data Creation and Processing Issues section to describe any dataset dependencies and/or intermediate datasets:

Data Flow Example 2:



To aid in traceability, any special data processing should be described in the DRGs, such as:

- Legacy data conversions (in the LDCP appendix of the cSDRG or ADRG)
- When SDTM was not the primary source used in the creation of ADaM data
- If there are lengthy or complex computational algorithms which do not fit well within the define.xml file (optimally with a reference from the define.xml to the DRG)
- Use of look-up tables to flag or transform data for standardization or analysis
- Look-up tables which were used (for example) for mapping of controlled terms from original data to standardized data or for converting data values from original units to standardized units (see below for an example)

SDTM Example:

Source Test Code	Source Test Name	Source Orig. Units	SDTM LBTESTCD	SDTM LBTEST	SDTM LBORRESU	SDTM LBSTRESU	Conversion Factor
ALBU	ALBUMIN	g/L	ALB	Albumin	g/L	g/L	1
ALP	ALKALINE PHOSPHATASE	IU/L	ALP	Alkaline Phosphatase	IU/L	U/L	1

ADaM Example:

PARAM	PARCAT1	Associated Event	Hypo/ Hyper	ATOXGR= 1	ATOXGR= 2	ATOXGR= 3	ATOXGR= 4
Albumin (g/L)	Chemistry	Hypoalbuminemia	HYPO	<LLN - 30	<30 - 20	<20	Not applicable
Alkaline phosphatase (IU/L)	Chemistry	Alkaline phosphatase increased	HYPER	>ULN - 3.0 x ULN	>3.0 - 5.0 x ULN	>5.0 - 20.0 x ULN	>20.0 x ULN

Dataset metadata is presented in both the define.xml and the DRGs, however there is different information presented within each document. The list of datasets in the define.xml summarizes the attributes of each dataset: label, purpose, class, key variables, etc. Only metadata for datasets which were actually submitted should be included in the define.xml. Any split datasets should have an entry for each dataset along with a description for how the data was split. Empty datasets should not be submitted nor documented in the define.xml, but can be noted in the DRG. Additional information may be provided in the "Documentation" column within the dataset-level metadata table in the define.xml, however anything lengthy which would be difficult to read in the define.xml format is better placed in the DRG. A hyperlink from define.xml to the DRG can be included in those cases.

The table that lists the datasets in the DRG includes the label, and optionally the class of each dataset. Split datasets should be listed separately and a description for how the data was split should be provided in the dataset summary paragraphs (described after the examples). There is indication whether safety and/or efficacy data are present in each SDTM or ADaM dataset. This table also informs with a quick glance if there is any related data within a SUPPQUAL (Supplemental Qualifiers) or RELREC (Related Records) domain (for SDTM data). We recommend including a column to identify custom domains in the cSDRG (note: this has been added in the Nov. 2018 release of the cSDRG template).

SDTM Example:

Dataset – Dataset Label	Efficacy	Safety	Other	Custom	SUPP-	Related Using RELREC	Observation Class*
AE – Adverse Events		X				DS	Events
DM – Demographics			X				Special Purpose
DS – Disposition			X			AE	Events
EX – Exposure			X				Interventions
XA – Event Adjudication	X			X			Findings

*Observation Class column has been removed as of the Nov. 2018 cSDRG template, however it is helpful to include to identify the structure for custom domains. If not included, it is advised to note the dataset class within the summary paragraph describing the custom domain.

ADaM Example:

Dataset Dataset Label	Class	Efficacy	Safety	Baseline or other subject characteristics	PK/PD	Primary Objective	Structure
ADSL Subject Level Analysis Dataset	ADSL			X			One record per subject
ADEFF Efficacy Analysis Dataset	OTHER	X				X	One record per subject
ADVS Vital Signs Analysis Dataset	BDS		X				One record per subject per parameter per visit per timepoint

The DRG allows the sponsor to expand upon the basic information in the tabular list of datasets by allowing inclusion of a hyperlink to a separate paragraph in which the sponsor can describe, address, or highlight anything additional they choose. Any datasets containing efficacy or safety data of interest for the study (e.g. reactogenicity events for vaccine studies or tumor measurements for oncology studies) should have summaries here. The ADRG should specifically indicate which datasets and variables contain the primary versus secondary efficacy endpoints and a description of their derivations. The text should highlight where to find parameters of interest, even specifying the categories, test codes, or other criteria which can be used to filter for those records. Examples from the cSDRG and ADRG packages[2] are shown here:

cSDRG Example:

3.3.7. RS – Response

The investigator's assessment of disease response is identified by RSEVAL equal to INVESTIGATOR. The IRC's assessment of disease response is identified by RSEVAL equal to INDEPENDENT ASSESSOR and RSACPTFL equal to Y.

ADRG Example:

5.2.2 ADTTE – Time to Event Analysis Dataset

The time to event dataset was used to support the primary study endpoints. It followed standard ADaM conventions for a TTE file. The events of interest were overall survival (paramcd='OS') and progression free survival (paramcd='PFS').

