STUDY DATA
TECHNICAL CONFORMANCE GUIDE

Technical Specifications Document

This Document is incorporated by reference into the following Guidance Document(s):

Guidance for Industry *Providing Regulatory Submissions in Electronic Format – Standardized Study Data*

For questions regarding this technical specifications document, contact CDER at cder-edata@fda.hhs.gov or CBER at cber.cdisc@fda.hhs.gov

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)

December 2014

STUDY DATA
TECHNICAL CONFORMANCE GUIDE

December 2014

Revision History

Date	Version	Summary of Revisions
January 2014	1.0	Initial Version
December 2014	2.0	Revisions based on the public comment period (February 2014 – May 2014); and CDER/CBER internal review May 2014 – December 2014

Table of Contents

STUDY DATA
TECHNICAL CONFORMANCE GUIDE

This technical specifications document represents the Food and Drug Administration's (FDA's) current thinking on this topic. It does not create or confer any rights for or on any person and does not operate to bind FDA or the public. You can use an alternative approach if the approach satisfies the requirements of the applicable statutes and regulations. If you want to discuss an alternative approach, contact the FDA staff responsible for implementing this guidance. If you cannot identify the appropriate FDA staff, send an email to cder-edata@fda.hhs.gov or cber.cdisc@fda.hhs.gov.

1. Introduction

1.1 Background

This Study Data Technical Conformance Guide (Guide) provides specifications, recommendations, and general considerations on how to submit standardized study data using FDA-supported[1] data standards located in the **Data Standards Catalog** (*Standards Catalog*).[2] The Guide supplements the guidance for industry *Providing Regulatory Submissions in Electronic Format — Standardized Study Data* (eStudy Data). The eStudy Data guidance will implement the electronic submission requirements of section 745A(a) of the FD&C Act with respect to standardized study data contained in certain investigational new drug applications (INDs), new drug applications (NDAs); abbreviated new drug applications (ANDAs); and certain biologics license applications (BLAs) that are submitted to the Center for Drug Evaluation and Research (CDER) or the Center for Biologics Evaluation and Research (CBER).[3]

1.2 Purpose

This Guide provides technical recommendations to sponsors[4] for the submission of animal and human study data and related information in a standardized electronic format in INDs, NDAs, ANDAs, and BLAs. The Guide is intended to complement and promote interactions between sponsors and FDA review divisions. However, it is not intended to replace the need for sponsors to communicate directly with review divisions regarding implementation approaches or issues relating to data standards. To better understand why the FDA is now emphasizing the submission of standardized data for all studies, please refer to the Appendix.

---

[1] For the purposes of this document, "supported" means the receiving Center has established processes and technology to support receiving, processing, reviewing, and archiving files in the specified file format.
[2] Available at http://www.fda.gov/forindustry/datastandards/studydatastandards/default.htm.
[3] See *Providing Regulatory Submissions in Electronic Format — Standardized Study Data* (section II.A) available at http://www.fda.gov/forindustry/datastandards/studydatastandards/default.htm.
[4] For the purposes of this document, the term "sponsor" refers to both "sponsors" and "applicants" who are submitting study data to the Agency.

Because of the inherent variability across studies and applications, it is difficult to identify all data needed by a review division prior to a scientific regulatory review. We recommend that as early as the pre-IND meeting, sponsors should use the established regulatory process to discuss with the review division the key data necessary to support a submission, the data elements that should be included in each dataset, and the organization of the data within the datasets.

Some data standards may not require the use of all defined data elements to be collected in any given study. For example, the Study Data Tabulation Model Implementation Guide (SDTMIG)[5] classifies variables as required, expected, or permissible. *What* data are collected and submitted is a decision that should be made based on scientific reasons, regulation requirements, and discussions with the review division. However, all study-specific data necessary to evaluate the safety and efficacy of the medical product should be submitted in conformance with the standards currently supported by FDA and listed in the *Standards Catalog*.

If there is a question regarding a specific submission or a particular data standard implementation, the sponsor should contact the review division for specific submission questions or the appropriate contact for data standards issues (cder-edata@fda.hhs.gov or cber.cdisc@fda.hhs.gov).

This Guide supersedes all previous Study Data Specifications documents (Versions 1.0 - 2.0) and CDER Study Data Common Issues Documents (Versions 1.0 -1.1).

1.3 Document Revision and Control

FDA intends to post updated versions of the Guide to the **Study Data Standards Resources Web page** (Standards Web page)[6] followed by *Federal Register* notices announcing updated versions. The revision history page of the Guide will contain sufficient information on the changes made by section.

1.4 Organization and Summary of the Guide

This document is organized as follows:

Section 1: **Introduction** – provides information on regulatory policy and guidance background, purpose, and document control.

Section 2: **Planning and Providing Standardized Study Data** – recommends and provides details on preparing an overall study data standardization plan, a study data reviewer's guide and an analysis data reviewer's guide.

[5] See http://www.cdisc.org.
[6] The Standards Web page can be accessed at http://www.fda.gov/forindustry/datastandards/studydatastandards/default.htm.

Section 3: **Exchange Format - Electronic Submissions** – presents the specifications, considerations, and recommendations for the file formats currently supported by FDA.

Section 4: **Study Data Submission Format: Clinical and Nonclinical** – presents general considerations and specifications for sponsors using, for example, the following standards for the submission of study data: Study Data Tabulation Model (SDTM), Analysis Data Model (ADaM), and Standard for Exchange of Nonclinical Data (SEND).

Section 5: **Therapeutic Area Standards** – presents supplemental considerations and specific recommendations when sponsors submit study data using FDA-supported therapeutic area standards (TA).

Section 6: **Terminology** – presents general considerations and specific recommendations when using controlled terminologies/vocabularies for clinical trial data.

Section 7: **Electronic Submission Format** – provides specifications and recommendations on submitting study data using the electronic Common Technical Document (eCTD) format.

Section 8: **Data Validation and Traceability** – provides general recommendations on conformance to standards, data validation rules, data traceability expectations, and legacy data conversion.

1.5 Relationship to Other Documents

This Guide integrates and updates information discussed previously in the Study Data Specifications and the CDER Common Data Standards Issues documents.[7] As noted above, this Guide supersedes all previous Study Data Specifications documents (Versions 1.0 - 2.0) and CDER Study Data Common Issues Documents (Versions 1.0 -1.1). The examples of issues and concerns discussed in the Guide are intended as examples only of common issues, and not an inclusive list of all possible issues.

This Guide is incorporated by reference into the Guidance to Industry *Providing Regulatory Submissions in Electronic Format: Standardized Study Data*. In addition, sponsors should reference the following:

- FDA Study Data Standards Resources Web page (See section 1.1)
- FDA Data Standards Catalog (See section 1.1)
- FDA Portable Document Format Specifications (See section 3.2)

[7] See http://www.fda.gov/Drugs/DevelopmentApprovalProcess/FormsSubmissionRequirements/ElectronicSubmissions/ucm248635.htm.

- Guidance to Industry Providing Regulatory Submissions in Electronic Format: *Submissions Under Section 745A(a) of the Federal Food, Drug, and Cosmetic Act*[8]
- Guidance to Industry *Providing Regulatory Submissions in Electronic Format: Certain Human Pharmaceutical Product Applications and Related Submissions Using the Electronic Common Technical Document Specifications*[9]

2. Planning and Providing Standardized Study Data

2.1 Study Data Standardization Plan

For clinical and nonclinical studies, sponsors should include a plan (e.g., in the IND) describing the submission of standardized study data to FDA. The Study Data Standardization Plan (*Standardization Plan*) assists FDA in identifying potential data standardization issues early in the development program. Sponsors may also initiate discussions at the pre-IND stage. For INDs, the *Standardization Plan* should be located in the general investigational plan. The *Standardization Plan* should include, but is not limited to the following:

1. List of the planned studies
2. Type of studies (e.g., phase I, II or III)
3. Study designs (e.g., parallel, cross-over, open-label extension)
4. Planned data standards, formats, and terminologies and their versions or a justification of studies that may not conform to the currently supported standards

The *Standardization Plan* should be updated in subsequent communications with FDA as the development program expands and additional studies are planned. Updates to the *Standardization Plan* should not be communicated each time a study is started. The cover letter accompanying a study data submission should describe the extent to which the latest version of the *Standardization Plan* was executed.

