PharmaSUG 2016 - Paper PO15

SDTM Metadata: The Output is only as Good as the Input

Sue Sullivan, d-Wise, Morrisville, North Carolina

ABSTRACT

Capturing all of the intricacies, exceptions to the rules, and finer points of the SDTM IG takes both attention to detail and manual effort. Many companies configure their own interpretation of SDTM metadata to meet their specific sponsor needs and ensure compliance with federal regulations. However, CDISC has created SHARE, "…an electronic repository for developing, integrating and accessing CDISC metadata standards in electronic format."[1] As part of SHARE, eSHARE files have been created that contain SDTM metadata for SDTM models v1.2-1.5, which are available to download only for CDISC Gold members.

The eSHARE files may be used in a number of ways. For one, the files may be used as a starting point to compile a sponsor's metadata by IG and model version. Additionally, the files may be used for compliance checking of sponsor standards, MDR content, or study level metadata.

However, the output is only as good as the input and the eSHARE files do not include all of the SDTM metadata contained as text in the IG which is fundamentally needed to completely describe your data. To create a comprehensive set of SDTM metadata based on these files, it is necessary to supplement the eSHARE files.

This paper defines a process to leverage the eSHARE files and extend those files to encapsulate all of the SDTM metadata needed to house or check a sponsor's SDTM metadata. Examples of what supplemental information is needed and how to develop comprehensive metadata and tools will also be provided.

INTRODUCTION

Complete and accurate metadata is the foundation of all processes that rely on that metadata. Metadata is 'data about data'. In the world of a clinical programmer, metadata tells you something about your datasets, your variables or the values of your variables. In the clinical space, complete and accurate metadata has become part of the critical path towards submission and approval of drugs and devices. With the publication of multiple Guidance for Industry documents and the Data Standards Catalogue by the FDA, sponsors will soon be required to submit clinical trial data and analysis using CDISC standards for New Drug Applications (NDAs)[2,3,4,5]. Included in this guidance is the requirement to provide the metadata associated with the clinical trial and analysis data as part of the submission package. Therefore, having accurate and complete study metadata is vital.

Many sponsors have chosen to create study level metadata from a set of sponsor defined metadata standards. This set of sponsor standards metadata is typically housed in some version of a Metadata Repository (MDR) (e.g., a commercial solution or Excel spreadsheets), and is used as the source for creating study level metadata. This allows for consistent use of the standards across studies as well as continued adherence to the standards.

CDISC SDTM and ADaM metadata is housed in various CDISC sources and file types: the Model documents, the Implementation Guides, the define.xml packages, and excel files (henceforth referred to as eSHARE). As part of the SHARE program, CDISC has made the above listed documents available; however, the eSHARE files are only available to its Gold Members and above. The eSHARE files contain SDTM and ADaM metadata held within unique excel spreadsheets by standard type (e.g., SDTM) and by version (e.g., SDTM IG v3.2). None of the sources listed above contain complete metadata on their own. The eSHARE files appear to be an attempt to house all of the metadata in a single file (one per version) in a format that could be used in a machine readable way; but unfortunately, the files are not complete. The metadata missing from these files is often found as text in the IG or model documents. Compiling all of the metadata from disparate sources into one file is critical for efficiency and accuracy.

This paper will focus on SDTM metadata, and will review the types of metadata that are held in the various SDTM SHARE files. An outline of how the metadata contained as text can be incorporated into a single file producing complete and accurate SDTM metadata for SDTM IG v3.2/SDTM model v1.4 will be presented.

UTILITY OF STANDARDS METADATA

Metadata provided by CDISC may be used in a variety of ways. It may be used as a starting point for sponsors to create their own standards by model and version for use in clinical studies. Alternatively, it may be used for compliance checking of a sponsor's standard metadata or study level metadata. Both of these types of checks may be built into a comprehensive Standards Compliance Check Tool (Figure 1). This tool allows the user to select

1. The standard to be checked: SDTM or ADaM
2. The version of the standard
3. NCI Controlled Terminology version
4. Input data: CDISC standard or sponsor standard
5. Comparator metadata to be used: sponsor standard or study level

These checks not only ensure that a sponsor's standards metadata are compliant with the published CDISC standards but also pave the way to metadata compliant study datasets for regulatory submission.