This is also a good place to explain any exceptions to rules/criteria outlined in the standards or protocol, any outliers or gaps in the data, or anything which may be unexpected in any of the datasets. It is highly recommended to include summaries for any custom SDTM domains, in order to provide a reviewer with information pertaining to the contents of the non-standard datasets which otherwise may not be clear based on the dataset naming and properties alone. The structure, contents, and any special terminology used should be noted. It may be helpful to include a table (similar to that used to list any supplemental

qualifiers – see value-level metadata information below) listing the tests or assessments included in the domain (for custom Findings domains) if the list is not overly long.

Additionally, it may be helpful to provide a summary of the types of relationships included in the RELREC domain in the cSDRG, if included in the submission.

Example:

Related Domains	Description of Relationship
AE, CM	Any medications taken in relation to an adverse event are linked to the relevant AE.
AE, MH	If an adverse event represents a worsening of a pre-existing condition, it is linked to the reported condition from the subject's medical history records.

Variable metadata exists primarily in the define.xml and lists the variables included in each dataset, along with the attributes of each variable: label, type, length/display format, controlled terminology, source (origin), as well as any sponsor-provided comments and any derivation algorithms. The information is presented by dataset and every variable present in the dataset must be listed and described in the define.xml. The variables and their attributes should adhere to the specified data standard and version, including the variable label, type, and code list. Additionally, the origin of the variable must be specified to indicate the source(s) of the data – whether CRF(s), protocol, assigned/derived, or from an external data source. If a variable has multiple origins across test names, qualifiers, parameters, or other criteria, this can be split out in the value-level metadata (see below). In all other cases, the sponsor should select a primary origin and include additional origin information in the Derivation/Comment column. When originating from CRF data, hyperlinks to the specific CRF page(s) in the aCRF file should be included. The DRGs may describe variables that are of particular importance to highlight to the reviewer, such as core/common variables applied across the datasets, treatment variables, subgrouping variables supporting analyses, etc. and may be used to provide additional details or clarity for more complex mappings or derivations.

Value-level metadata also resides primarily in the define.xml and enables the specification of information when a set of variables (typically in a vertically-structured dataset) is used for multiple/different tests, procedures, assessments, qualifiers, or other criteria, hereafter referred to as "subsets". This information is included in the define.xml on a subset-specific level to help a reviewer better understand and know what to expect for the various subsets of that set of variables. Any relevant formatting and code lists specific to that subset can be noted, along with the subset-specific origins. In the LB (laboratory test results) domain for example, results may come from both CRFs and external data transfers, but the value-level metadata notes "CRF" or "eDT" as the source for each individual test (a single origin should be listed for each subset, or test in this case). Value-level metadata is particularly helpful for the non-standard information provided in the form of SUPPQUAL in SDTM.

Additional information on supplemental qualifiers, or non-standard variables, in the SUPPQUAL domains may also be provided in the cSDRG. A table for this is included in the template. However, the tables there only include the QNAM/QLABEL combinations in order to give a reviewer an idea of the additional information available for each domain. One must go to the define.xml to obtain the full metadata for those additional qualifiers. If a sponsor includes the SUPPQUAL tables in the cSDRG, it is recommended to

include an additional column (similar to the one included in the CBER SDSP Appendix[1]) in each of the tables to provide additional explanation of the qualifier. This will help convey further information as to the meaning, relevance, and importance of the qualifiers the sponsor has chosen to include in the datasets.

Example:

QNAM	Description	Corresponding CRF Question or Derivation information
CECLTYP	Collection Type	Requested by CBER OVRR to identify source of data at the record level as part of chain of custody documentation. Values will be "CRF" or "DIARY" to denote the source of each specific value.

Codelists/controlled terminology is listed in the define.xml and information on CT sources and versions is provided in both the define.xml and the DRGs. Lists of the possible values for a particular variable or value for the study are included in the define.xml file with hyperlinks from the associated variables and/or values that utilize the list to the corresponding controlled terminology list. These lists should utilize standard controlled terminology wherever possible, with the source of the terminology documented in the XML. In addition to the codes/decodes and source information, the code list type will indicate whether the codes are numeric or text and the codes can be ordered for display purposes and/or be ranked to indicate importance, severity, etc. Any sponsor-specific terminology (custom code lists or extensions to CDISC CT[16]) or deviations from CDISC CT should be explained in the DRGs. Additionally, any issues with mapping legacy data to controlled terminology should be noted within the LDCP in the cSDRG and/or ADRG.

For **computational algorithms**, there is a dedicated section of the define.xml used to describe the derivation of variables. A computational algorithm should be provided for any variable or value which is derived from other data[3]. The algorithm should be clearly written in English language and avoid any programming lingo or code. If there are variables that are used across multiple datasets, they should have a consistently worded derivation algorithm. There may be situations where the derivation is complicated, contains special characters, or the description of the derivation is very long. It can be helpful to document complex logic and lengthy descriptions in the DRGs where it may be easier to read and visual diagrams may be used when beneficial (see example below). This is also true for some integrated safety or efficacy analysis variables where the derivation is not only complex but also involves multiple contributing studies. Variable derivations in the DRGs can be hyperlinked or referenced from the corresponding variables in the define.xml**.**

Example:

Calculation of Mean Daily Dose

a. the sum across each study treatment dosing record (except those from the titration period) of [(stop date - start date + 1) * number of capsules taken] * mgs per capsule
   divided by
b. (the last study treatment date - the first study treatment date + 1 - the length of the titration period).