2.2 Study Data Reviewer's Guide

The preparation of a Study Data Reviewer's Guide (*SDRG)*[10,] is recommended as an integral part of a standards-compliant study data submission. The *SDRG* should describe, for each study, any special considerations or directions that may facilitate an FDA reviewer's use of the submitted data and may help the reviewer understand the relationships between the study report and the data.[11] For each study, the *SDRG* should include, but is not limited to the following:

---

[8] http://www.fda.gov/downloads/Drugs/Guidances/UCM384686.pdf

[9] http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM333969.pdf

[10] A specific template for a Study Data Reviewer's Guide is not specified. However, an example of a Study Data Reviewer's Guide (template, completion guidelines and examples) can be found at http://www.phusewiki.org/wiki/index.php?title=Study_Data_Reviewer's_Guide.

[11] For submissions to CBER, sponsors and applicants should continue to provide the Data Interpretation and Validation Report (DIVR). The DIVR can be incorporated into the Study Data Reviewer's Guide. The

1. Study protocol title, number, and version
2. Study design
3. Standards, formats, and terminologies and their versions
4. Description of study datasets
5. Data standards validation rules, versions, and issues
6. Description of all sponsor decisions related to data standard implementations

A *SDRG* for each nonclinical study and for each clinical study should be placed in Module 4 and 5, respectively, of the Electronic Common Technical Document (eCTD).[12]

2.3 Analysis Data Reviewer's Guide

The preparation of an Analysis Data Reviewer's Guide (ADRG)[13] is recommended as an important part of a standards-compliant analysis data submission. The ADRG provides FDA reviewers with context for analysis datasets and terminology, received as part of a regulatory product submission, additional to what is presented within the data definition file (i.e., define.xml). The ADRG also provides a summary of ADaM conformance findings. The ADRG purposefully duplicates limited information found in other submission documentation (e.g., the protocol, statistical analysis plan, clinical study report, define.xml) in order to provide FDA reviewers with a single point of orientation to the analysis datasets. It should be noted that the submission of an ADRG does not eliminate the requirement to submit a complete and informative define.xml file corresponding to the analysis datasets.

An *ADRG* for clinical study should be placed in Module 5 of the Electronic Common Technical Document (eCTD).

3. Exchange Format – Electronic Submissions

3.1 Extensible Mark-up Language

Extensible Mark-up Language (XML), as defined by the World Wide Web Consortium (W3C), specifies a set of rules for encoding documents in a format that is both human-readable and machine-readable.[14,15] XML's primary purpose is to facilitate the sharing of structured data across different information systems. An XML use case is CDISC's define.xml file. All XML files should use .xml as the file

---

DIVR can be found at
http://www.fda.gov/BiologicsBloodVaccines/DevelopmentApprovalProcess/ucm209137.htm

[12] The Study Data Reviewer's Guides are separate from an overall reviewer's guide that may be placed in Module 1 of the eCTD.

[13] A specific template for an Analysis Data Reviewer's Guide is not specified. However, an example of an Analysis Data Reviewer's Guide (template, completion guidelines and examples) can be found at http://www.phusewiki.org/wiki/index.php?title=Analysis_Data_Reviewer's_Guide.

[14] See http://en.wikipedia.org/wiki/XML.

[15] See http://www.w3.org/XML/.

extension. Although XML files can be compressed, the define.xml should not be compressed.

3.2 Portable Document Format

Portable Document Format (PDF) is an open file format used to represent documents in a manner independent of application software, hardware, and operating systems.[16] A PDF use case includes, e.g., the annotated CRF (aCRF / blankcrf), and other documents that align with the International Conference on Harmonization (ICH) M2.[17] FDA PDF specifications are located on FDA's Electronic Common Technical Document (eCTD) Web site.[18] The *Standards Catalog* lists the PDF version(s) that are supported by FDA. All PDF files should use .pdf as the file extension.

3.3 File Transport Format

3.3.1 SAS Transport Format

The SAS Transport Format (XPORT) Version 5, is the file format for the submission of all electronic datasets.[19] The XPORT is an open file format published by SAS Institute for the exchange of study data. Data can be translated to and from XPORT to other commonly used formats without the use of programs from SAS Institute or any specific vendor. There should be one dataset per transport file, and the dataset in the transport file should be named the same as the transport file (e.g., "AE" and AE.xpt, "SUPPAE" and SUPPAE.xpt).

XPORT files can be created by the COPY Procedure in SAS Version 5, Version 6 and higher of the SAS Software. SAS Transport files processed by the SAS CPORT cannot be reviewed, processed, or archived by FDA. Sponsors can find the record layout for SAS XPORT transport files through SAS technical document TS-140.[20] All SAS XPORT transport files should use .xpt as the file extension. There should be one dataset per XPORT file and the files should not be compressed.

3.3.2 Dataset Size

Each dataset should be provided in a single transport file. The maximum size of an individual dataset that FDA can process depends on many factors. Datasets greater than 1 gigabyte (gb) in size should be split into smaller datasets no larger than 1 gb. Sponsors should submit these smaller datasets in addition to the larger non-split

---

[16] Adobe Systems Incorporated, PDF Reference, sixth edition, version 1, Nov. 2006, p. 33.
[17] See http://www.ich.org/products/electronic-standards.html.
[18] Available at http://www.fda.gov/Drugs/DevelopmentApprovalProcess/FormsSubmissionRequirements/ElectronicSubmissions/ucm153574.htm
[19] See http://www.sas.com
[20] Available at http://support.sas.com/techsup/technote/ts140_2.pdf

datasets. The split datasets should be placed in a separate sub-directory labeled “split” (See section 7).

3.3.3 Dataset Column Length

The allotted length for each column containing character (text) data should be set to the maximum length of the variable used across all datasets in the study. This will significantly reduce file sizes. For example, if USUBJID has a maximum length of 18, the USUBJID’s column size should be set to 18, not 200.

3.3.4 Variable and Dataset Descriptor Length

The length of variable names, descriptive labels, and dataset labels should not exceed the maximum permissible number of characters described below.

Table 1: Maximum Length of Variables and Dataset Elements

Element	Maximum Length in Characters
Variable Name	8
Variable Descriptive Label	40
Dataset Label	40

3.3.5 Special Characters: Variables and Datasets

Variable names, as well as variable and dataset labels should include American Standard Code for Information Interchange (ASCII) text codes only.

3.3.6 Variable and Dataset Names

Variable and dataset names should not contain punctuation, dashes, spaces, or other non-alphanumeric symbols. In addition, the variable and dataset names should not contain special characters, including:

\ / * ,? < > | “ ‘ : % # + ( ) { } [ ]

3.3.7 Variable and Dataset Labels

Variable and dataset labels can include punctuation characters. However, special characters should not be provided, such as,

1. Unbalanced apostrophe, e.g., Parkinson's.
2. Unbalanced single and double quotation marks.
3. Unbalanced parentheses, braces or brackets, e.g., ‘(‘, ‘{‘and ‘[‘.
4. ‘<’ less-than sign and ‘>’ greater-than sign.

4. Study Data Submission Format – Clinical and Nonclinical

4.1 Clinical Data Interchange Standards Consortium

Clinical Data Interchange Standards Consortium **(**CDISC) is an open, multidisciplinary, neutral, nonprofit standards development organization (SDO) that has been working through consensus-based collaborative teams to develop global data standards for clinical and nonclinical research.[21]

Data format specifications for the tabulation datasets of clinical and nonclinical toxicology studies are provided by SDTM and SEND, respectively. ADaM provides the clinical data format specifications for analysis datasets. Currently, ADaM specifications for SEND have not been developed. As noted in Section 1.1, the *Standards Catalog* provides a listing of the currently supported data standards with links to reference materials.

Although the SDTM and SEND formats facilitate review of the data, they do not always provide the data structured in a way that supports all analyses needed for review. Analysis files are critical for FDA to understand, on a per subject basis, how the specific analyses contained in the study report have been created. Therefore, sponsors should supplement the SDTM with ADaM analysis datasets as described below.

There may be instances in which current implementation guides (e.g., SDTMIG, SENDIG) do not provide specific instruction as to how certain study data should be represented. In these instances, sponsors should discuss their proposed solution with the review division and submit supporting documentation that describes these decisions or solutions in the *SDRG* at the time of submission.

4.1.1 Study Data Tabulation Model

4.1.1.1 Definition

The Study Data Tabulation Model (SDTM) defines a standard structure for human clinical trials tabulation datasets.

