CDER Common Data Standards Issues Document

(Version 1.0/May 2011)

The Center for Drug Evaluation and Research (CDER) is strongly encouraging sponsors to submit data in standard form as a key part of its efforts to continue with advancement of review efficiency and quality. CDER has been collaborating with CDISC, a standards development organization (SDO), in the development of standards to represent study data submitted in support of regulatory application. Study data standards are vendor-neutral, platform-independent, and freely available via the CDISC website (**www.CDISC.org**). CDISC study data standards include SDTM (Study Data Tabulation Model) for representation of clinical trial tabulations, ADaM (Analysis Data Model) for clinical trial analysis files, and SEND (Standard for Exchange of Non-clinical Data) for representation of nonclinical animal toxicology studies tabulations.

CDER has accepted SDTM datasets since 2004; however, due to differences in sponsor implementation of the standard, CDER has observed significant variability in submissions containing "standardized" electronic clinical trial data. CDER has received numerous "SDTM-like" applications over the past several years in which sponsors have not followed the SDTM Implementation Guide. Furthermore, aspects of particular sponsor implementations have actually resulted in increased review difficulty for CDER reviewers. The goal of this document is to communicate general CDER preferences and experiences regarding the submission of standardized data in order to aid sponsors in the creation of standardized datasets.

This document is not intended to replace the need for sponsors to communicate with review divisions regarding data standards implementation approaches or issues, but instead, it is designed to complement and facilitate the interaction between sponsors and divisions. Because of specialized needs in different divisions, it is likely that divisions may have additional requests or preferences. When uncertainty exists regarding a particular data standards implementation or submission issue, the sponsor should contact the review division to discuss further.

This document will be **updated periodically** based on division/reviewer feedback and experience. Therefore, it is important that sponsors refer to the CDER data standards website (**http://www.fda.gov/Drugs/DevelopmentApprovalProcess/FormsSubmissionRequirements/ElectronicSubmissions/ucm248635.htm**) to ensure that they are using the most up-to-date version. Questions regarding the content of this document should be submitted to cder-edata@fda.hhs.gov.

GENERAL CONSIDERATIONS

Introduction:

Sponsors should refer to **http://www.cdisc.org** for the latest version of the SDTMIG, SENDIG and ADaMIG, in addition to other documentation related to the study data standards.

It is understood that CDISC data standards are evolving and that there may be instances in which the current implementation guides do not provide specific instruction as to how certain study data should be represented. In this instance, sponsors should discuss their proposed solution with the review division and submit supporting documentation as part of the reviewer's guide at the time of submission that describes these decisions/solutions.

Additionally, sponsors should refer to the Amendment 1 to SDTM V1.2 located at the following CDISC.org site (**http://www.cdisc.org/sdtm**). This document has been posted for trial use and public comment. Unless otherwise instructed by a division, these represent CDER-desired changes that sponsors should account for. It is expected that these changes will proceed through the usual CDISC data standards development process for evaluation and adoption as part of the SDTM standard.

A properly functioning define.xml file is an important part of the submission of electronic datasets and should not be considered optional. As a transition step, CDER prefers that sponsors submit both the define.pdf and define.xml formats. CDER will advise when it is ready to only receive define.xml.

Additionally, sponsors should make certain that every data variable's codelist, origin, and derivation is clearly and easily accessible from the define file. An insufficiently documented define file is a common deficiency that reviewers have noted.

Please include the variables EPOCH, ELEMENT, and ETCD (element code) for every subject-level observation (e.g., adverse events, laboratory, concomitant medications, exposure, vital signs). This will allow the reviewer to easily determine during which phase of the trial the observation occurred (e.g., screening, on-therapy, follow-up), as well as actual intervention the subject experienced during that phase. In the case where an event duration spans multiple Epochs and Elements, the entry should contain the Epoch and Element at the onset of the event.

SDTM:

The ideal time to implement SDTM standards for representation of clinical trial tabulation data is prior to the conduct of the study. The use of case report forms that incorporate SDTM-standard data elements (such as with CDASH-style case report forms) allows for a simplified process for creation of SDTM domains. This approach is preferred to the alternative of collecting data in a non-standard format and then converting to SDTM format after the trial (legacy data conversion). Legacy data conversion is often complex and difficult, and CDER has received submissions in which reviewers have occasionally encountered difficulties while reviewing converted data.

It is very important that the results presented in the Clinical Study Report be traceable back to the original data elements as they were collected in the case report form and represented in the SDTM datasets. The SDTM datasets must be able to support the results in the Clinical Study Report, either directly for some results, or, for other results, indirectly through analysis datasets that are derivable from the SDTM datasets.