Comments (variable-level or value-level) are included in the define.xml, which help to provide additional information or context beyond what can be found in the implementation guides. This is particularly useful for sponsor-, program-, or study-specific implementation decisions and especially when the usage requirements for a particular variable are not specifically defined by the standards model and guide. The define.xml comments section is best utilized for shorter, succinct comments which can be expressed simply and within 1-2 short sentences. Longer or more detailed or complex comments are generally better presented in the DRGs where the layout is more flexible and paragraphs and visuals can be utilized. In these cases, the define.xml comment field can be used to refer the reviewer to the specific section of the DRG (it is helpful to include a hyperlink) where additional information can be found (see example below).

Example:

define.xml:

Variable	Label	Type	Length / Display Format	Controlled Terms or Format	Source/Derivation/Comment
DRGCOMP	Drug Compliance (%)	float	8		Derived: Calculated based on drug formulation, age group, and weight. Actual total dose received / Expected total dose * 100. See ADRG section 5.2.1 for further details.

ADRG:

5.2.1 ADSL – Subject Level Analysis Dataset

Drug compliance was calculated based on drug formulation (tablets or oral solution), age group, and subject weight at baseline, where Drug X was provided as 5 mg tablets and as 0.4 mg/mL solution for oral administration:

Age	5 years - < 18 years				3 months - < 5 years			
Weight	≥ 35 kg		< 35 kg		> 35 kg	<35 to 18 kg	<18 to 9 kg	< 9 kg
Formulation	Oral Solution	Tablets	Oral Solution	Tablets	Oral Solution	Oral Solution	Oral Solution	Oral Solution
Days 1-7	10 mg twice daily	Two 5 mg tablets twice daily	8 mg twice daily	N/A	10 mg twice daily	8 mg twice daily	6 mg twice daily	4 mg twice daily
Days 8 and on	5 mg twice daily	One 5 mg tablet twice daily	4 mg twice daily	N/A	5 mg twice daily	4 mg twice daily	3 mg twice daily	2 mg twice daily

Analysis Results Metadata (ARM) is an important component of data traceability, providing information on variables/parameters used in the analysis, analytical methods and procedures, and linking results, data, and documentation. The ADaM define.xml has a specific section (which is highly recommended to support traceability from data to study results) for this information. Additional information (beyond what is

in define.xml) could be provided in the ADRG if needed, however there is not a single specific section for this in the ADRG template as there is in the define.xml structure. Traceability information is provided in the ADRG when noting the data sources used for ADaM, within the section comparing SDTM and ADaM, and when describing processing and imputations/derivations. Additionally, the "Submission of Programs" section is required when including any programs in the submission. Optimally, the list of programs should identity the corresponding inputs and outputs in order to provide the complete linkage from data to results (particularly if Define-XML ARM is not provided).

Analysis considerations can be included in define.xml and/or the ADRG. While the define.xml includes the computational algorithms (and potentially comments) for the derived variables within a dataset, the ADRG provides a section in which to include a summary of analysis rules or considerations applied across the datasets (such as baseline definitions and visit windows), highlighting differences between the SDTM and ADaM datasets (e.g. differences between baseline flags), identifying common (or core) variables across datasets, and describing how treatment variables were used in the analysis datasets. Any key analysis rules or definitions should be documented within this section. Some information from the protocol or SAP may need to be repeated so that this important information is readily available, rather than having to search through multiple external documents. When information is exactly duplicated from a source document, a link to the source document can be provided instead of (or in addition to) copying the text (as long as the document is included in the submission).

Conformance/compliance findings (results of checking the datasets against a set of validation rules) are documented only in the DRGs. A data conformance summary section is present in both the cSDRG and the ADRG templates[1] with a table format for listing and documenting compliance findings. Findings which cannot be corrected in the datasets should have explanations provided. Severity levels of "Error" and "Warning" from the FDA validation checks should be documented in the DRGs, but "Notice" level findings do not necessarily need to be included unless there is an impact to the results or reviewability of the data. It may be helpful to include the relevant rule ID by inserting a column in the issues summary table from the template (see "FDA ID" column in the example below).

It is important to provide a complete and sufficiently detailed explanation for each finding. Generic responses such as "per the data" or "data is locked" are not sufficient. Information on why the issue is present and why it could not be resolved should be provided. For subject or site level findings, the specific subject and/or site numbers should be noted. For example, a good explanation for a missing expected data element should outline that the data point was not collected at the site for a particular subject, rather than just acknowledging the data point as missing, or stating that the data element was not used in the analysis.

Insufficient explanation:

Rule ID	***FDA ID***	Dataset	Diagnostic Message	Severity	Count	Explanation
SD0021	FDAC117	AE	Missing End Time-Point value	Warning	1	The value is missing in the database.