4.1.1.2 SDTM General Considerations

It is recommended that sponsors implement the SDTM standard for representation of clinical trial tabulation data prior to the conduct of the study. The use of case report forms that incorporate SDTM standard data elements (e.g., Clinical Data Acquisition Standards Harmonization (CDASH)) allows for a simplified process for the creation of SDTM domains.

The SDTMIG should be followed unless otherwise indicated in this Guide or in the *Standards Catalog*. The conformance criteria listed in the SDTMIG should not be

[21] See http://www.cdisc.org.

interpreted as the sole determinant of the adequacy of submitted data. If there is uncertainty regarding implementation, the sponsor should discuss application-specific questions with the review division and general standards implementation questions with the specific center resources identified elsewhere in this Guide (See section 1.2). When data imputation is utilized, sponsors should submit imputed data in an analysis dataset, and the relevant supporting documentation (e.g., ADRG, define.xml) explaining the imputation methods.

Except for variables that are defined in the SDTMIG as being coded, no numerically coded variables should typically be submitted as part of the SDTM datasets. Numeric values generated from validated scoring instruments or questionnaires do not represent codes, and therefore have no relevance for this issue. There may be special instances when codes are preferred, hence sponsors should refer to the review division for direction, if there are any questions.

Subject Identifier (SUBJID)

The SUBJID is an ID of the entity (i.e., person) that participates in a trial. If the same subject is screened more than once in a trial, then the subject's SUBJID should be different.

Unique Single Identifier (USUBJID)

Each individual subject should be assigned a single unique identifier across the entire application. This is in addition to the subject ID (SUBJID) used to identify subjects in each study and its corresponding study report. An individual subject should have the exact same unique identifier across all datasets, including between SDTM and ADaM datasets. Subjects that participate in more than one study should maintain the same USUBJID across all studies. It is important to follow this convention to enable pooling of a single subject's data across studies (e.g., a randomized control trial and an extension study).

Sponsors should not add leading or trailing spaces to the USUBJID variable in any dataset. For example, applications have been previously submitted in which the USUBJID variable for each individual subject appeared to be the same across datasets; however, in certain datasets, the actual entry had leading zeros added, or zeros added elsewhere in the entry. This does not allow for machine-readable matching of individual subject data across all datasets. Improper implementation of the USUBJID variable is a common error with applications and often requires sponsors to re-submit their data.

Adjudication Data

There are no existing standards or best practices for the representation of adjudication data as part of a standard data submission. Until standards for adjudication data are developed, it is advised that sponsors discuss their proposed approach with the review division and also include details about the presence, implementation approach, and location of adjudication data in the *SDRG*.

4.1.1.3 SDTM Domain Specifications

SUPPQUAL (Supplemental Qualifier)
A SUPPQUAL dataset is a special SDTM dataset that contains non-standard variables which cannot be represented in the existing SDTM domains. SUPPQUAL should be used only when key data cannot be represented in SDTM domains. In general, variables used to support key analyses should not be represented in SUPPQUAL. Discussion with the review division should occur if the sponsor intends to include important variables (e.g., that support key analyses) in SUPPQUAL datasets, and reflected in the *SDRG*.

DM Domain (Demographics)
In the DM domain, each subject should have only one single record per study.

Screen failures, when provided, should be included as a record in DM with the ARM field left blank. For subjects who are randomized in treatment group but not treated, the planned arm variables (ARM and ARMCD) should be populated, but actual treatment arm variables (ACTARM and ACTARMCD) should be left blank.[22]

DS Domain (Disposition)
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 a death of any type occurs, it should be the last record and should 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.

SE Domain (Subject Elements)
The Subject Elements domain should be included to aid in the association of subject data (e.g., findings, events, and interventions) with the study element in which they occurred.

AE Domain (Adverse Events (AE))
Currently, there is no variable in the AE domain that indicates if an AE was "treatment-emergent." The AE domain should include all adverse events that were recorded in the subjects' case report forms, regardless of whether the sponsor determined that particular events were or were not treatment-emergent.

The entry of a "Y" for the serious adverse event variable, AESER, should have the assessment indicated, (e.g., as a death, hospitalization, or disability/permanent damage). Frequently, sponsors omit the assessment information, even when it has been collected on the CRF. The criteria that led to the determination should be

[22] Although this convention is inconsistent with the SDTMIG, FDA recommends its use so that "Screen Failure" is not specified as a treatment arm.

provided. This information is critical during FDA review to support the characterization of serious AEs.

Custom Domains
The SDTMIG permits 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 into an existing domain. If it is necessary to create custom domains, sponsors should follow the recommendations in the STDMIG. In addition, sponsors should present their implementation approach in the *SDRG*.

LB Domain (Laboratory)
The size of the LB domain dataset submitted by sponsors is often too large to process (See section 3.3.2). This issue can be addressed by splitting a large LB dataset into smaller datasets 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 lb1.xpt for chemistry, lb2.xpt for hematology, and lb3.xpt for urinalysis. Splitting the dataset in other ways (e.g., by subject or file size) makes the data less useable. Sponsors should submit these smaller files in addition to the larger non-split standard LB domain file. Sponsors should submit the split files in a separate sub-directory/split that is clearly documented in addition to the non-split standard LB domain file in the SDTM datasets directory (See section 7).

Trial Design
Trial Design datasets provide a way to describe the planned conduct of a clinical trial and should be included in SDTM submissions.

4.1.2 Analysis Data Model

4.1.2.1 Definition

Specifications for analysis datasets for human drug product clinical studies are provided by the Analysis Data Model (ADaM) and its implementation by the ADaMIG. Analysis datasets should be used to create and to support the results in clinical study reports, Integrated Summaries of Safety (ISS), and Integrated Summaries of Efficacy (ISE), as well as other analyses required for a thorough regulatory review. Analysis datasets can contain imputed data or data derived from SDTM datasets.

4.1.2.2 General Considerations

Generally, ADaM analysis datasets facilitate FDA review. However, it does not always provide data structured in a way that supports all of the analyses that should be submitted for review. For example, ADaM does not support simultaneous analysis of multiple dependent variables or correlation analysis across several response variables. Therefore, sponsors should, as needed, supplement their ADaM datasets after discussions with the specific review division.

One of the expected benefits of analysis datasets that conform to ADaM is that they simplify the programming steps necessary for performing an analysis. As noted above, ADaM datasets should be derived from the data contained 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. To ensure traceability, all SDTM variables utilized for variable derivations in ADaM should be included in the ADaM datasets. Each analysis dataset that is submitted should be described accordingly with adequate metadata in the define.xml file.

4.1.2.3 Key Efficacy and Safety Variables

Sponsors should submit ADaM datasets to support key efficacy and safety analyses. At least one dataset should be referenced in the define file as containing the primary efficacy variables. Further, primary and secondary variables and their derivations (as applicable) should be provided, as well as documented in the define file.

4.1.2.4 Timing Variables

A variable for relative day of measurement or event, along with timing variables for visit should be included when an analysis dataset contains multiple records per subject (i.e., repeated measures data). In addition to a protocol-scheduled visit variable, sponsors should include at least two additional timing variables: a character variable describing the visit (e.g., WEEK 8) and a corresponding numeric variable (e.g., 8). These two variables may represent measures of real time from randomization. The reason for this request is related to a common analysis (i.e., the portrayal of data over the duration of a study). These data are often presented as means or medians by treatment group and by "time on study treatment." The "time on study treatment" variable is defined by windowed visits that might be represented in ADaM datasets as AVISIT (a character variable) and its numeric analog AVISITN. AVISIT / AVISITN are adequate for sub-setting the data. However, in certain circumstances, such as extension studies, AVISIT / AVISITN are not an adequate measure of real time on study treatment, and so it is not sufficient for plotting data where the x-axis measures real time.

4.1.2.5 Core Variables

Core variables, including all covariates presented in the study protocol, should be listed after the key variables (USUBJID and visit) and included in each ADaM dataset. These core variables are typically included in the Analysis Data Subject Level (ADSL) dataset (See section 4.1.2.9.1). Examples of key core variables include study/protocol number, center/site number, geographic region, country, treatment assignment information, sex, age, race, analysis population flags (e.g., Intent-to-Treat (ITT), Full Analysis Set (FAS), Safety, Per-Protocol (PP)), and other important baseline demographic variables. Note that all variables that contain coded data should be accompanied by a variable that provides the decoded information.

