CDISC Submission Metadata Model

by
Dave Christiansen and Wayne Kubick

Version 2.0 - 26 November 2001

The CDISC Submission Data Standards Metadata Approach

The CDISC Submission Data Model has focused on the use of effective metadata as the most practical way of establishing meaningful standards applicable to electronic data submitted for FDA review. Metadata is defined as "data about the data"; in other words, metadata includes description of the content, context, structure, and/or purpose of a database. It is important to recognize that the metadata provided by the model is intended to be the minimum required to meet the need of FDA users, and is not intended to fully meet all of the needs of the sponsor's data management, statistics, or other internal groups. Additional internal metadata standards will be desirable within most organizations to govern the ways that data is captured, cleaned, and analyzed statistically.

The CDISC Submission Metadata Model was created to help ensure that the supporting metadata for these submission datasets should meet the following objectives:

- Provide FDA reviewers with clear descriptions of the usage, structure, contents, and attributes of all datasets and variables
- Allow reviewers to replicate most analyses, tables, graphs, and listings with minimal or no transformations
- Enable reviewers to easily view and subset the data used to generate any analysis, table, graph, or listing without complex programming.

The above objectives of minimizing transformations and complex programming are enhanced by including certain common variables such as Sex, Treatment, etc. in multiple datasets (i.e., variable redundancy). This model does not address specific content issues such as how to populate individual datasets for a particular study. The data collected and reported for a study should be based on scientific and medical considerations. However, the model does guide sponsors toward certain common conventions by identifying certain minimum core data elements that should normally be included with basic safety datasets. As a result, the model should 1) provide greater consistency and uniformity among all future submissions and, 2) begin to reduce the range of possibility for diversity in data that is provided for regulatory submissions. The model helps ensure that those data domains, elements, and attributes that are common to all submissions will be represented in the same manner in every case. By concentrating on metadata, the Submissions Group hopes to achieve many of the benefits of dictionary standardization while ensuring that scientific objectives are not compromised.

The CDISC Submission Data Standards are based on the three FDA guidance documents: "Guidance for Industry, **Providing Regulatory Submissions in Electronic Format — General Considerations (IT-2, January, 1999), Providing Regulatory Submissions in Electronic Format – NDAs (IT-3, January, 1999)**, and **Providing Regulatory Submissions in Electronic Format – Biologics Marketing Applications (November, 1999)** and on the Example of an Electronic New Drug Application Submission dated 2/17/99.

Metadata Definitions for Domain Datasets

In the FDA guidance documents, the core safety data are divided into 12 domain datasets and the data elements are described in the data definition file (define.pdf). Using the FDA Sample NDA as a starting point, the CDISC Submission Data Standards Group has made a number of improvements and proposes the following dataset definitions.

The first table in the data definition file provides basic information about each of the datasets in four columns (items that are already requested in the FDA guidance are marked with a *):

- ***DATASET NAME** - The 8-character name of the dataset is provided by the sponsor in this column. Note: the FDA may be specifying standard names to be used for some of the more common datasets in a future guidance.. For this example, the names used in the FDA Sample NDA will be used whenever possible.
- ***DESCRIPTION** - A more detailed description of the information contained in the dataset is included in this column. This column should also be linked to the domain variable descriptions table.
- ***LOCATION** - The file location including the folder and file name is included in this column. This column may be linked to the dataset.
- **STRUCTURE** - The level of detail represented by each record in a dataset. This describes the "shape" of the dataset. Examples of dataset structures include one record per patient, one record per patient per visit, one record per patient per event, one record per patient per visit per event, etc. A single data domain may need more than one structure to facilitate understanding of the data (for example, labs may be reported as one record per lab measurement in a CRT dataset, or as one record per patient visit in an analysis dataset). The structure(s) for a dataset may depend on the type of data, the indication, and/or reviewer preferences. This information was not described in the guidance.
- **PURPOSE** – The purpose of the dataset: either CRT or Analysis. Both NDA and BLA guidance documents require "Case Report Tabulations" (CRT) datasets containing both raw and derived variables pertaining to the domain. Although the main purpose of these CRT datasets is to list the data from a particular domain, these datasets may also be used for some analyses. Additional analysis datasets may also be required, and should be discussed with the reviewing division prior to the submission. This information was not described in the guidance. Guidelines on the use of the metadata model for analysis datasets are being formed separately by the CDISC Analysis Dataset Modeling group (ADaM).
- **KEY FIELDS** – Used to uniquely identify and index each record in a dataset. Most datasets will have between 3 and 5 key variables. Key variables are often used when combining