One can see that the accuracy and completeness of the CDISC standards metadata input directly correlates with the ability to create accurate and complete sponsor metadata and/or study level metadata used for submissions. Incomplete or inaccurate input standards metadata leads to incomplete or inaccurate output.

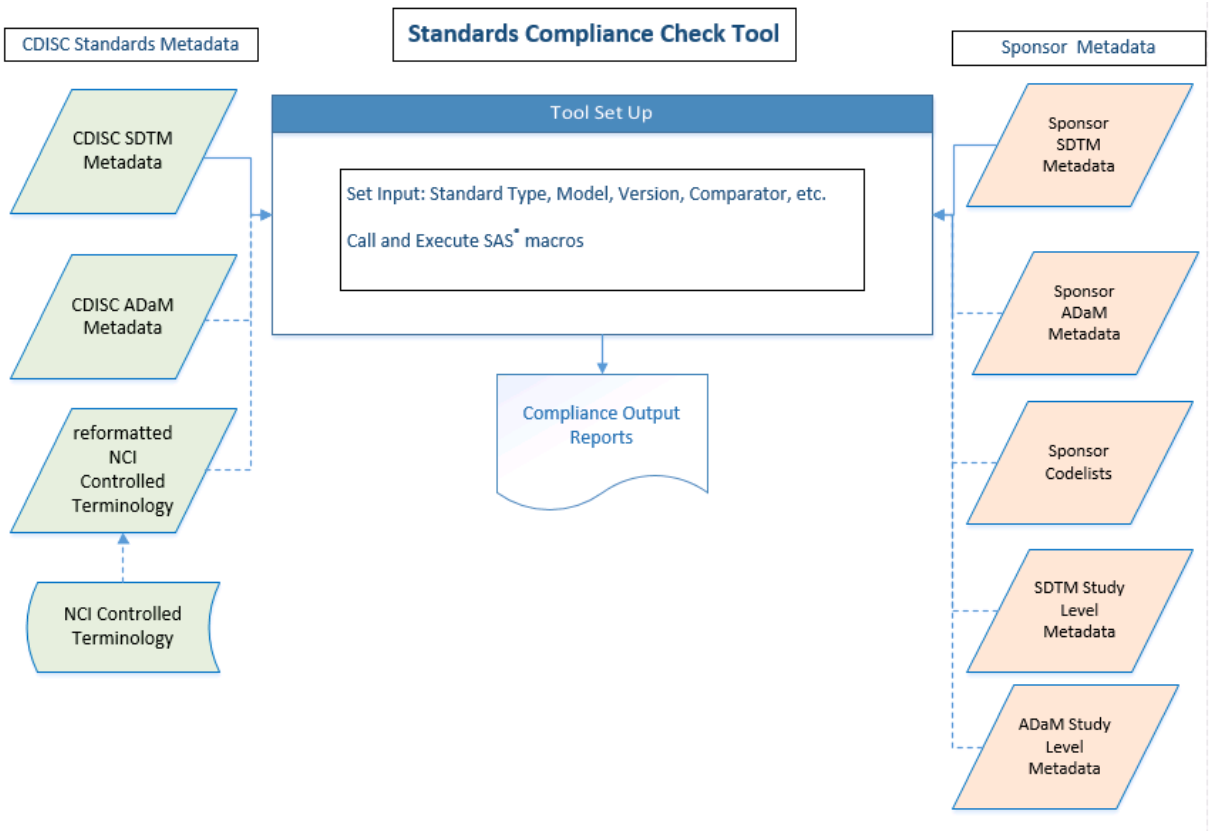

Figure 1. Standards Compliance Tool Design

A comprehensive Standards Compliance Check Tool allows for the checking of multiple standards as well as multiple types of input data. The standards include SDTM, ADaM and CDISC Controlled Terminology (CT). The input data includes sponsor standards metadata as well as study level metadata. The tool would allow for all versions of the different standards to be used.

CREATING COMPLETE SDTM METADATA

While the tool shown in Figure 1 uses SDTM, ADAM, and NCI Controlled Terminology CDISC standards as input metadata, Controlled Terminology (CT) and ADaM metadata are beyond the scope of this paper. This section will narrow the focus to the topic of creating complete CDISC compliant SDTM metadata.