In addition, it is important to note that SDTM datasets do not have core variables (such as demographic and population variables) repeated across the different

domains. The duplication of core variables across various domains can be fulfilled through their inclusion in the corresponding 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 should be included in an adverse event analysis dataset.

4.1.2.6 Dates

Dates should be formatted as numeric in the analysis datasets. In addition to including dates formatted as International Standards Organization (ISO) 8601, corresponding numeric dates should be included in the analysis datasets. The specific date of reference used to calculate numeric dates varies by software; hence this date of reference should be specified within define.xml. In the event of partial dates, imputation should be performed and appropriate corresponding ADaM imputation flags utilized.

4.1.2.7 Labels

Each dataset should be described by an internal label that is shown in the define.xml file. The label names of analysis datasets should be different from 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 dataset label, such as "Adverse Events."

4.1.2.8 Software Programs

Sponsors should provide the software programs used to create ADaM datasets and corresponding tables and figures to help reviewers to better understand how the datasets, tables and figures were created. Any submitted programs (scripts) generated by an analysis tool should be provided as ASCII text files or PDF files and should include sufficient documentation to allow a reviewer to understand the submitted programs. If the submitted programs created by the analysis tool uses a file extension other than .txt or .pdf, then the file name should include the native file extension.

4.1.2.9 ADaM Domain Specifications

4.1.2.9.1 Analysis Data Subject Level

Analysis Data Subject Level (ADSL) is the subject-level analysis dataset for ADaM. All submissions containing standard data should contain an ADSL file for each study. In addition to the variables specified for ADSL in the ADaMIG such as those previously listed in the core variables section (See section 4.1.2.5) above, the sponsor should include multiple additional variables representing various important baseline subject characteristics / covariates presented in the study protocol. Some examples of baseline characteristics / covariates include, but are not limited to, disease severity scores such as Acute Physiology and Chronic Health Evaluation (APACHE) scores,[23] baseline organ function measurements such as calculated creatinine clearance or

[23] Knaus WA, Draper EA, Wagner DP, Zimmerman JE (1985). "APACHE II: a severity of disease classification system." Critical Care Medicine, 13 (10): 818–829.29.

Forced Expiratory Volume in 1 second (FEV1); range categories for continuous variables; and numeric date variables in non-ISO formats.

4.1.2.9.2 Imputed Data

When data imputation is utilized, sponsors should submit imputed data in an analysis dataset, and the relevant supporting documentation (e.g., define.xml and ADRG) explaining the imputation methods.

4.1.3 Standard for Exchange of Nonclinical Data

4.1.3.1 Definition

The Standard for Exchange of Nonclinical Data (SEND) provides the organization, structure, and format of standard nonclinical (animal toxicology studies) tabulation datasets for regulatory submission. Currently, the SEND Implementation Guide (SENDIG) supports single-dose general toxicology, repeat-dose general toxicology, and carcinogenicity studies.

4.1.3.2 General Considerations

The SENDIG provides specific domain models, assumptions, conformance and business rules, and examples for preparing standard tabulation datasets that are based on the SDTM. If there is uncertainty regarding SEND implementation, the sponsor should discuss the issue with the review division.

The ideal time to implement SEND is prior to the conduct of the study as it is very important that the results presented in the accompanying study report be traceable back to the original data collected.

4.1.3.3 SEND Domain Specification

<u>SUPPQUAL (Supplemental Qualifier)</u>

A SUPPQUAL dataset is a special SEND dataset that contains non-standard variables which cannot be represented in the existing SEND domains. Discussion with the review division should occur if the sponsor intends to include important variables (i.e., that support key analyses) in SUPPQUAL datasets and this should be reflected in the *SDRG*.

Currently, SUPPQUAL should be used to capture some collected information (e.g., pathology modifiers) until the SEND is further refined to adequately represent such information.

<u>Microscopic Findings (MI) Domain</u>

Sponsors should ensure that the transformation of findings from MIORRES to MISTRESC closely adheres to the instructions in the SENDIG. Modifiers for which there are variables available (e.g. MISEV, MILAT, etc.) should be placed appropriately. There should be no severities (e.g., minimal, mild, etc.) included in MISTRESC. Sponsors should use the VISITDY variable if postmortem findings

which were intended to be analyzed together were collected across multiple study days.

Macroscopic Findings (MA) Domain
Sponsors should use the VISITDY variable if postmortem findings which were intended to be analyzed together were collected across multiple study days.

Custom Domains
The SENDIG allows for the creation of custom domains if the data do not fit into an existing domain.

Tumor Dataset
Carcinogenicity studies should include an electronic dataset of tumor findings to allow for a complete review. At this time sponsors should include a tumor.xpt file while following the specification in the SENDIG for its creation regardless of whether or not the study is in SEND format (See www.cdisc.org/send).

4.1.4 General Considerations: SDTM, SEND, ADaM

4.1.4.1 Variables: Required, Expected, and Permissible

CDISC data standards categorize SDTM and SEND variables as being Required, Expected, and Permissible. In some instances, sponsors have interpreted Permissible variables as being optional and, in other cases, sponsors have excluded Expected variables. For the purposes of SDTM and SEND submissions, all Required, Expected, and Permissible variables that were collected, plus any variables that are needed to compute derivations, should be submitted.[24]

SDTM datasets should not contain imputed data. FDA recognizes that SDTM contains certain operationally derived variables that have standard derivations across all studies (e.g., --STDY, EPOCH). If the data needed to derive these variables are missing, then these variables cannot be derived and the values should be null. The following are examples of some of the Permissible and Expected variables in SDTM and SEND that should be included, if available:

1. Baseline flags (e.g., last non-missing value prior to first dose) for Laboratory results, Vital Signs, ECG, Pharmacokinetic Concentrations, and Microbiology results. Currently, for SDTM, baseline flags should be submitted if the data were collected or can be derived.

2. EPOCH designators. Please follow CDISC guidance for terminology.[25] The variable EPOCH should be included for clinical subject-level observation (e.g., adverse events, laboratory, concomitant medications, exposure, vital signs). This

[24] See CDISC SDTM Implementation Guide at www.cdisc.org for additional information on variables referenced throughout this Guide
[25] See http://www.cancer.gov/cancertopics/terminologyresources/page6.

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 the actual intervention the subject experienced during that phase.

3. Whenever --DTC, --STDTC or --ENDTC, which have the role of timing variables, are included, the matching Study Day variables (--DY, --STDY, or --ENDY, respectively) should be included. For example, in most Findings domains, --DTC is Expected, which means that --DY should also be included. In the Subject Visits domain, SVSTDTC is Required and SVENDTC is Expected; therefore, both SVSTDY and SVENDY should be included.

4.1.4.2 Dates

Dates in SDTM and SEND domains should conform to the ISO 8601 format. Examples of how to implement dates are included in the SDTMIG and SENDIG.[26]

4.1.4.3 Naming Conventions

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

4.1.4.4 SDTM / SEND / ADaM Versions

When submitting clinical and nonclinical data, sponsors should not mix versions within a study. As noted above, the *Standards Catalog* lists the versions that are supported by FDA.

Conversions to one standardized version should be described in the *SDRG*, including the rationale for the conversion.

4.1.4.5 Data Definition File

The data definition file, define.xml, describes the metadata of the submitted SDTM, SEND, and ADaM datasets.

A properly functioning define.xml file is an important part of the submission of electronic study datasets. In addition to the define.xml, a printable define.pdf should be provided if the define.xml cannot be printed.[27] To confirm that a define.xml is printable within the CDER IT environment, it is recommended that the sponsor submit a test version to cder-edata@fda.hhs.gov prior to application submission. If a define.xml version 2.0 or later version is submitted, then a define.pdf does not need to be included in the submission.

---

[26] See http://www.cdisc.org

[27] Detailed FDA PDF specifications are located on FDA's Electronic Common Technical Document (eCTD) Web site, http://www.fda.gov/drugs/developmentapprovalprocess/formssubmissionrequirements/electronicsubmissions/ucm153574.htm

Sponsors should make certain that the code list, origin, and derivation for each variable are clearly and easily accessible from the define file. An insufficiently documented define file is a common deficiency that reviewers have noted. The version of any external dictionary should be clearly stated both in the define.xml and, where possible, in the updated Trial Summary (TS) domain (SDTMIG 3.1.2 or greater; SENDIG 3.0 or greater).