datasets of a compatible structure. For example, when merging two datasets in SAS, the SAS variables would be selected from the common key variables of the datasets that are being merged. The key variables for each dataset will depend on the dataset structure and the sponsor's conventions. For this example, each subject will be uniquely identified by a single variable, USUBJID and each visit by VISIT. See the Variable Definition section below for more details. This information was not described in the guidance.

Dataset Definition

Dataset	Description	Location	Structure [optional]	Purpose	Keys
DEMO	Demographics and Subject Characteristics DM	Demo.xpt	1 Rec per subject	CRT	USUBJID
AE	Adverse Events AE	Adv.xpt	1 Rec per subject per adverse event occurrence	CRT	USUBJID, AESEQ
CONMEDS	Concomitant Medication CM	Conmed.xpt	1 Rec per subject per medication per event	CRT	USUBJID, CMSEQ
DISPOSIT	Disposition DS	Disposit.xpt	1 Rec/subject	CRT	USUBJID
ECG	ECG EG	Ecg.xpt	1 Rec per subject per visit per finding (per timepoint/position/qualifier) (horizontal model); 1 Rec per subject per visit per test, measurement, or finding (vertical model)	CRT	USUBJID, VISIT
EXPOSE	Drug Exposure EX	Expose.xpt	1 Rec per subject per constant dosing interval	CRT	USUBJID, EXSEQ
CHEM	Labs – Chemistry Detail LB	Chem.xpt	1 Rec per subject per visit per measurement		USUBJID, VISIT, Measure
CHEMSUM	Labs – Chemistry Summary LB	Chemsum.xpt	1 Rec per subject [per visit]		USUBJID, [VISIT]
HEMAT	Labs – Hematology LB	Hemat.xpt	1 Rec per subject per visit per measurement	CRT	See CHEM
URINE	Labs – Urinalysis LB	Urine.xpt	1 Rec per subject per visit per measurement	CRT	See CHEM

MEDHIST	Medical History MH	Medhist.xpt	1 Rec per subject per condition/procedure	CRT	USUBJID, HXSEQ
PE	Physical Examination PE	Pe.xpt	1 Rec per subject per exam (or per body system, or per finding)	CRT	USUBJID, PESEQ
VITAL	Vital Signs VS	Vital.xpt	1 Rec per subject per visit(per position/qualifier)] (horizontal model); 1 Rec per subject per visit per measurement (per position/qualifier) (vertical model)	CRT	USUBJID, VISIT, [Position]

Metadata Definitions for Domain Variables

These tables describe the specific variables included in each dataset. The CDISC metadata model intends to create a superset of possible variables that might be included in a submission over time – not a standard list of required elements. While there are some data elements that would normally be expected to exist in the datasets for most submissions, there are many others that are indication-specific or optional, and the latter should only be included when pertinent to the submission at hand.

The metadata definitions for each domain include the following metadata columns (items that are already requested in the FDA guidance are marked with a *):

***VARIABLE NAME** - This column should include the 8-character field name the sponsor used for its analyses. The 8-character limitation is currently required due to a limitation of the SAS V5 transport format currently required by the FDA; a more flexible format is expected to replace this in the near future. CDISC has proposed variable names that should be considered for use as standards by sponsors; however, CDISC recognizes that the sponsor may choose to use their own internal variable names until such time that the FDA requests standard names. If a common standard for naming variables included in a submission is defined by the FDA in a future guidance document, the sponsor's internal names can still be included as aliases.