CDISC SDTM metadata is housed in various CDISC sources; the SDTM IGs, the SDTM Model documents, the define.xml packages, and the eSHARE files (Figure 2). The SDTM model and IG documents house the largest amount of metadata, however this information is contained as text in pdf files or pdf portfolios, and therefore is not readily consumable as part of a compliance check tool. The approach to compile a complete set of CDISC SDTM metadata for SDTM IG v3.2 is to start with the eSHARE file for the same version, and add metadata from other sources as needed.

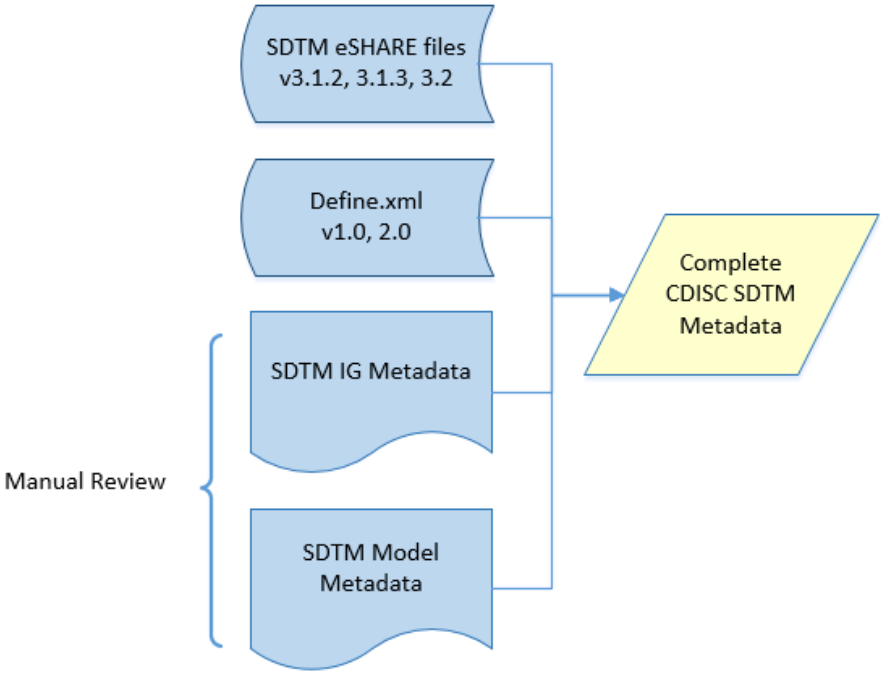

Figure 2. Sources of SDTM Metadata

CDISC SDTM standards metadata is housed in various CDISC eSHARE documents, datasets and files. No single source contains all of the metadata, and therefore construction of a complete set of SDTM metadata necessitates multiple input sources.

The eSHARE files are available for download from the CDISC website http://www.cdisc.org/cdisc-share. Each version of the SDTM model is contained in a unique file. All of the eSHARE files published to date (model versions 1.2 -1.5* IG v3.1.2 - 3.3) are formatted similarly and contain the same types of metadata. Table 1 contains an abbreviated example of the information held in a sample eSHARE file. Some of the metadata in this file is required for submission to regulatory agencies (e.g., Variable Name), while other metadata is not (e.g., Variable Name (minus domain prefix)).

Seq. For Order	Observation Class	Domain Prefix	Variable Name (minus domain prefix)	Variable Name	Variable Label	Type	….	Role	… ..	Core
8	Findings	LB	TESTCD	LBTESTCD	Lab Test or Examination Short Name	Char		Topic		Req
9	Findings	LB	TEST	LBTEST	Lab Test or Examination Name	Char		Synonym Qualifier		Req

Table 1. eSHARE for SDTM model v1.5, SDTM IG v3.2

Snapshot of the metadata housed in a SDTM eSHARE file. Column Headers reflect the metadata fields, while the rows contain the metadata values. The example shown is the metadata for variables LBTESTCD and LBTEST. Columns "Controlled Terms, Codelist or Format" and "CDISC Notes (for domains) Description (for General Classes)" were removed and replaced with "…" to enable the table to fit within the specified margins.