If a sponsor decides to convert trial data to SDTM that was originally collected in non-SDTM format, it is important to note that the resulting SDTM data should support the accompanying analysis data sets and sponsor's reports (study reports, etc.). CDER has received applications in which the converted SDTM data sets were not consistent with the submitted analysis datasets and study report analyses, thus causing confusion during application review.

The SDTM Implementation Guide (SDTMIG) (**http://www.cdisc.org**) should be followed carefully. Section 3.2.2 of the SDTMIG provides general criteria conformance with the SDTM data model. These criteria should not be interpreted as the sole indication of the adequacy of submitted data. However, they should be followed unless otherwise indicated. If there is uncertainty with regards to implementation, the sponsor should discuss with the review division.

ADaM:

For **analysis datasets**, sponsors should refer to the published ADaM Implementation Guide (**www.CDISC.org**) as well as the CDER Study Data Specifications Document. It is expected that significant discussion between the sponsor and CDER reviewers will be necessary to appropriately determine which analysis datasets and associated content are needed to support application review.

The label names of analysis data sets should not be identical to those of the SDTM datasets. For example: the SDTM adverse event dataset (AE) and the analysis adverse event dataset (e.g., ADAE) should not share the exact same label, such as "Adverse Events".

Analysis datasets should be derivable from the SDTM datasets, in order to enable traceability from analysis results presented in the study reports back to the original data elements collected in the case report form and represented in the SDTM datasets. There are features built into the ADaM standard that promote traceability from analysis results to ADaM datasets, and from ADaM datasets to SDTM.

SEND:

CDER currently accepts regulatory submissions of tabulated nonclinical toxicology data in electronic format provided that a complete study report is also submitted. If a sponsor elects to submit datasets in the SEND format, they should clearly note in the define file what version of the data standard was used in addition to definitions of variables and data submitted. Sponsors should continue to consult the Study Data Specifications for information regarding submission of tumor datasets from rodent carcinogenicity studies (**http://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM199599.pdf)**. Sponsors should contact the appropriate division with any additional questions. CDER is

currently piloting SEND as a standard for the submission of nonclinical data for general toxicology and carcinogenicity studies.

TERMINOLOGY

CDISC Controlled Terminology

Data values for CDISC standards-specified variables should use the CDISC Controlled Terminology, which can be found at the NCI Enterprise Vocabulary Services (**http://www.cancer.gov/cancertopics/terminologyresources/page6**).

No Standard Terminology Exists

For variables for which no standard terminology exists, or if the available terminology is insufficient and needs to be extended, the sponsor may propose their own terminology. Please provide supporting documentation that describes the non-standard terminology that is used. The define.xml is the preferred mechanism for documenting such terminology issues.

Consistent Drug Dictionary

It is strongly preferred that a consistent drug dictionary (for example, the WHO Drug Dictionary) terminology be used for coding of the concomitant medications. The generic preferred term for a drug should be used for the SDTM standardized medication name variable, CMDECOD.

MedDRA

When using MedDRA for adverse events and medical history terms, sponsors should exactly follow the spelling and case of the MedDRA terms. A common error that has been seen is a misspelling of a System Organ Class term or other MedDRA term. Sometimes trials are conducted at different times during the development cycle which results in the use of different versions of MedDRA from one study to the next. It is expected that the Adverse Event dataset for the Integrated Summary of Safety include MedDRA Preferred Terms from a single version of MedDRA. The reason for this request is that reviewers often want to analyze adverse events across trials, including the use of Standardised MedDRA Queries. If different dictionary versions are used for data included in the same analysis, there is the potential for confusion or incorrect results.

Generally, no numerically ***coded variables*** should be submitted as part of the SDTM datasets. Numeric values generated from validated scoring instruments or questionnaires do not represent codes, and so have no relevance for this issue. There may be special instances were codes are preferred, so sponsors should refer to the review division for direction, if there is any question. CDER is currently exploring the potential use of codes for particular variables (example: UNI codes for concomitant medications or MedDRA codes for adverse events), possibly as a pilot.

Common Dictionaries

It is expected that common dictionaries are used across trials and throughout the submission for each the following: adverse events, concomitant medications, procedures, indications, study drug names, and medical history. Implementation of such dictionaries should be careful to exactly follow the terminology conventions (e.g., spelling and case) specified by the dictionary or according to a single

consistent sponsor specification if no pre-existing terminology exists. CDER has frequently received data in which terminology conventions were not followed, for example, misspelling of MedDRA or WHO Drug terms or lack of conformance to upper/lower case or the use of hyphens. This makes it difficult to use or develop tools to analyze this data in an automated fashion. Again, the use of a standard dictionary that is sponsor defined or extended should be documented using the define.xml.