The data definition specification for submitted datasets should be included in the define file in XML format.[28] The specification defines the metadata structures that should be used to describe the datasets and variables. The *Standards Catalog* lists the currently supported version(s) of define.xml. Sponsors should include a reference to the style sheet as defined in the specification and place the corresponding style sheet in the same submission folder as the define.xml file.

The internal dataset label should clearly describe the contents of the dataset. For example, the dataset label for an efficacy dataset might be "Time to Relapse (Efficacy)."

4.1.4.6 Annotated Case Report Form (aCRF)

An Annotated Case Report Form (aCRF) is a PDF document that maps the data collection fields used to capture subject data (electronic or paper) to the corresponding variables or discrete variable values contained within the datasets. Regardless of whether the clinical database is legacy or SDTM compliant, an aCRF should be submitted. The aCRF should be provided as a PDF with the file name "acrf.pdf."[29] The SDTM Metadata Submission Guidelines should be used for additional information on annotated CRFs.[30]

The aCRF should include treatment assignment forms, when applicable, and should map each variable on the CRF to the corresponding variables in the datasets (or database). The aCRF should include the variable names and coding for each CRF item.

When data are recorded on the CRF but are not submitted, the CRF should be annotated with the text "NOT SUBMITTED." There should be an explanation in the *SDRG* stating why data have not been submitted.

[28] See http://www.cdisc.org/define-xml.
[29] Previously acrf.pdf was called blankcrf.pdf.
[30] See Study Data Tabulation Model Metadata Submission Guidelines (SDTM-MSG) (http://www.cdisc.org/sdtm).

5. Therapeutic Area Standards

This section is reserved for future comments, recommendations, and preferences on therapeutic area data standards.

6. Terminology

6.1 General

Common dictionaries should be used across all clinical studies and throughout the submission for each of the following: adverse events, concomitant medications, procedures, indications, study drug names, and medical history. FDA recommends that sponsors use, where appropriate, the terminologies supported and listed in the *Standards Catalog*. It is important that coding standards, if they exist, be followed (e.g., ICH MedDRA Term Selection: Points-to-Consider document). Frequently, sponsors submit data that do not conform to terminology standards, for example, misspelling of MedDRA or WHO Drug terms, lack of conformance to upper / lower case, or the use of hyphens. All controlled terms submitted in datasets should conform to the exact case and spelling used by the terminology maintenance organization (e.g., MedDRA, CDISC controlled terminology). These conformance issues make it difficult to use or develop automated review and analysis tools. The use of a dictionary that is sponsor-defined or an extension of a standard dictionary should be avoided if possible, but, if essential, its use should be documented in the define.xml file and the *SDRG*.

6.1.1 Controlled Terminologies

Controlled terminology standards are an important component of study data standardization and are a critical component of achieving semantically interoperable data exchange (See Appendix). Generally, controlled terminology standards specify the key concepts that are represented as definitions, preferred terms, synonyms, codes, and code system.

The analysis of study data is greatly facilitated by the use of controlled terms for clinical or scientific concepts that have standard, predefined meanings and representations. Standard terminology for adverse events perhaps represents the earliest example of using standards for study data. For example, *myocardial infarction* and *heart attack* are synonyms, and as such should be mapped to the same term in a standard dictionary. This level of standardization facilitates an efficient analysis of events that are coded to the standard term. In electronic study data submissions, sponsors should provide the actual verbatim terms that were collected (e.g., on the case report form), as well as the coded term.

Controlled terminology is also useful when consistently applied across studies to facilitate integrated analyses (that are stratified by study) and cross-study comparative analyses (e.g., when greater statistical power is needed to detect important safety signals). Cross-study comparisons and pooled integrated analyses occasionally provide critical information for regulatory decisions, such as statistical results that

support effectiveness,[31] as well as important information on exposure-response relationships[32] and population pharmacokinetics[33].

6.1.2 Use of Controlled Terminologies

FDA recognizes that studies are conducted over many years, during which time versions of a terminology may change. Generally, FDA expects sponsors to use the most current version of an FDA-supported terminology available at the time of coding. It is acceptable to have different studies use different versions of the same dictionary within the same application. There are some situations where it may be acceptable to use a single older version of a dictionary across multiple studies, even though that version may not be the most current for the later studies. The study data submission should describe the impact, if any, of the older version on the study results in the SDRG. For example, if the sponsor anticipates pooling coded data across multiple studies, then it may be desirable to use a single version across those studies to facilitate pooling. If a sponsor selects this approach, then the approach and the justification should be documented in the *Standardization Plan*, or in an update to the plan.

Regardless of the specific versions used for individual studies, pooled analyses of coded terms across multiple studies (e.g., for an integrated summary of safety) should be conducted using a single version of a terminology. This will ensure a consistent and coherent comparison of clinical and scientific concepts across multiple studies. Sponsors should specify the terminologies and versions used in the study in the *SDRG*.

6.1.3 Maintenance of Controlled Terminologies

The use of supported controlled terminologies is recommended wherever available. If a sponsor identifies a concept for which no standard term exists, FDA recommends that the sponsor submit the concept to the appropriate terminology maintenance organization as early as possible to have a new term added to the standard dictionary. FDA considers this *good terminology management practice.* The creation of custom terms (i.e., so-called *extensible* code lists) for a submission is discouraged, because this does not support semantically interoperable study data exchange. Furthermore,

---

[31] See the guidance for industry *Providing Clinical Evidence of Effectiveness for Human Drugs and Biological Products*, available at http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm072008.pdf. We update guidances periodically. To make sure you have the most recent version of a guidance, check the FDA Drugs guidance Web page at http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm.

[32] See the guidance for industry *Exposure-Response Relationships — Study Design, Data Analysis, and Regulatory Applications*, http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm072109.pdf.

[33] See the guidance for industry *Population Pharmacokinetics*, available at http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm072137.pdf.

the use of custom or "extensible" code lists should not be interpreted to mean that sponsors may substitute their own nonstandard terms in place of existing equivalent standardized terms. Terminology maintenance organizations generally have well-defined change control processes. Sponsors should allow sufficient time for a proposed term to be reviewed and included in the terminology, as it is desirable to have the term incorporated into the standard terminology before the data are submitted. If custom terms cannot be avoided, the submitter should clearly identify and define them within the submission, reference them in the *SDRG*, and use them consistently throughout the application.

If a sponsor identifies an entire information domain[34] for which FDA has not accepted a specific standard terminology, they may select a standard terminology to use, if one exists. FDA recommends that sponsors include this selection in the *Standardization Plan* (See section 2.1) or in an update to the existing plan, and reference it in the *SDRG*. If no controlled terminology exists, the sponsor may define custom terms. The non-FDA supported terms (whether from a non-supported standard terminology or sponsor-defined custom terms) should then be used consistently throughout all relevant studies within the application.

6.2 CDISC Controlled Terminology

Sponsors should use the terminologies and code lists in the CDISC Controlled Terminology, which can be found at the NCI (National Cancer Institute) Enterprise Vocabulary Services.[35] For variables for which no standard terminology exists, or if the available terminology is insufficient, the sponsor should propose its own terminology. The sponsor should provide this information in the define.xml file and in the *SDRG*.

6.3 Adverse Events

6.3.1 MedDRA

6.3.1.1 General Considerations

MedDRA should be used for coding adverse events. The spelling and capitalization of MedDRA terms should match the way the terms are presented in the MedDRA dictionary (e.g., spelling and case). Common errors that have been observed include the incorrect spelling of a System Organ Class (SOC) and other MedDRA terms.

Generally, the studies included in an application are conducted over many years and may have used different MedDRA versions. To avoid potential confusion or incorrect results, the preparation of the adverse event dataset for the ISS should include MedDRA Preferred Terms from a single version of MedDRA. The reason for an ISS based on a single version of MedDRA is that reviewers often analyze adverse

[34] By *information domain*, we mean a logical grouping of clinical or scientific concepts that are amenable to standardization (e.g., adverse event data, laboratory data, histopathology data, imaging data).
[35] See http://www.cancer.gov/cancertopics/terminologyresources/page6.

events across studies, including the use of Standardized MedDRA Queries.[36] In addition, sponsors should use the MedDRA-specified hierarchy of terms. The SDTM variables for the different hierarchy levels should represent MedDRA-specified primary SOC-coded terms.