To begin assembling complete SDTM standards metadata, one may start with the eSHARE file for a particular model and IG version, add metadata that is contained in other sources, and possibly remove metadata that is not required or needed for a submission or for programming. The eSHARE file contains the following required define.xml

* The model v1.5 and IGv3.3 are draft documents. At the time this paper was authored, these documents have not been finalized and released.

metadata: Domain Class, Domain Prefix, Variable Name, Variable Label, Type, Role and Core. The main sources for the additional information are the SDTM model and the SDTM IG, and text held within the eSHARE file which corresponds to text in the IG (see Table 4, CDISC Notes). One must read through these documents carefully to find information related to metadata, as not all of the metadata is specifically called out in these documents.

Table 2 illustrates the metadata that is contained within the CDISC eSHARE file, as well as the metadata that must be found in other sources. Additionally, there is information contained in the eSHARE file that needs either manual or programmatic manipulation to make it consumable by a tool. The types of metadata requiring this type of manipulation are:

1. Additional allowed variables for each of the General Observations Class Domains
2. Custom Domain Metadata for each of the General Observations Classes. (Custom domains are domains created to house data that does not otherwise fit into one of the published General Observation Class of domains: Events, Interventions and Findings)

Metadata Contained in eSHARE	Metadata Contained in SDTM IG or SDTM Model Not in eSHARE
Domain Class	Domain Label
Domain Prefix	Variable Maximum Length (all variables)
Variable Name	Variable Order
Variable Label	Additional Allowed Variables for Special Purpose Class Domains
Type	Disallowed Variables by Domain
Role	
Core	
Variable Maximum Length (subset of variables)	

Table 2. Sources for SDTM Metadata.

SDTM metadata from the eSHARE files must be supplemented with metadata contained in other sources. Metadata contained in the SDTM model and the SDTM IG can be used to supplement the metadata contained in the eSHARE file.

METADATA CONTAINED IN THE SDTM MODEL AND THE SDTM IG DOCUMENTS

Capturing all of the intricacies, exceptions to the rules, and finer points of the SDTM IG and SDTM model takes both attention to detail and manual effort. While the SDTM model v1.4 is a fairly short document (40 pages), the SDTM IGv3.2 is broken out into a pdf portfolio hundreds of pages long. There is no one section in either document that contains all the metadata types and allowed values, and therefore the bulk of the documents must be read in their entirety. One must have an understanding of what constitutes standards metadata, and then proceed to read the documents and extract the necessary information.

Domain Level Metadata

The domain level metadata that is captured in the eSHARE file is:

1. The valid values for Domain Class and
2. Domain Prefixes for the domains published in the SDTM IG

However, there is no information regarding the valid values for Domain Labels. This information is fairly easy to find within the SDTM IG, and is also contained within the define.xml and the NCI CT. By navigating to either the Section Index for each of the Domain Classes, or by navigating to the domains themselves, one can locate the correct Domain Label for each domain within the IG. In the table below, the correct domain label for Domain Prefix = CO is Comments. Likewise, Domain Label = Demographics where Domain Prefix = DM. This information is then added to the eSHARE file as a new column with the associated metadata values.

Domain Code	Domain Description	Domain Document Name
CO	**Comments** The Comments dataset accommodates two sources of comments: 1) those collected alongside other data on topical case report form (CRF) pages such as Adverse Events and 2) those collected on a separate page specifically dedicated to comments.	Section 5 – CO Domain
DM	**Demographics** The Demographics domain includes a set of essential standard variables that describe each subject in a clinical study. It is the parent domain for all other observations for human clinical subjects.	Section 5 - DM Domain
SE	**Subject Elements** The subject element table describes the actual order of elements followed by the subject, together with the start date/time and end date/time for each element.	Section 5 – SE Domain
SV	**Subject Visits** The subject visits table describes the actual start and end data/time for each visit of each individual subject.	Section 5 - SV Domain

Table 3. Index from Section 5 of the SDTM IG v3.2

Domain Prefix (Domain Code) and Domain Label metadata (from Domain Description) contained in the SDTM IG v3.2, Section 5 index: Models for Special-Purpose Domains.