SDTM DOMAINS

SUPPQUAL Datasets

SUPPQUAL represents a series of datasets in both SDTM and SEND submissions. It is intended to include data variables that are not specified in the SDTM. SUPPQUAL datasets are often used as a "waste basket" for data elements that the sponsor is not sure how to allocate. Discussion needs to occur if the sponsor intends to include important variables (that support key analyses) in the SUPPQUAL datasets. For clinical trial data in SDTM, one way to deal with this issue for important data elements that are likely to be needed to support review work, is to ensure that analysis datasets include these and other relevant data elements. However, these data elements should not then be excluded from SUPPQUAL, since there is a need to maintain traceability from ADaM back to SDTM.

DM Domain

In the DM domain, each subject should have only one single record per study. Integrated summaries may contain more than one record per unique subject in the case that an individual subject was enrolled in more than one study.

Amendment 1 to SDTM V1.2 provides specifications for additional reference start and end dates, an actual arm variable, as well as a death flag and date of death variable. It is strongly preferred that sponsors follow the specifications in this amendment.

DS Domain:

When there is more than one disposition event, the EPOCH variable should be used to aid in distinguishing between them. This will allow identification of the Epoch in which each event occurred. If "DEATH" occurs, it should be the last record and include its associated EPOCH. It is expected that EPOCH variable values will be determined based on the trial design and thus should be defined clearly and documented in the define.xml.

AE Domain (Adverse Events):

There is currently no variable in the AE domain that indicates if an AE was "treatment emergent." CDER would like the AE domain to include all adverse events recorded in any way in the patients' case report forms, regardless of whether the sponsor has determined that particular events were treatment emergent or not. Since judgment is often used to make a determination of whether an event is treatment-emergent, reviewers need to be able to easily identify all reported adverse events

to allow for an evaluation of the sponsor's assessment. It is not acceptable that the AE domain only include those adverse events that a sponsor has decided are "treatment-emergent." A treatment-emergent flag should be included in AE to indicate if the event was or was not treatment emergent, as determined by the sponsor and as used in the sponsor's primary adverse event analyses. For example, depending on the study design, one definition of treatment-emergent might be all adverse events (new events and worsening of pre-existing events/conditions) that occurred after the first dose of study drug exposure and within a certain amount of time following the final dose. Specifications for the treatment emergent variable are contained in Amendment 1 to SDTM V1.2.

In accordance with Amendment 1 to SDTM V1.2, CDER would prefer that sponsors provide the variables for the different levels of MedDRA hierarchy.

The SDTM Implementation Guide states that sponsors have the choice to use secondary mapped SOC in place of primary mapped SOC as they wish in the AEBODSYS variable. CDER does not generally agree with this, however, and needs a dedicated variable for the primary System Organ Class. The AESOC variable entry, as described in Amendment 1 to SDTM V1.2, should contain the MedDRA-defined, primary mapped SOC. If the sponsor used a secondary SOC for a subsequent analysis, then this should be present in the AEBODSYS variable. The inclusion of both AESOC and AEBODSYS allows reviewers to easily determine whether the use of a secondary SOC for analysis was appropriate.

Custom Domains:

Both the SDTM and SEND Implementation Guides allow for the creation of custom domains if the data do not fit into an existing domain. Prior to creating a custom domain, sponsors should confirm that the data do not fit an existing domain and also check the CDISC website for domains added after the most recent published implementation guide. If necessary, sponsors should follow the recommendations in the implementation guides for how to create a custom domain.

LB Domain (Laboratory):

The size of the LB domain is often quite large and can exceed the reviewers' ability to open the file using standard-issue computers. This size issue can be addressed by splitting the large LB dataset into smaller data sets according to LBCAT and LBSCAT, using LBCAT for initial splitting. If the size is still too large, then use LBSCAT for further splitting. For example: use the dataset name lbc.xpt for chemistry lbh.xpt for hematology and lbu.xpt for urinalysis. Splitting it other ways (by subject or file size, etc) makes the data less useable. Sponsors should submit these smaller files **in addition to** the larger non-split standard LB domain file. Individual file size up to 400 megabytes is usually fine; however, it is recommended to confirm this with the review division.

ADAM DATASETS (ANALYSIS DATASETS)

General Comments:

In determining how to create ADaM analysis datasets for submission to CDER, sponsors should refer to **three documents**: the Analysis Data Model and the ADaM Implementation Guide (**www.CDISC.org**), and the FDA Study Data Specifications Document (**http://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM199599.pdf**). Close adherence to the ADaM Implementation Guide is expected and any specific questions that result from attempts to adhere to these documents should be discussed with the review division.