Improved explanation:

Rule ID	***FDA ID***	Dataset	Diagnostic Message	Severity	Count	Explanation
SD0021	FDAC117	AE	Missing End Time-Point value	Warning	1	The event for subject 1111-001_123456789 is still ongoing as of the interim data cut, therefore AEENRF is missing. The value will be entered prior to final database lock.

Additional recommendations and examples can be found in the paper "Best Practice for Explaining Validation Results in the Study Data Reviewer's Guide" by Kristin Kelly[17].

Findings for the define.xml file should also be documented in the compliance summary of the corresponding DRG (in addition to dataset compliance findings).

Legacy data conversion information is provided in the LDCP appendix within the cSDRG and/or the ADRG, depending on which types of data were converted. It is important to clearly document the flow of data and at what point the conversion occurred, as well as which datasets were the sources for the study results. Any issues encountered during the conversion or differences between the legacy and standardized data should be described along with any decisions made to address them or any resulting compliance findings.

Conversion Scenario	**LDCP(s) Needed**
Converted legacy tabulation to SDTM*	LDCP in cSDRG
Converted legacy analysis to ADaM*	LDCP in ADRG
Converted legacy tabulation to SDTM which was then used as source for ADaM	LDCPs in cSDRG and ADRG

*If both legacy tabulation and analysis datasets are converted independently to SDTM and ADaM, then both LDCPs are required.

Where to Document Other Important Information:

If there is additional information that would be useful to help to understand the data (i.e. something that would not be readily apparent to someone who is unfamiliar with the study and its datasets) but is not specifically included in the sections of the DRG or the define.xml, or if it is unclear whether to include the information in one or both documents, then the team should evaluate the impact and scope to identify where best to include the information.

It is important to document anything which may be unexpected or not easily found in the data. Any special situations should be clearly documented in the spirit of transparency. Anything of special note in the data can be described within the comments in the define.xml file, however longer or more detailed comments/algorithms or anything requiring complex explanations are better presented in the DRGs where the layout is more flexible and paragraphs and visual aids can be utilized.

Additionally, if there is information pertaining to key safety or efficacy data, there is a large amount of data affected, or there is impact to the analyses or interpretation of the data, then best practice is to note it within the DRGs. It should be written in a concise but informative manner and, if possible, include links from the define.xml where relevant to the corresponding section in the DRG.

Information specific to a variable or value:

- Best practice: This may best be noted in the comments and/or computational algorithm for that particular variable or value, however if the notes are lengthy or a visual of some sort is helpful, then the information can be included in the DRGs. Notes related to mapping or other programming decisions can be explained in the DRGs while items which are very apparent in the data (e.g. outliers, missing data) may be commented on in both the define.xml and the DRGs. Furthermore, if there is any impact to analyses, it is best to describe what happened in the DRGs (including the potential impact on study analyses and results).
    - Example 1: description of how a specific xxSPID has been assigned
        - Describe in define.xml comments
    - Example 2: certain subjects received the wrong study drug, hence DM.ACTARM (actual treatment arm) differs from DM.ARM (planned treatment arm)
        - Provide explanations and indicate impact on analyses in the DRGs
    - Example 3: test results from one site could not be used due to instrumentation issues
        - Document in DRG and define.xml

Information specific to a dataset:

- Best practice: Define-XML allows for some documentation for each dataset, however anything lengthy and/or detailed should be included in the DRG. Section 3.4.x in the cSDRG or 5.2.x in the ADRG is useful for providing additional information for a dataset. If there is something important to call attention to, it is advised to include that information in both the define.xml and the DRG.
    - Example 1: derivation of exposure from combination of collected data and protocol
        - This is best explained in the DRG.
    - Example 2: description of custom domain contents or how a dataset was split
        - A brief description may be provided in the dataset-level table in the define.xml with more detailed notes provided in the DRG.

Information that affects or applies to multiple datasets:

- Best Practice: This is best addressed in the DRGs, however links from the documentation column of the dataset table in define.xml to this information is helpful. Section 3.1 in the cSDRG or 5.1 in the ADRG (Additional Content of Interest area) can be used to describe the issue in general and then references to this from the specific dataset summaries in section 3.4.x (cSDRG) or 5.2.x (ADRG) may be included as well.
    - Example 1: for maternal immunization studies, fetal exposure is derived from the mother's exposure data
        - The Documentation column in define.xml may be used to indicate where the mother's and fetus's exposure resides (e.g. for EC domain: "Vaccination data collected for maternal subjects" and for EX: "Derived maternal and fetal exposure data"). The cSDRG can be used to describe what data was collected from the sites and how the fetal exposure records were derived.