6.4 Medications

6.4.1 FDA Unique Ingredient Identifier

6.4.1.1 General Considerations

The FDA Unique Ingredient Identifier (UNII)[37] should be used to identify active ingredients (specifically, active moieties) that are administered to investigational subjects in a study (either clinical or nonclinical). This information should be provided in the SDTM Trial Summary (TS) domain. UNIIs should be included for all active moieties of investigational products (TSPARM=TRT or TRTUNII), active comparators (TSPARM=COMPTRT), and any protocol-specified background treatments (TSPARM=CURTRT).

If a medicinal product has more than one active moiety, then multiple records in TS should be provided, one for each active moiety. For example, if the investigational product is Bactrim (a combination of sulfamethoxazole and trimethoprim), then TS will contain two records for TSPARM=TRT: one for sulfamethoxazole and one for trimethoprim.

The preferred substance names and UNII codes can be found by searching FDA's Substance Registration System, hosted by the National Library of Medicine.[38] We recognize that unapproved substances may not yet have registered UNII codes. We recommend that sponsors obtain UNII codes for unapproved substances as early in drug development as possible, so that relevant information, such as study data, can be unambiguously linked to those substances.

6.4.2 WHO Drug Dictionary

6.4.2.1 General Considerations

World Health Organization (WHO) Drug Dictionary[39] is a dictionary maintained and updated by Uppsala Monitoring Centre. WHO Drug Dictionary contains unique product codes for identifying drug names and evaluating medicinal product information, including active ingredients and therapeutic uses.

Typically, WHO Drug is used to code concomitant medications. --DECOD should be populated with the generic name from the WHO dictionary, and --CLAS populated with the drug class, if the utilized dictionary codes drugs to a single class. When

[36] See http://www.meddra.org/standardised-meddra-queries.
[37] See http://www.fda.gov/ForIndustry/DataStandards/SubstanceRegistrationSystem-UniqueIngredientIdentifierUNII/
[38] The Substance Registration System can be accessed at http://fdasis.nlm.nih.gov/srs
[39] See http://www.who-umc.org/

using WHODRUG, generally, --CLAS would not be filled because a drug may have multiple classes. However, one Anatomic Therapeutic Code (ATC) level 4 code could be mapped to --CLAS and the remainder of the ATC codes could be placed in SUPPCM.

6.5 Pharmacologic Class

6.5.1 National Drug File -- Reference Terminology

6.5.1.1 General Considerations

The Veterans Administration's National Drug File – Reference Terminology (NDF-RT)[40] should be used to identify the pharmacologic class(es) of all active investigational substances that are used in a study (either clinical or nonclinical). This information should be provided in the SDTM Trial Summary (TS) domain. The information should be provided as one or more records in TS, where TSPARM=PCLAS.

Pharmacologic class is a complex concept that is made up of one or more component concepts: mechanism of action (MOA), physiologic effect (PE), and chemical structure (CS).[41] The established pharmacologic class is generally the MOA, PE, or CS term that is considered the most scientifically valid and clinically meaningful. Sponsors should include in TS the established pharmacologic class of all active moieties of investigational products used in a study. FDA maintains a list of established pharmacologic classes of approved moieties.[42] If the established pharmacologic class is not available for an active moiety, then the sponsor should discuss the appropriate MOA, PE, and CS terms with the review division. For unapproved investigational active moieties where the pharmacologic class is unknown, the PCLAS record may not be available.

6.6 Indication

6.6.1 SNOMED CT

6.6.1.1 General Considerations

The International Health Terminology Standards Organization's (IHTSDO) Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT)[43] should be used to identify the medical condition or problem that the investigational product in a study is intended to affect (treat, diagnose or prevent, i.e., the indication). This information should be provided in the SDTM Trial Summary (TS) domain as a record where TSPARM=INDIC and TSPARM=TDIGRP. SNOMED CT was chosen to

---

[40] See http://mor.nlm.nih.gov/download/rxnav/NdfrtAPIs.html#

[41] See the guidance for industry and review staff Labeling *for Human Prescription Drug and Biologic Products —Determining Established Pharmacologic Class for Use in the Highlights of Prescribing Information*, available at http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm186607.pdf.

[42] Available at http://www.fda.gov/downloads/ForIndustry/DataStandards/StructuredProductLabeling/UCM346147.zip

[43] http://www.ihtsdo.org/snomed-ct/.

harmonize with Indication information in Structured Product Labeling (SPL)[44]. A reviewer should be able to take the indication term from product labeling and readily search for clinical or nonclinical studies of that indication without having to translate.

7. Electronic Submission Format

Study datasets and their supportive files should be organized into a specific file directory structure when submitted in the eCTD[45] format (see Figure 1 and Table 2 below). Note that this structure is distinct from the eCTD headings and hierarchy folder structure, and does not affect it. Submission of files within the appropriate folders allows automated systems to detect and prepare datasets for review, and minimizes the need for manual processing.

The define.xml and supportive style sheet should reside in the same folder as the datasets they pertain to (e.g., for SDTM, place in "tabulations\sdtm\"). Do not submit empty file folders. Do not submit additional subfolders. If you feel that additional folders are needed, please consult with the appropriate center in advance for guidance.

If you need to split a file that exceeds file size limits (See section 3.3.2), you should submit the smaller split files in the "split" sub-folder in addition to the larger non-split file in the original data folder.

All datasets should be referenced in the eCTD XML backbone. Datasets included within the eCTD should be accurately tagged within a study tagging file to ensure proper identification and organization.[46] The file folder structure for study datasets is summarized in Figure 1. Table 2 provides the study dataset and file folder structure and associated description.

---

[44] See http://www.fda.gov/ForIndustry/DataStandards/StructuredProductLabeling/ucm163377.htm.
[45] See http://www.ich.org/products/ctd.html.
[46] See "The eCTD Backbone File Specification for Study Tagging Files" (http://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/FormsSubmissionRequirements/ElectronicSubmissions/UCM163560.pdf) for further details.

Figure 1: Folder Structure for Study Datasets

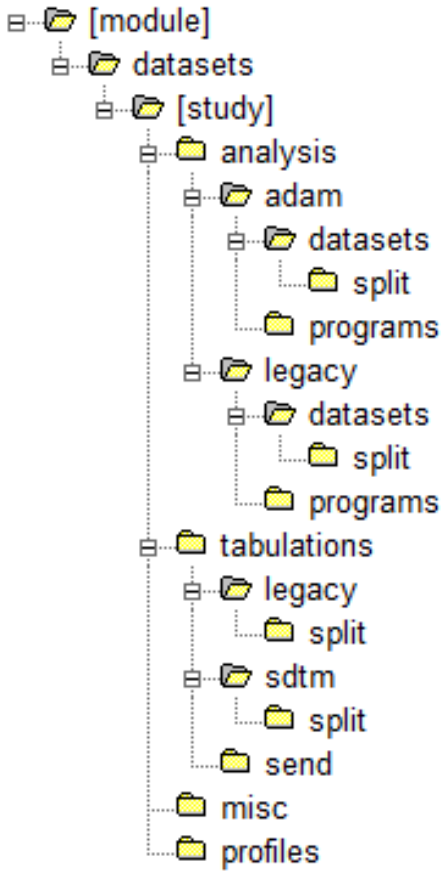

Table 2: Study Dataset and File Folder Structure and Description

Folder Name	Folder Level	Description/Contents
[module]	1	Refers to the eCTD module in which study data is being submitted. Name this folder m4 for nonclinical data and m5 for clinical data. Do not place files at this level.
datasets	2	Resides within the module folder as the top-level folder for study data (nonclinical or clinical) being submitted for the specified module (m4 or m5). Do not place files at this level.
[study]	3	Name this folder with the study identifier or analysis type performed (e.g., study123, iss, ise). Do not place files at this level.
analysis	4	Contains folders for analysis datasets and software programs; arrange in designated level 6 subfolders. Do not place files at this level.
adam	5	Contains subfolders for ADaM datasets and corresponding software programs. Do not place files at this level.
datasets	6	Place ADaM datasets in this subfolder.
split	7	Place any split ADaM datasets in this subfolder.
programs	6	Place software programs for ADaM datasets, tables and figures in this subfolder.
legacy	5	Contains legacy formatted analysis datasets and corresponding software programs. Do not place files at this level.
datasets	6	Place legacy analysis datasets in this subfolder.
split	7	Place split legacy analysis datasets in this subfolder.
programs	6	Place software programs for legacy analysis datasets, tables and figures in this subfolder.
misc	4	Place miscellaneous datasets that don't qualify as analysis, profile, or tabulation datasets in this subfolder. This subfolder was formerly named "listings".
profiles	4	Place patient profiles in this subfolder.
tabulations	4	Contains subfolders for tabulation datasets. Do not place files at this level.
legacy	5	Place legacy (non-standardized) tabulation datasets in this folder.
split	6	Place any split legacy tabulations datasets in this subfolder.
sdtm	5	Place SDTM tabulation datasets in this subfolder. Should only be used in m5 for clinical data.
split	6	Place any split SDTM files in this subfolder.
send	5	Place SEND tabulation datasets in this subfolder. Should only be used in m4 for animal data.