A careful assessment of **which analysis datasets** will be needed should occur. Sponsors must submit analysis datasets with their application to support key efficacy and safety analyses. Additionally, it is important to remember that SDTM datasets do not have core variables (such as demographic and population variables) repeated across the different domains. The need for such duplication of core variables across various domains can be fulfilled through their inclusion in the corresponding analysis datasets. This need is sufficient for the purposes of justifying a request for analysis datasets. For example, the SDTM adverse event dataset does not allow for the inclusion of variables such as treatment arm, sex, age, or race. These and other variables may be included in an adverse event analysis dataset.

ADSL

ADSL is the subject-level analysis dataset for ADaM. CDER expects this ADaM-defined dataset along with the other supporting analysis datasets. All submissions containing standard data are expected to contain an ADSL file for each study. In addition to the variables specified for ADSL in the ADaM Implementation Guide, it is expected that the sponsor will include multiple additional variables representing various important baseline patient characteristics. A few examples could include: disease severity scores such as APACHE scores or FINE scores; baseline organ function measurements such as calculated creatinine clearance or FEV1; range categories for continuous variables; numeric date variables in non-ISO format such as SAS or Oracle.

VARIABLES

Required vs. Expected vs. Permissible

CDISC data standards categorize SDTM and SEND variables as being Required, Expected, and Permissible. Some sponsors have interpreted Permissible variables as being optional. However, for the purposes of SDTM and SEND submissions to CDER, all Permissible variables for which data

were collected or for which derivations are possible should be submitted. **Examples** of some of the Permissible variables in SDTM that CDER expects to see include:

- Baseline flags for Laboratory results, Vital Signs, ECG, Pharmacokinetic Concentrations, and Microbiology results
- EPOCH designators. An extensible codelist for EPOCH will be developed. Please follow CDISC guidance for terminology.
- --DY and --STDY variables in SE or other Findings domains. --STDY should be calculated based on first treatment date.

Naming Conventions

Naming conventions (variable name and label) and variable formats should be followed as specified in the implementation guides.

Dates

Dates in SDTM and SEND domains should conform to the ISO 8601 format. Examples of how to implement this are included in the Implementation Guides. Missing components should be handled by right truncation, and should not be padded with extra zeros.

USUBJID

Each individual subject must be assigned a single unique identifier (USUBJID) across the entire submission. An individual subject should have the exact same unique identifier across all datasets, including SDTM and ADaM. Do not add leading or trailing spaces to the USUBJID variable in any dataset. In some cases, sponsors will submit data in which the USUBJID variable for each individual subject appears to be the same across datasets, however, in certain datasets, the actual entry will have leading zeros added, or zeros added elsewhere in the entry. This does not allow for machine readable matching of individual subject data across all data sets. Improper implementation of the USUBJID variable is a common error that is seen with many applications, and often requires sponsors to re-submit their data.

Imputed Data

SDTM should not include any imputed data. If there is a need for data imputation, this should occur in an analysis dataset, and the relevant supporting documentation to explain the imputation methods must be provided**.**

COMMON ERRORS

In this section, some of the most common data errors that CDER finds when performing conformance checking are described. Each error listed in this section lists the name of the actual

error as the title and then provides some brief comment and/or description. It is expected that has CDER increasing the scope of its data checking activities, this section will be updated based on which errors are seen most frequently.

Define Doesn't Validate

Please refer to **www.cdisc.org/define-xml** for instructions. Here sponsors can find the white paper for XML Schema Validation for Define.xml, which provides guidance on validating define.xml version 1.0 documents against the define.xml XML schemas. Prior to submission, a sponsor may submit their define.xml for testing to determine whether it validates. The submission of a define.xml is expected with all CDISC applications. If sponsors would like to also include a define.pdf document additionally, this would be acceptable.

Invalid ISO 8601 Date Format

All dates in the SDTM domains must conform to the ISO 8601 format.

Begin Date Must be ≤ End Date

This is a common error. Examples include a concomitant medication or adverse event begin date that is after the end date.

Required Variable Not Found

A Required variable is any variable that is basic to the identification of a data record (such as the unique subject identifier) or is necessary to make the record meaningful. Required variables must always be included in the dataset and cannot be null for any record.

Inconsistent Value for Standard Units

For a given test, all values of --STRESU should be the same. In some cases --TESTCD may not be sufficient to uniquely indentify a test.

Invalid value for MedDRA Term

This occurs when the sponsor has not accurately represented the MedDRA term as it appears in the MedDRA terminology. This seems to be more common with the System Organ Class terms, which are commonly longer and more prone to misspellings.