In general, if something does not "fit" in the define.xml (due to the strict requirements of the Define-XML structure and contents) or is not easy to read there, then it is suggested to include it in the DRGs (note that something may easily apply to multiple sections of the DRGs and should be included in all of the relevant sections, at least in the form of a reference to another section, in those cases). Useful sections (section numbers are based on the current templates as of this writing) for including additional information include:

- Section 2 (**Protocol Description**)– anything related to protocol/trial design
- Sections 3.1/3.3 in cSDRG (**Subject Data Description – Overview and Annotated CRFs**) or Sections 4 through 5.1 in ADRG (**Analysis Data Creation and Processing Issues, Analysis Dataset Descriptions**) – anything related to data collection/extraction/processing, data cuts (particularly for interim data cuts for an ongoing trial), or key safety/efficacy data
    - Note: the "**Additional Content of Interest**" subsection under section 3.1 in the cSDRG is a good area to make use of for items which may not fit well into the other sections
- Section 3.4 in cSDRG (**SDTM Subject Domains**) or Section 5.2 in ADRG (**Analysis Datasets**) – items related to specific datasets
- Section 4 in cSDRG or Section 6 in ADRG (**Data Conformance Summary**) – anything related to compliance/validation against standards and submission requirements
- Alternately, if there is a lot of information to include or large or many visuals, then an appendix could be added at the end of the DRG to contain this information.

In terms of whether to include the information in the cSDRG and/or corresponding define.xml (tabulation metadata files) or the ADRG and/or corresponding define.xml (analysis metadata files), one should consider the following:

If the information….	Include in Tabulation Metadata Files	Include in Analysis Metadata Files
Impacts analyses or interpretation of study results and applies to (or originates from) the tabulation datasets	X	X
Is simply a collection/tabulation issue and does not affect the analyses	X	
Is highlighting a computational algorithm or derivation but there is nothing special to bring to attention to for the source tabulation data		X

When in doubt, we recommend to document the information in both sets of files.

Consideration of Differences between Agencies and Divisions:

It is worth noting that there are different requirements and expectations within (e.g. FDA CDER versus CBER) and between regulatory agencies (e.g. FDA, EMEA, PMDA, etc.). It is recommended that agency-specific considerations are researched and addressed within the datasets and the metadata documents as early as possible prior to any regulatory submission. The U.S. and Japanese authorities currently have the most developed data standards requirements[18,19], with the U.S. requiring standardized data for any studies starting after Dec. 17, 2016[6] and Japan requiring for studies submitted after March 2020[20], and are highlighted here. It is expected that other countries will begin to require standardized data and metadata documents in the near future. There are various points to consider when planning submissions and some important considerations are reviewed here. Most important is ensuring adherence to the data standards requirements, including supported versions of standards and any related deliverables which are required.

There a few differences worth noting between FDA divisions. One concerns providing data to support site selection for inspection planning. Guidance from CDER requests Office of Scientific Investigation (BIMO) deliverables and requires a summary level clinical site dataset to be provided[21]. CBER does not request this deliverable but instead may ask that the SITEID variable be included in all SDTM dataset domains[22]. This conflicts with the SDTM data standard in which SITEID is included only in the DM (demographics) domain and introduces compliance findings which would need to be explained in the cSDRG.

There is another important difference related to the SDSP. Both CDER and CBER expect a SDSP to be provided, however CBER additionally requests an appendix[1, 23] which includes tables of proposed SDTM domain/variable usage, supplemental domain usage, and proposed analysis datasets for each study, as well as the planned datasets for any integrated analyses (ISS/ISE). For submissions to Japan, the PMDA requires a document which is similar to the SDSP (referenced as Form 8, Attachment 8, or Appendix 8; available only in Japanese) for discussion during the Consultation on Data Format of the Submission of Electronic Study Data meeting[20,24,25]. This meeting is held prior to the pre-NDA meeting to focus on the electronic data submission requirements. Sponsors are expected to be aware of data compliance findings and prepared to discuss them with the agency at this meeting.

Data validation rules and severities, as well as recommended validation tools and versions may differ between countries. For example, FDA uses two severity codes (Warning and Error) to classify conformance findings whereas PMDA uses three severity codes (Warning, Error, Reject). As mentioned earlier in this paper, conformance findings that cannot be corrected must be addressed in the DRGs, however for PMDA a severity level of Reject will cause the review process to be suspended until corrections have been made and the issue is resolved. PMDA expects sponsors to execute validation checks prior to the Sponsor Consultation Meeting and review severity codes of Error at the meeting to determine whether the issues will affect the review. If agreed upon, the errors should be documented in the Conformance Section of the DRGs and will have no impact to the review cycle. Note that a severity of Error which has not been discussed with the agency in advance of the submission will cause the review to be suspended until corrections have been made[26,27]. Similarly, the FDA has technical rejection criteria[28] which, if violated, could result in a Refuse to File (RTF) or Refuse to Receive (RTR).

Another example concerns the Analysis Results Metadata (ARM) for ADaM. FDA does not as yet require the ARM in submissions although it is acceptable to include it either within the ADaM define.xml or as a separate file in PDF format. PMDA desires the ARM to be included in submissions and prefers that the ARM be included as a section within the ADaM define.xml[26], following the Analysis Results Metadata specification[13], rather than a separate file in PDF format. They additionally ask for the programs used to create the ADaM datasets and the tables, listings, and figures (TLFs)[26,27] whereas the FDA requests the programs to create ADaM datasets and tables and figures which summarize the primary and secondary efficacy analyses[4]. A table of programs that include outputs, inputs, and macros used is suggested for inclusion in the ADRG.