***VARIABLE LABEL** - This is a 40-character description of the variable (the maximum length allowed by SAS V5 transport datasets). The label should adequately explain the content of the variable. The label should always be defined by the sponsor to be consistent with the submission document. This is a correction to the 32-character limit noted in the Guidance documents.

***TYPE** - This describes whether the variable is a character string or numeric value to conform with existing SAS conventions. Values that can be either character or numeric are listed as "Char or Num", but should be consistent throughout a submission. Since date, time, and date-

time values are stored by SAS as numeric values, sponsors should specify SAS formats in the DECODE/FORMAT column where appropriate.

***DECODES/FORMATS** - This column describes the character or number codes used in the dataset for each variable. Wherever possible, self-explanatory text should be used instead of codes (e.g., use "Y" instead of "1"). In addition, this column can list all allowable values for the field. (This column could contain standard SAS formats). CDISC has recommended certain standard formats for specific variables, including: Sex, Yes/No Variables and Date Variables.

ORIGIN - This column shows the source or point of origin for each variable included within the current domain.

- **Current Domain -** For variables that originate within the current domain, the column should indicate whether it was collected intact from a CRF or electronic device (a source variable), or whether it was computed or derived.
  - **Source Variables** - For variables captured from CRFs, the location of the information on the case report form is provided as a link to the annotated CRF pdf file.
  - **Derived Variables** - For derived variables, a description of how the variable was derived, including any algorithms used, should be provided (this may be done by providing a hyperlink to another document describing the derivation algorithm or process especially when the algorithm is complex).
- **External Domain -** For variables that originate in other, external domains but are merged into the current domain (e.g., Age, Sex), this column should list the dataset where it originated and provide a link to the appropriate domain variable definition. For example, the variable "AGE" would be described in the AE variable definition table as DEMO.AGE; and a hyperlink to the DEMO definition table would then show how age was computed in the Demographics Domain.

ROLE - This attribute provides information on how a variable is used in a particular dataset. In the current domain descriptions provided by CDISC, the Key and Selection variables are normally presented first in a dataset. Other roles may be included by the sponsor at their discretion. Each variable may have multiple roles. For example, a Key variable may also be a Selection variable. Other roles are being considered by the CDISC ADaM group, and additional custom roles may be required in the future to support standard FDA review tools.

- **Key Variables** - used to uniquely identify and index each record in a dataset: Study, Center, Subject ID, Visit. Most datasets will have between 3 and 5 key variables. Key variables are often used when combining datasets of a compatible structure. For example, when merging two datasets in SAS, the SAS variables would be selected from the common key variables of the datasets that are being merged. Key variables should **always** be clearly identified in the metadata for each dataset and are generally presented first.

- **Selection Variables** – variables that are frequently used to subset, sort or, group data for reporting purposes. These fields are frequently included in other datasets to simplify queries and facilitate simple analyses. For example: Sex, Age, and Race are often merged into other datasets from Demographics, and the FDA requests that Treatment Group be merged from the Drug Exposure table into every submitted domain. CDISC

has assigned a group of core selection variables that should normally be included in all datasets to simplify review tasks. These core selection variables are shown in Table 2 below.

- **Review Variables** –source or derived variables directly related to the primary objectives of the study. These variables are usually tabulated and or summarized for analysis purposes. They may be continuous, categorical or both (e.g., mean age and frequency by age group).

- **Support Variables**– these variables are usually not Key, Selection, or Review variables, but provide other useful background or reference information, or provide input for deriving Review variables (e.g., Date of Birth).

- ***COMMENTS** – This field should be used by the sponsor to present other useful information that may assist the reviewer in understanding the data. For example, it might list which coding dictionary (COSTART, MedDRA, etc.) was used to populate a coded value.

Two additional columns (which are not intended to be provided in the metadata submitted to the FDA) are included in the sample domain datasets defined by CDISC in order to help the sponsor prepare their metadata definitions:

- **CDISC Notes** – additional clarifications or notations about how to interpret or use a variable in the dataset.