8. Data Validation and Traceability

8.1 Definition of Data Validation

For purposes of this Guide, data validation is a process that attempts to ensure that submitted data are both compliant and useful. *Compliant* means the data conform to the applicable and required data standards. *Useful* means that the data support the intended use (i.e., regulatory review and analysis).

Data validation is one method used to assess submission data quality. Standardized data do not ensure quality data, but they do make it easier to assess some aspects of data quality by facilitating the automation of various data checks.

Data validation relies on a set of validation rules that are used to verify that the data conform to a minimum set of quality standards. The data validation process can identify data issues early in the review that may adversely affect the use of the data. FDA recognizes that it is impossible or impractical to define *a priori* all the relevant validation rules for any given submission. Sometimes serious issues in the submitted data are only evident through manual inspection of the data and may only become evident once the review is well under way. Often these issues are due to problems in data content (i.e., *what* was or was not submitted, or issues with the collection of original source data), and not necessarily *how* the data were standardized.

8.2 Study Data Validation Rules

8.2.1 Types of Data Validation Rules

Generally, FDA recognizes two types of validation rules:

Conformance validation: These rules help ensure that the data conform to the data standards. For example, a conformance validation rule for CDISC SDTM data would check that the value in the Domain column of all datasets matches the name of the domain.

Quality checks: These checks help to ensure the data will support meaningful analysis. For example, a quality check for a particular human study may require that each value for AGE fall within a pre-specified human physiologic range.

Once a data standard is defined, the conformance validation rules are generally static. They are not expected to change substantially unless the standard itself changes. However, new analysis requirements or specific studies may suggest additional quality checks and these will be incorporated into data validation processes.

8.2.2 Support on Data Validation Rules

The Standards Web page[47] provides links to the currently available validation rules, i.e. both conformance rules and quality checks.

Sponsors should validate their study data before submission using the published validation rules and either correct any validation errors or explain in the *SDRG* why certain validation errors could not be corrected. The recommended pre-submission validation step is intended to minimize the presence of validation errors at the time of submission.

Data validation is used by FDA to inform review staff of potential problems in using the data, and to assess the usefulness of the rules. If applicable, FDA may report important data validation errors to the sponsor for correction.

8.3 Study Data Traceability

8.3.1 Overview

An important component of a regulatory review is an understanding of the provenance of the data (i.e., traceability of the sponsor's results back to the CRF data). Traceability permits an understanding of the relationships between the analysis results, analysis datasets, tabulation datasets, and source data. Traceability enables the reviewer to accomplish the following:

- Understand the construction of analysis datasets
- Determine the observations and algorithm(s) used to derive variables
- Understand how the confidence interval or the p-value was calculated in a particular analysis

Based upon reviewer experience, establishing traceability is one of the most problematic issues associated with legacy study data converted to standardized data. If the reviewer is unable to trace study data from the data collection of subjects participating in a study to the analysis of the overall study data, then the regulatory review of a submission may be compromised. Traceability can be assured when studies are prospectively designed to collect data using a standardized CRF, e.g., CDASH.

As noted in Section 1.1, the submission of standardized study data will be required according to the timetable specified in the eStudy Data guidance. During the transition period to required study data standards, FDA recognizes that some study data (i.e., legacy data) submissions may not conform to FDA-supported study data standards and may need to be converted.

[47] The Standards Web page can be accessed at http://www.fda.gov/forindustry/datastandards/studydatastandards/default.htm.

8.3.2 Legacy Study Data Conversion to Standardized Study Data

Sponsors should use processes for legacy data conversion that account for traceability. Generally, a conversion to a standard format will map every data element as originally collected to a corresponding data element described in a standard. Some study data conversions will be straightforward and will result in all data converted to a standardized format. In some instances, it may not be possible to represent a collected data element as a standardized data element. In these cases, there should be an explanation in the *SDRG* as to why certain data elements could not be fully standardized or were otherwise not included in the standardized data submission. The legacy data (i.e., aCRF, legacy tabulation data, and legacy analysis data) may be needed in addition to the submission of converted data.

In cases where the data were collected on a Case Report Form (CRF) or electronic CRF but were not included in the converted datasets, the omitted data should be apparent on the annotated CRF and described in the *SDRG*. The tabular list of studies in the *Standardization Plan* should indicate which studies contained previously collected non-standard data that were subsequently converted to a standard format.

8.3.2.1 Traceability Issues with Legacy Data Conversion

FDA does not recommend a particular approach to legacy study data conversion, but rather explains the issues that should be addressed so that the converted data are traceable and adequate to support review.

Table 3 presents some of the issues that can be observed during a review when legacy study data are converted to SDTM and submitted with legacy analysis datasets.

Table 3: Traceability Issues: Legacy Data Conversion to SDTM Only

1. Limited ability to determine location of collected CRF variables in the converted SDTM data unless the legacy aCRF is re-annotated.
2. Limited traceable path from SDTM to the legacy analysis data.
3. Limited ability to replicate/confirm legacy analysis datasets (i.e., analysis variable imputation or derived variables) using SDTM datasets.
4. Limited ability to confirm derivation of intermediate analysis datasets or custom domains.
5. Difficulty in understanding the source or derivation methods for imputed or derived variables in integrated/pooled data, supplemental qualifiers, and related records.

Table 4 presents the issues when legacy study data and legacy analysis data are independently converted to SDTM and ADaM formats, respectively, rather than ADaM datasets being created directly from the SDTM datasets (converted from legacy study data).

Table 4: Traceability Issues: Independent Legacy Data Conversion to SDTM and ADaM

Issues
1. Limited ability to determine location of collected CRF variables in the converted SDTM data unless the legacy aCRF is re-annotated.
2. Limited traceable path from SDTM to the legacy analysis data.
3. Limited ability to replicate/confirm legacy analysis datasets (i.e., analysis variable imputation or derived variables) using SDTM datasets.
4. Limited ability to confirm derivation of intermediate analysis datasets or custom domains.
5. Limited traceable path from SDTM to the ADaM datasets.
6. Limited ability to replicate ADaM datasets (i.e., analysis variable imputation or derived variables) using SDTM datasets.
7. Limited traceable path from ADaM to the Tables, Figures and the Clinical Study Report (CSR).
8. Difficulty in understanding the source or derivation methods for imputed or derived variables in integrated/pooled data, supplemental qualifiers, and related records.

Table 5 presents the issues when legacy data are converted to SDTM and ADaM formats in sequence (i.e., converting legacy study data to SDTM and then creating ADaM from the SDTM). The key concern is the traceability from ADaM to the Tables, Figures and CSR.

Table 5: Traceability Issues: Legacy Data Conversion to SDTM and ADaM in Sequence

1. Limited ability to determine location of collected CRF variables in the converted SDTM data unless the legacy aCRF is re-annotated.
2. Limited traceable path from SDTM to the legacy analysis data.
3. Limited ability to replicate/confirm legacy analysis datasets (i.e., analysis variable imputation or derived variables) using SDTM datasets.
4. Limited ability to confirm derivation of intermediate analysis datasets or custom domains.
5. Limited traceable path from ADaM to the Tables, Figures and the CSR.
6. Difficulty in understanding the source or derivation methods for imputed or derived variables in integrated/pooled data, supplemental qualifiers, and related records.

8.3.2.2 Legacy Data Conversion Plan and Report

Sponsors should evaluate the decision involved in converting previously collected non-standardized data (i.e., legacy study data) to standardized data (i.e., SDTM, SEND, and ADaM). Sponsors should provide the explanation and rationale for the study data conversion in the *SDRG*. To mitigate traceability issues when converting legacy data, FDA recommends the following procedures:

1. Prepare and Submit a Legacy Data Conversion Plan and Report.
   - The plan should describe the legacy data and the process intended for the conversion.
   - The report should present the results of the conversions, issues encountered and resolved, and outstanding issues.
   - The plan and report should be provided in the *SDRG*.