When preparing metadata documents, ensure the acceptability of English or the appropriate language for the country of submission. The DRGs should describe language translations needed to ensure understanding of the information provided. If data is provided in a language other than English, an extended character set may be needed. It is useful to include the character set used on your system for the SAS datasets, e.g. ENCODING = Latin1 for Western (ISO), in the appropriate DRG within the table in section 1.3 that outlines the standards used.

Due to differences noted above, plus others, sponsors may need to create different submission packages with slightly different content when submitting to multiple regulatory authorities. Care should be taken to ensure that information remains accurate and consistent between the different versions while meeting the various country requirements. Sponsors should also be aware of the file naming conventions to be used within the eCTD folder structures between countries. For example, different naming conventions are required for FDA versus PMDA for the DRGs[4, 26].

Recommendations for Submitting Quality Documents:

Prior to submitting the metadata documents to a regulatory agency, several levels of review are recommended to check for technical functionality, adherence to technical specifications, accuracy and completeness of content, and cross-document consistency.

It is recommended to run the define.xml through Pinnacle21 Validator[29] (or similar tool) on its own and with the corresponding datasets. It is best to use the latest available version of the tool, as checks are regularly added and updated, unless a specific version is required by the reviewing agency. You should

be sure to select the appropriate configuration file (set of validation checks) based on the data standard and version used, as well as the intended reviewing agency. Findings from the tool should be reviewed and addressed (either through correcting the issue or providing an appropriately detailed explanation within the DRG's data conformance summary section). Explanations should be reserved for issues which cannot be corrected due to how data was collected, inability to contact a site, and such reasons; otherwise the issue should be corrected. If a sponsor regularly experiences certain findings due to the nature of their studies, data collection, etc., it may be helpful to institute some standard language which teams may use to provide explanations within the conformance summaries (and/or other sections of the DRGs or define.xml files if helpful and appropriate). It is important to note that certain issues could lead to a refusal to file from the agency (see previous section).

When using the PhUSE templates for the DRGs, it is recommended to maintain all sections from the templates and their numbering. If any particular section is not relevant/not used, then it is best to keep the section in the document and just include a statement explaining why it is not applicable. Alternately, if there is additional material to include which warrants a new section, it can be added at the end or as an appendix. We recommend that teams start populating the DRG contents as early as possible to ensure that items which may impact review get documented as they come up (while details are still fresh in your mind) rather than trying to recall everything at the end.

Any references within the documents should be verified and bookmarks and hyperlinks checked to ensure that they are functioning correctly (and not only working, but pointing to the correct location). Additionally, the define.xml file should be opened in multiple browsers and browser versions to ensure that it renders correctly.

For both the define files and the DRGs, the contents of the files should be reviewed by the team members who are familiar with the data and analyses, to ensure that the material and information is accurate and complete and it reflects the data which is being submitted (i.e. does not include any references to data which is not included in the submission, with exception of documenting any domains which were planned but not submitted in the data overview section of the cSDRG). An additional review by someone who is unfamiliar with the study is also recommended to check that everything is clear and makes sense. Internal information (such as names of people or references to other files [including raw data files] or other studies which are not part of the submission, change logs, etc.), programming/mapping specifications, and abbreviations which would not be known across the industry should not be used within these documents. Care should be taken if copying specifications or documentation from another study. Snippets of programming logic may be included if the team thinks it would be helpful to a reviewer, but the information should also be provided in plain language so that it can be understood by anyone using the document.

Information should provide sufficient detail to be able to understand the data, without being overly wordy and confusing. It can be difficult to judge the appropriate level of information to include, so having multiple reviewers for the document can help to gauge if the level of information and detail is appropriate. It is important to stick to the facts and not provide extraneous information (ask yourself if it is something a regulatory reviewer needs to know).

Example with Extraneous Information:

Variable	Label	Type	Length / Display Format	Controlled Terms or Format	Source/Derivation/Comment
EFFEVT	Efficacy Event (Y/N)	text	1	["N" = "No", "Y" = "Yes"] <No Yes Response>	If VTESTAT.THRMSTT in (1, 2) and VISITN = 84 then EFFEVT = 'Y'. Else 'N'. UDPATED 01JAN2018: per Stewart, we need to change this to check the status at last available visit as some subjects may not have imaging done at Day 84. E.g. subject 8274-2847's last image is at Day 63, so we need to check that visit in that case. Also, need to exclude subjects who do not have any post-baseline imaging.

Improved Version:

Variable	Label	Type	Length / Display Format	Controlled Terms or Format	Source/Derivation/Comment
EFFEVT	Efficacy Event (Y/N)	text	1	["N" = "No", "Y" = "Yes"] <No Yes Response>	Sort XA domain by USUBJID, XADTC and keep the last available date where XATESTCD = "THRMSTT". If XA.XASTRESC is "REGRESSION" or "UNCHANGED" then EFFEVT = "Y". Else if XA.VISITDY > 0 then EFFEVT = "N". Else if XA.VISITDY <= 0 then EFFEVT = " " (null).