- **CDISC Core Variable** – variables marked with a "Y" should normally be present in all submitted datasets. In some cases, this column will indicate when several alternatives may be considered for filling the need for a core variable in this area. Variables marked with an "N" should be supplied only when appropriate. CDISC generally includes examples of the most commonly used non-core variables in their datasets, but sponsors should include any other relevant data.

The following variables are considered core selection variables for use in all CDISC domain models. These variable roles may also be defined with other roles (such as Key), and roles may differ from dataset to dataset.

Variable Name	Variable Label	Comments	Included in:
STUDYID	Study ID	Uniquely identifies a study within a particular submission.	All files
SITEID	Site ID	Some sponsors may use INVID instead of or in addition to a SITEID.	At least one of these variables must be included in all files
INVID	Investigator ID		
USUBJID	Unique Subject ID	Must be unique subject identifier within a submission (previously defined as PID; should be consistent with PID references used elsewhere in the submission)	All files

SUBJID	Subject ID	Number ID captured on CRF to identify a subject	All files
AGE	Age in Years at Baseline	An additional variable for age units may be necessary for age collected in other units.	All files
SEX	Sex	Standard codes are assigned as F for Female, M for Male	All files
RACE	Race	Sponsor defined; should be consistent across submission	All files
TRTCD	Treatment Code	Numeric version of assigned treatment/sequence group to aid in sorting. Originates in Exposure domain.	One or more in all files
TRTGRP	Treatment Group	Sponsor-defined assigned treatment group. Should include description of drug and dose, and be consistent across all domains. Usually originates in Exposure domain.	
COUNTRY	Country	Should use standard codes from ISO 3166	Include in all files for submissions comprising multinational studies.
VISITNUM	Visit Number	1. Clinical encounter number. 2. Numeric version of Visit	One of these visit/study identifiers must be core to reference a clinical encounter or event.
VISITDY	Visit Day	Planned study day of visit.	
VISIT	Visit Name	1. Clinical encounter description. 2. May used in addition to VISITNUM and VISITDY as a text description of a clinical encounter, or to indicate whether a visit was unplanned, unscheduled, or a repeat.	

Preparing Dataset Metadata

The CDISC Submissions Data Standards (SDS) group has defined metadata models for the twelve safety domains specified in the FDA Guidances. These models represent a starting point for creating metadata for any safety dataset. Metadata prepared for a domain (such as an efficacy domain) which has not been described in a CDISC model should follow the general format of the safety domains, including the same set of core selection variables and all of the metadata attributes specified for the safety domains. Additional examples and usage guidelines are available on the CDISC web site at www.cdisc.org.

Data Conventions

The CDISC Metadata Model describes the structure and form of data, not the content. However, the varying nature of clinical data in general will require the sponsor to make some decisions about how to represent certain real-world conditions in the dataset. Therefore, it is useful for a metadata document to give the reviewer an indication of how the datasets handle certain special cases. Some of these special cases to be considered and suggestions for handling them are described below:

- In general, duration values (regardless of units) should be calculated using the convention stop – start + 1.
- Case sensitivity of text values -- The metadata document should indicate if uppercase is used consistently for text data.
- Missing/uncollected values – Since data will be submitted as SAS transport datasets, the convention used for missing values (or values that were not collected for certain records in a CRT) should be described. The conventions should be used consistently in all datasets and explained in the metadata.
- Non-numeric data in numeric fields – In general, a character field type must be used for a field that contains numeric as well as non-numeric data, but it is useful if the metadata describes when this occurs.
- Partial dates - the metadata should indicate whether dummy dates or incomplete values are used for specified dates, and the same convention should be used consistently.
- Partial date/times - these should be handled in the same manner as partial dates.

Conclusions

The CDISC Submission Metadata Model has been defined to guide sponsors in the preparation of data that is to be submitted to the FDA. By following the principles of this model, sponsors will help reviewers to accurately interpret the contents of submitted data and work with it more effectively, without sacrificing the scientific objectives of clinical development.