Variable Level Metadata

At the variable level the eSHARE file houses information regarding variable names, labels, their type (Num or Char), the role they have, and whether they are Required, Expected or Permissible. The eSHARE file also contains information in the CDISC Notes section that relates to additional metadata. The CDISC Notes in the eSHARE file is the same text that is held in the CDISC Notes sections of the IG. Unfortunately, this information is not held as a consumable piece of data, as it is part of a text string.

Variable level metadata housed within the SDTM model and IG but not in eSHARE are:

1. Maximum allowed character variable lengths for all variables
2. Additional allowed variables by domain
3. Variables disallowed by domain
4. Variable order within a domain

Once again, the process to find this information is to read the model and the IG paying attention to sections and text that relate to metadata.

Variable Maximum Length

The eSHARE file contains maximum length information for some variables as text. Capturing the variable maximum length metadata as a programmatically consumable values allows for metadata checks to be programmed. As the eSHARE file is not complete, it is wise to go to the SDTM IG to determine the value for maximum length for all variables. The information regarding variable maximum length is housed in different locations within the IG. Currently the FDA only accepts SAS® v5 transport files which have a maximum length of 200 for character variables[6].By default 200 char is the maximum length for all character variables until the time at which the FDA is able to accept different types of datasets. One must now read the CDISC Notes section (either from the IG or eSHARE) and determine if specific variables have maximum lengths other than 200 char.

This information can be defined in a more structured way, and therefore is added to the eSHARE file as a new column with the correct values entered for each variable. Each domain must be reviewed to capture all of the variable length maximum information contained in the CDISC Notes. Table 4 below shows a section of the updated eSHARE file containing maximum length metadata for a subset of the VS (Vital Signs) domain.

Domain Prefix	Variable Name	Variable Label	Type	Role	CDISC Notes (for domains) Description (for General Classes)	Core	Maximum Length
VS	STUDYID	Study Identifier	Char	Identifier	Unique identifier for a study.	Req	200
VS	DOMAIN	Domain Abbreviation	Char	Identifier	Two-character abbreviation for the domain.	Req	200
VS	USUBJID	Unique Subject Identifier	Char	Identifier	Identifier used to uniquely identify a subject across all studies for all applications or submissions involving the product.	Req	200
VS	VSSEQ	Sequence Number	**Num**	Identifier	Sequence Number given to ensure uniqueness of subject records within a domain. May be any valid number.	Req	**8**
VS	VSGRPID	Group ID	Char	Identifier	Used to tie together a block of related records in a single domain for a subject.	Perm	200
VS	VSSPID	Sponsor-Defined Identifier	Char	Identifier	Sponsor-defined reference number. Perhaps pre-printed on the CRF as an explicit line identifier or defined in the sponsor's operational database.	Perm	200
VS	VSTESTCD	Vital Signs Test Short Name	Char	Topic	Short name of the measurement, test, or examination described in VSTEST. It can be used as a column name when converting a dataset from a vertical to a horizontal format. **The value in VSTESTCD cannot be longer than 8 characters**. …….	Req	**8**
VS	VSTEST	Vital Signs Test Name	Char	Synonym Qualifier	Verbatim name of the test or examination used to obtain the measurement or finding. **The value in VSTEST cannot be longer than 40 characters.** ……	Req	**40**

Table 4. Maximum Length Metadata added to eSHARE in numeric format

Subset of column metadata shown for domain VS. Column "Maximum Length" has been added to the original eSHARE file. Numeric variables have a maximum length set to 8, character variables called out in the IG as having a maximum length have those values entered, and all other variables have the maximum length defaulted to 200 char per the SAS® transport file requirement.

TS Example and eSHARE Short Comings

Additional variable metadata that is housed in the IG but not as consumable metadata in eSHARE relates to variables that are allowed to be added to specific domains. In the case of the Trial Summary domain, Assumption 6 in the domain specification states "If TSVAL is > 200 characters, then it should be split into multiple variables, TSVAL-TSVALn"[7]. Variables TSVAL1, TSVAL2, etc. are not in the eSHARE file, and thus must be added with all of their associated domain and variable level metadata.