Along with each document being reviewed as an individual document, there should also be some level of cross-document review, to ensure that any repeated information is consistent, optimally in terms of both content/meaning and phrasing/verbiage; however there is no need to repeat all content within every document, so teams should try to limit any repetition to key information. It is also recommended to use a consistent voice (active or passive), point of view (first-, second-, or third-person) and tense (i.e. "was" versus "is") throughout and across documents. Furthermore, there should be alignment between the SDTM/tabulation documents and the ADaM/analysis documents - they should not come across as being written by two entirely different, independent teams, but should have a similar layout, definitions/algorithms, and use of language. Similar checks should be applied across all of the submission metadata documents (these are not limited to define.xml and the DRGs).

A "clean" set of documents should be generated following review – with correct spelling and grammar and with any edits, tracked changes, comments, etc. finalized prior to submission. If there are updates to the define.xml, it should subsequently be re-run through the validation tool to ensure that no new issues have arisen. Be sure to conduct a final review of the entire data package to verify that everything is truly "submission-ready". Final documents should be polished and professional, including formatting checks for pagination, indentation, spacing, and font size, color and effects.

Conclusion:

Health authority data standards resource pages should be reviewed to become familiar with the applicable regulations, guidance, and accepted standards. Additionally, industry organizations such as CDISC and PhUSE are actively working to create and maintain data and documentation standards, templates, examples, and completion guidelines in support of initiatives to streamline regulatory reviews. The dataset metadata documents mentioned in this paper are essential submission components to support the regulatory review process. This documentation helps reviewers, and any users who are unfamiliar with the study details, to more quickly understand and work with the study data. Preparing effective, quality dataset metadata documents can be difficult – knowing how to complete sections of the documents, what information to put in each document, and the level of detail to include are often a challenge because this level of guidance is not provided in current standards, guidelines and templates. The best practices presented in this paper promote the creation of sufficiently detailed, quality dataset metadata, utilizing the appropriate documents and components in order to provide complete and clear information to the reviewers.

Remember that the templates are just a starting point and can be added to as needed. Be sure to include any additional information that would not be readily apparent to someone who is unfamiliar with the study and its datasets and document anything which may be unexpected or difficult to find in the data. Special situations should be clearly documented in the spirit of transparency and traceability.

Although the pre-specified sections and completion guidelines are provided to help you include the appropriate information, one should avoid completing only the minimal required elements but instead fully consider what extent of information is needed to tell the story of the study and understand the data. Preparing these documents should not be a matter of "going through the motions", but rather requires thought and consideration towards your end users and how the documents will be used.

Since these deliverables are complex and highly related, a final point to stress is the need to implement a quality review to ensure technical functionality, readability, compliance, and consistency within and between the documents. Ultimately, the documents should work together to help a reviewer navigate through the data and to communicate key information.

Disclaimer:

The opinions expressed in this document are those of the authors and do not necessarily represent the opinions of PhUSE, members' respective companies or organizations, or regulatory authorities. The content in this document should not be interpreted as a data standard and/or information required by regulatory authorities. Additionally, the authors do not endorse any specific commercial products or services.

All information is believed current and accurate at the time of the writing of this paper however is subject to change as regulations, statutes, and guidance are updated or replaced. It is recommended to discuss submission plans with the reviewing agency in advance of any regulatory submission.

References:

[1] PhUSE. *CS Final Deliverables Catalog - cSDRG Package, ADRG Package, SDSP Package* (current versions at time of publishing – cSDRG template version November 2018, ADRG template version January 2015, SDSP template version January 2018). Available at: https://www.phuse.eu/css-deliverables

[2] Clinical Data Interchange Standards Consortium (CDISC): https://www.cdisc.org/

[3] Clinical Data Interchange Standards (CDISC). *Define-XML Specification* (current version at time of publishing – v2.0/April 2014). Available at: https://www.cdisc.org/standards/data-exchange/define-xml

[4] U.S. Food & Drug Administration (FDA). *FDA Technical Conformance Guide* (current version at time of publishing – v4.2.1/January 2019). Available at: https://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm#guides

[5] U.S. Food & Drug Administration (FDA). *FDA Data Standards Catalog* (current version at time of publishing – v5.2/December 2018). Available at: https://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm#catalog

[6] U.S. Food & Drug Administration (FDA). *Providing Regulatory Submissions In Electronic Format — Standardized Study Data* (December 2014). Available at: https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM292334.pdf

[7] Sviglin, Helena; Navarro, Eileen; Allard, Crystal; Rosario, Lilliam. *Implementing the SDRG: Reflections from the Reviewer Community* (PhUSE CSS 2015). Available at: https://www.phusewiki.org/docs/CSS2015Presentations/PP18FINAL.pdf

[8] Allard, Crystal et al. *JumpStarting Review: Highlights (PhUSE CSS 2015).* Available at: https://www.phusewiki.org/docs/CSS2015Presentations/PP21FINAL.pdf

[9] Doi, Mary. *How Good is Your SDTM Data? Perspectives from JumpStart* (PhUSE CSS 2016). Available at: http://www.phusewiki.org/docs/CSS%202016%20Presentations/SDTM%20Mary%20Doi.pptx