2. Provide an aCRF, for clinical data, that maps the legacy data elements.
   - Sponsors should provide two separate CRF annotations, one based on the original legacy data, and the other based on the converted data (i.e., SDTM) when legacy datasets are submitted. The legacy CRF tabulation data should include all versions and all forms used in the study.

3. Record significant data issues, clarifications, explanations of traceability, and adjudications in the *SDRG*. For example, data were not collected or were collected using different/incompatible terminologies, or were collected but will not fit into, for example, SDTM format.

4. Legacy data (i.e., legacy aCRF, legacy tabulation data, and legacy analysis data) may be needed in addition to the converted data.

Appendix: Data Standards and Interoperable Data Exchange

This appendix provides some of the guiding principles for the Agency's long-term study data standards management strategies. An important goal of standardizing study data submissions is to achieve an acceptable degree of *semantic interoperability* (discussed below). This appendix describes different types of interoperability and how data standards can support interoperable data exchange now and in the future.

At the most fundamental level, study data can be considered a collection of data elements and their relationships. A data element is the smallest (or *atomic*) piece of information that is useful for analysis (e.g., a systolic blood pressure measurement, a lab test result, a response to a question on a questionnaire).

A data value is by itself meaningless without additional information about the data (so called *metadata*). Metadata is often described as *data about data*. Metadata is structured information that describes, explains, or otherwise makes it easier to retrieve, use, or manage data.[48] For example, the number *44* itself is meaningless without an association with Hematocrit and the unit of measurement (e.g. "%"). Hematocrit in this example is metadata that further describes the data.

Just as it is important to standardize the representation of data (e.g., M and F for male and female, respectively), it is equally important to standardize the metadata. The expressions Hematocrit = 44; Hct = 44, or Hct Lab Test = 44 all convey the same information to a human, but an information system or analysis program will fail to recognize that they are equivalent because the metadata is not standardized. It is also important to standardize the definition of the metadata, so that the meaning of a Hematocrit value is constant across studies and submissions.

In addition to standardizing the data and metadata, it is important to capture and represent relationships (also called associations) between data elements in a standard way. Relationships between data elements are critical to understand or interpret the data. Consider the following information collected on the same day for one subject in a study:

Systolic Blood Pressure = 90 mmHg
Position = standing
Systolic Blood Pressure = 110 mmHg
Time = 10:23 a.m.
Time = 10:20 a.m.

[48] Metadata is said to "give meaning to data" or to put data "in context." Although the term is now frequently used to refer to XML (extensible markup language) tags, there is nothing new about the concept of metadata. Data about a library book such as author, type of book, and the Library of Congress number, are metadata and were once maintained on index cards. SAS labels and formats are a rudimentary form of metadata, although they have not historically been referred to as metadata.

Position = lying

When presented as a series of unrelated data elements, they cannot reliably be interpreted. Once the relationships are captured, as shown below using arrows, the interpretation of a drop in systolic blood pressure of 20 mmHg while standing, and therefore the presence of clinical orthostatic hypotension, is possible. Standardizing study data therefore involves standardizing the data, metadata, and the representation of relationships.

Time = 10:20 a.m. ←→ Position = lying ←→ Systolic Blood Pressure = 110 mmHg
Time = 10:23 a.m. ←→ Position = standing ←→ Systolic Blood Pressure = 90 mmHg

With these fundamental concepts of data standardization in mind, data standards can be considered in the context of interoperable data exchange.

Interoperability

Much has been written about interoperability, with many available definitions and interpretations within the health care informatics community. In August 2006, the President signed an Executive Order mandating that the Federal Government use interoperable data standards for health information exchange.[49] Although this order was directed at Federal agencies that administer health care programs (and therefore not the FDA), it is relevant to this guidance because it defined interoperability for use by Federal agencies:

"Interoperability" means the ability to communicate and exchange data accurately, effectively, securely, and consistently with different information technology systems, software applications, and networks in various settings, and exchange data such that clinical or operational purpose and meaning of the data are preserved and unaltered.

Achieving interoperable study data exchange between sponsors and applicants and FDA is not an all-or-nothing proposition. Interoperability represents a continuum, with higher degrees of data standardization resulting in greater interoperability, which in turn makes the data more useful and increasingly capable of supporting efficient processes and analyses by the data recipient. It is therefore useful to understand the degree of interoperability that is desirable for standardized study data submissions.

In 2007, the Electronic Health Record Interoperability Work Group within Health Level Seven issued a white paper that characterized the different types of interoperability based on an analysis of how the term was being defined and used in actual practice.[50] Three types of interoperability were identified: technical, semantic, and process interoperability. A review of these three types provides insight into the desired level of interoperability for standardized study data submissions.

---

[49] See http://www.cga.ct.gov/2006/rpt/2006-R-0603.htm.

[50] See Coming to Terms: Scoping Interoperability for Health Care http://www.hln.com/assets/pdf/Coming-to-Terms-February-2007.pdf.

Technical interoperability describes the lowest level of interoperability whereby two different systems or organizations exchange data so that the data are useful. The focus of technical interoperability is on the conveyance of data, not on its meaning. Technical interoperability supports the exchange of information that can be used by a person but not necessarily processed further. When applied to study data, a simple exchange of nonstandardized data using an agreed-upon file format for data exchange (e.g., SAS transport file) is an example of technical interoperability.

Semantic interoperability describes the ability of information shared by systems to be understood, so that nonnumeric data can be processed by the receiving system. Semantic interoperability is a multi-level concept with the degree of semantic interoperability dependent on the level of agreement on data content terminology and other factors. With greater degrees of semantic interoperability, less human manual processing is required, thereby decreasing errors and inefficiencies in data analysis. The use of controlled terminologies and consistently defined metadata support semantic interoperability.

Process interoperability is an emerging concept that has been identified as a requirement for successful system implementation into actual work settings. Simply put, it involves the ability of systems to exchange data with sufficient meaning that the receiving system can automatically provide the right data at the right point in a business process.

An example of process interoperability in a regulatory setting is the ability to quickly and automatically identify and provide all the necessary information to produce an expedited adverse event report in a clinical trial upon the occurrence of a serious and unexpected adverse event. The timely submission of this information is required by regulation to support FDA's mandate to safeguard patient safety during a clinical trial. Process interoperability becomes important when particular data are necessary to support time-dependent processes.

Because the vast majority of study data are submitted after the study is complete, achieving process interoperability for study data submissions in a regulatory setting is relatively unimportant, at least for the foreseeable future. It is reasonable to conclude that it is most desirable to achieve *semantic interoperability* in standardized study data submissions.

In summary, the goal of standardizing study data is to make the data more useful and to support semantically interoperable data exchange between sponsors, applicants, and the FDA such that it is commonly understood by all parties.

Glossary

The following list of acronyms and terms used in this Guide:

aCRF:	Annotated Case Report Form
ANDA:	Abbreviated New Drug Application
ADaM:	Analysis Data Model
ADSL:	Analysis Data Subject Level
ASCII:	American Standard Code for Information Interchange
CBER:	Center for Biologics Evaluation and Research
CDASH:	Clinical Data Acquisition Standards Harmonization
CDER:	Center for Drug Evaluation and Research
CDISC:	Clinical Data Interchange Standards Consortium
CS:	Chemical Structure
Domain:	A collection of observations with a topic-specific commonality
eCTD:	Electronic Common Technical Document
ICH:	International Conference on Harmonisation
IND:	Investigational New Drug
ISE:	Integrated Summary of Efficacy
ISO:	International Organization for Standardization
ISO 8601:	ISO character representation of dates, date/times, intervals, and durations of time.
ISS:	Integrated Summary of Safety
ITT:	Intent-To-Treat
MedDRA:	Medical Dictionary for Regulatory Activities
MOA:	Mechanism of Action
NDA:	New Drug Application
NDF-RT:	National Drug File – Reference Terminology
PDF:	Portable Document Format
PE:	Physiologic Effect
SDTM:	Study Data Tabulation Model
SNOMED:	Systematized Nomenclature of Medicine
UNII:	FDA Unique Ingredient Identifier
WHO:	World Health Organization
XML:	eXtensible Markup Language
XPORT:	SAS Transport Version 5