- By virtue of being in the TS domain, the Domain Name is TS.
- Based on Domain Controlled Terminology, the Domain Label = Trial Summary.
- As stated in the IG, the Variable Name = TSVAL1 which is incremented by n+1 for each additional variable that is needed, based on a 200 Character maximum.
- The user must determine the correct value for the Variable Label. The variable Label for TSVAL = Parameter Value. It would be logical to assume that the same Variable Label could be used for each TSVALn. However, the relationship between a Variable Name and its associated Variable Label is required to be one to one. Therefore, a compliant naming convention for TSVAL1 would be "Parameter Value 1".
- The Role would be the same as the Role for TSVAL, "Result Qualifier".
- Since TSVALn variables are only to be used if needed based on the length of the text for the value, the Core would equal Permissible or "Perm".
- Type is the same as TSVAL; 'Char'.
- As stated previously, the maximum length = 200.
- Additionally, a sponsor would need to determine the number of TSVALn variables they would need to create to adequately capture their Trial Summary data.

Table 5 shows an abbreviated eSHARE file for the updated TS domain, based on the metadata housed in the IG.

Observation Class	Order v3.2	Allowed v3.2	Allowed v3.1.2	Domain Prefix	Domain Label	Variable Name	Variable Label	Type	…	Role	Core	Max Length
Trial Design	1	Y	Y	TS	Trial Summary	STUDYID	Study Identifier	Char		Identifier	Req	200
Trial Design	2	Y	Y	TS	Trial Summary	DOMAIN	Domain Abbreviation	Char		Identifier	Req	200
Trial Design	3	Y	Y	TS	Trial Summary	TSSEQ	Sequence Number	Num		Identifier	Req	8
Trial Design	4	Y	Y	TS	Trial Summary	TSGRPID	Group ID	Char		Identifier	Perm	200
Trial Design	5	Y	Y	TS	Trial Summary	TSPARMCD	Trial Summary Parameter Short Name	Char		Topic	Req	8
Trial Design	6	Y	Y	TS	Trial Summary	TSPARM	Trial Summary Parameter	Char		Synonym Qualifier	Req	40
Trial Design	7	Y	Y	TS	Trial Summary	TSVAL	Parameter Value	Char		Result Qualifier	Exp	200
Trial Design	8	Y	Y	TS	Trial Summary	TSVAL1	Parameter Value 1	Char		Result Qualifier	Perm	200
Trial Design	9	Y	Y	TS	Trial Summary	TSVAL2	Parameter Value 2	Char		Result Qualifier	Perm	200
Trial Design	10	Y	Y	TS	Trial Summary	TSVAL3	Parameter Value 3	Char		Result Qualifier	Perm	200
Trial Design	11	Y	Y	TS	Trial Summary	TSVAL4	Parameter Value 4	Char		Result Qualifier	Perm	200
Trial Design	12	Y	Y	TS	Trial Summary	TSVAL5	Parameter Value 5	Char		Result Qualifier	Perm	200
Trial Design	13	Y	Y	TS	Trial Summary	TSVAL6	Parameter Value 6	Char		Result Qualifier	Perm	200
Trial Design	14	Y	N	TS	Trial Summary	TSVALNF	Parameter Null Flavor	Char		Result Qualifier	Exp	200
Trial Design	15	Y	N	TS	Trial Summary	TSVALCD	Parameter Value Code	Char		Result Qualifier	Exp	8
Trial Design	16	Y	N	TS	Trial Summary	TSVCDREF	Name of the Reference Terminology	Char		Result Qualifier	Exp	200
Trial Design	17	Y	N	TS	Trial Summary	TSVCDVER	Version of the Reference Terminology	Char		Result Qualifier	Exp	200

Table 5. TS metadata

Excerpt of Trial Summary metadata housed in an updated eSHARE file, based on TS metadata information from the SDTM IGs. Items in green denote additions to the original file. In this example, the sponsor chose to include six additional variables for TSVAL.

eSHARE Issues with Disallowed Variables in AE

The SDTM IG also spells out specific variables that are disallowed for some domains. Adverse Events (AE) is one example of this. The AE domain is an Events domain, which is one of the three General Observation Classes. The SDTM model contains a list of allowed variables for each of the three General Observations Classes: Findings, Events, and Interventions. However, in the SDTM IG Section 6.3, AE Domain specification, Assumption 8 states "...the following Qualifiers would not be used in AE: --OCCUR, --STAT, and--REASND. They are the only Qualifiers from the SDTM Events Class not in the AE domain. They are not permitted because the AE domain contains only records for adverse events that actually occurred."[8] Therefore, the variables AEOCCUR, AESTAT, and AEREASND are not allowed for use in an AE domain.