[10] Sviglin, Helena. *The State of Data Reviewer Guides* (PhUSE CSS 2016). Available at: https://www.phusewiki.org/docs/CSS%202016%20Presentations/The%20state%20of%20the%20Data%20Reviewers%20Guide%20Helena%20Svilgin.pptx

[11] Law, DeYett et al. *Data Quality Findings from JumpStart* (PhUSE CSS 2017). Available at: https://www.phusewiki.org/docs/2017_CSS_US/PP29_Draft.pdf

[12] Chen, Huanyu. *Common Data Related Review Issues and Prevention: A Statistical Reviewer's Thoughts* (PharmaSUG 2018). Available at: https://www.pharmasug.org/proceedings/2018/REG/PharmaSUG-2018-REG01.pdf

[13] Clinical Data Interchange Standards (CDISC). *Analysis Results Metadata (ARM) v1.0 for Define-XML v2.0* (current version at time of publishing – January 2015). Available at: https://www.cdisc.org/standards/foundational/adam

[14] VanPelt Nguyen, Sandra; Asam, Ellen; Dong, Wei; & Travalent, Annette. *Sorting Out the Paperwork – Define.xml versus Reviewer's Guide and other Submission Documents* (PhUSE CSS 2017). Available at: https://www.phusewiki.org/docs/2017_CSS_US/PP19_Final.pdf

[15] Clinical Data Interchange Standards (CDISC). *Study Data Tabulation Model Implementation Guide (SDTMIG)* (version referenced – 3.2/Nov. 2013). Available at: https://www.cdisc.org/standards/foundational/sdtmig

[16] Clinical Data Interchange Standards (CDISC). *Controlled Terminology* (latest version at time of publishing – 29Jun2018). Available at: https://www.cdisc.org/standards/terminology

[17] Kelly,Kristin. *Best Practice for Explaining Validation Results in the Study Data Reviewer's Guide* (PhUSE US Connect 2018). Available at: https://phusewiki.org/docs/2018_US%20Connect18/DS%20STREAM/ds13%20final%20.pdf

[18] U.S. Food & Drug Administration (FDA). *Study Data Standards Resources*. Available at: https://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm

[19] Japan Pharmaceuticals and Medical Devices Agency (PMDA). *Office of Advanced Evaluation with Electronic Data* (English). Available at: http://www.pmda.go.jp/english/review-services/reviews/advanced-efforts/0002.html

[20] Japan Pharmaceuticals and Medical Devices Agency (PMDA). *Notification on Practical Operations of Electronic Study Data Submissions* (English translation). April 2015. Available at: https://www.pmda.go.jp/files/000206451.pdf

[21] U.S. Food & Drug Administration (FDA). *Bioresearch Monitoring Technical Conformance Guide* (current version at time of publishing – Feb. 2018). Available at: https://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/FormsSubmissionRequirements/UCM332468.pdf

[22] U.S Food & Drug Administration (FDA). *Optimizing Your Study Data Submissions to FDA – Updates from CDER and CBER* (CDER SBIA Webinar July 13th 2017). Available at: https://www.fda.gov/Drugs/DevelopmentApprovalProcess/SmallBusinessAssistance/ucm565138.htm

[23] U.S. Food & Drug Administration (FDA). *Study Data for Submission to CDER and CBER.* Available at: https://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/ucm587508.htm

[24] Japan Pharmaceuticals and Medical Devices Agency (PMDA). *FAQs on Electronic Study Data Submission*. Available at: https://www.pmda.go.jp/english/review-services/reviews/advanced-efforts/0007.html

[25] Nakajima, Yuichi; Kitahara, Takashi; and Hara, Ryan. *Japanese Electronic Study Data Submission in CDISC Formats* (PhUSE Annual Conference 2016). Available at: https://www.phusewiki.org/docs/Conference%202016%20RG%20Papers/RG03.pdf

[26] Japan Pharmaceuticals and Medical Devices Agency (PMDA). *Technical Conformance Guide on Electronic Study Data Submissions* (English translation). April 2015. Available at: https://www.pmda.go.jp/files/000206449.pdf.

[27] Ando, Yuki. *Advanced Review with Electronic Data and CDISC Implementation in PMDA* (PhUSE Annual Conference 2015). Available at: https://www.pmda.go.jp/files/000208573.pdf

[28] U.S. Food & Drug Administration (FDA). *Technical Rejection Criteria for Study Data* (current version at time of publishing – May 2018). Available at: https://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/FormsSubmissionRequirements/ElectronicSubmissions/UCM523539.pdf

[29] Pinnacle21 Validator: https://www.pinnacle21.com/products/validation

Project Contact Information:

Sandra VanPelt Nguyen
sandra.vanpeltnguyen@pfizer.com

Ellen Asam
ellen_asam@merck.com

Acknowledgments:

We would like to acknowledge all of the people who have contributed toward the content in this paper over the years: Kiran Bonda, Wei Dong, David Fielding, Patty Hegarty, Mina Hohlen, Jacques Lanoue, Catherine Luckstone, Janet Low, Donna Sattler, Annette Travalent, Steve Wong, and Yan Ying.

We would also like to thank Lisa Brooks, Jane Lozano, and Sandra Minjoe for their valuable input and feedback.