The variables captured in the eSHARE file for the AE domain, are only those listed in the IG, they do not include all the variables allowed for an Events domain per the model. In order to capture the disallowed variables, the eSHARE AE domain would need to be updated to include all of the allowed variables in an Events domain, minus AEOCCUR, AESTAT, and AEREASND. This process would be done for each applicable domain, in order to capture the complete list of allowed variables by domain.

Variable Order

Once the final list of variables is assembled for all domains, the metadata for variable order would be added. This ensures that not only are all of the allowed variables added with their associated metadata, but the order that the variables are expected to be in the datasets is also captured. This is the last piece of metadata that would need to be added at the variable level for the specific SDTM version.

Capturing SDTM version metadata

As of yet, the variable and domain level metadata in the SDTM versions has not changed, other than the addition and removal of some variables. Therefore, an additional set of metadata attribute could be added to the file for each SDTM version to flag each variable as allowed per a specified version (Table 5, columns "Allowed v3.2", "Allowed v3.1.2"). This allows for the metadata from multiple SDTM versions to be housed in a single file.

CONCLUSION

As part of a submission to the FDA, sponsors of clinical trials are expected to, and will soon be required to, submit complete study metadata using CDISC standards for NDAs. This necessitates the need for sponsors to have complete and accurate metadata. In addition, sponsor held standards serve as the foundation for study level datasets. Assuring that the source metadata is complete, accurate, and compliant leads to more compliant study data and metadata, and may increase the quality of a submission packages to regulatory agencies.

Metadata for the SDTM model is found in multiple sources in multiple formats, and may be accessed via CDISC SHARE. The ideal behind SHARE is a good one; however as shown, there are still a number of gaps in the eSHARE files that require manual updates to obtain a complete set of metadata. Before using the eSHARE files out of the box, be sure to supplement these documents with information found in the various documents identified throughout this paper.

REFERENCES

---

[1] CDISC SHARE. January 13, 2016. Available at http://www.cdisc.org/cdisc-share

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

[3] FDA. "Providing Regulatory Submissions In Electronic Format — Standardized Study Data Guidance for Industry". December 2014. Available at http://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm

[4] FDA. "Study Data Technical Conformance Guide v2.3". October 26, 2015. Available at http://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm

[5] FDA. "Data Standards Catalog v4.4". September 3, 2015. Available at http://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm

[6] CDISC Submission Data Standards Team. “Study Data Tabulation Model Implementation Guide: Human Clinical Trials version 3.2”, Section 4.1.5.3.2.

[7] CDISC Submission Data Standards Team. “Study Data Tabulation Model Implementation Guide: Human Clinical Trials version 3.2”, Section 7.4, Trial Summary Domain Specification, Assumption 6

[8] CDISC Submission Data Standards Team. “Study Data Tabulation Model Implementation Guide: Human Clinical Trials version 3.2”, Section 6.3, Adverse Events Domain Specification, Assumption 8

ACKNOWLEDGMENTS

The author would like to acknowledge the following individuals for their contributions to this paper: Mike Molter, Senior Life Sciences Consultant, d-Wise, Morrisville North Carolina and Chris Decker, Vice President Life Sciences, Morrisville North Carolina

RECOMMENDED READING

- SDTM Model v1.5, available from the CDISC web site at: http://www.cdisc.org/sdtm.
- SDTM IG v3.2 available, from the CDISC web site at: http://www.cdisc.org/sdtm.
- Define.xml v2.0 package, available from the CDISC web site at: http://www.cdisc.org/sdtm.
- Providing Regulatory Submissions In Electronic Format — Standardized Study Data Guidance for Industry, available from the FDA website at: http://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.html
- Study Data Technical Conformance Guide v2.3, available from the FDA website at: http://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.html

CONTACT INFORMATION

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

Name: Sue Sullivan
Enterprise: d-Wise
Address: 1500 Perimeter Park Drive, Suite 150
City, State ZIP: Morrisville, NC 27560
Work Phone: 919-334-6090
Fax: 888-563-0931
E-mail: sue.sullivan@d-wise.com
Web: www.d-wise.com