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)**

February 2014

STUDY DATA
TECHNICAL CONFORMANCE GUIDE

Revision History

Date	Version	Summary of Revisions
February 2014	1.0	Initial Version

Table of Contents

STUDY DATA TECHNICAL CONFORMANCE GUIDE

This technical specifications document, when finalized, will represent 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 Guidance). The eStudy Data guidance, when finalized, 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).

1.2. Purpose

This Guide provides technical recommendations to sponsors[3] 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.

Because of the inherent variability across studies and applications, it is difficult to identify all data needed by a review division for a scientific regulatory review. Therefore, prior to submission, sponsors should discuss with the review division the data necessary to support a submission, the data elements that should be included in each dataset, and the organization of

[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] For the purposes of this document, the term "sponsor" refers to both "sponsors" and "applicants" who are submitting study data to the Agency.

the data within the datasets. 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.

1.3. Document Revision and Control

FDA intends to issue a *Federal Register* notice announcing revisions to and seeking public comment on this Guide as necessary and will post the revised Guide on the **Study Data Standards Resources Web page** (Standards Web page).[4] The revision history page of this document will contain sufficient information to indicate which sections of the Guide have been revised.

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 and a study data reviewer's guide.

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 Non-Clinical** – presents general considerations and specifications for sponsors using, for example, the following standards for the submission of study data: Clinical Data Interchange Standards Consortium (CDISC), Study Data Tabulation Model (SDTM), Analysis Data Model (ADaM), and Standard for Exchange of Nonclinical Data (SEND).

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

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

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

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

Section 8: **Data Fitness** – provides general recommendations on standards compliance, data traceability expectations, legacy data conversion, versioning, and data validation rules.

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.[5] The examples of issues and concerns discussed in the Guide are intended as examples only of common issues, not an inclusive list of all possible issues.

This Guide, when finalized, supersedes all previous Study Data Specifications documents (Versions 1.0 - 2.0) and CDER Study Data Common Issues Documents (Versions 1.0 -1.1). This Guide should be considered a companion document to the following:

- Guidance to Industry *Providing Regulatory Submissions in Electronic Format – Standardized Study Data*
- FDA Study Data Standards Resources Web page
- FDA Data Standards Catalog

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
5. List of and justification for studies that may not conform to the standards

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

The Standardization Plan should be updated in subsequent communications with FDA as the development program expands and additional studies are planned. The cover letter accompanying a study data submission should describe the extent to which the Standardization Plan was executed.

2.2. Study Data Reviewer's Guide

Some data standards may not require the use of all data elements defined by the standard to be collected in any given study. For example, the Study Data Tabulation Model (SDTM)[6] classifies variables as required, expected, or permissible. *What* data are collected and submitted is the subject of science, regulation, and discussions with the review division. However, all study-specific data necessary to evaluate the safety and efficacy of the product should be submitted with the highest level of standardization possible.

When using a data standard, there may be occasional ambiguity resulting in more than one way to implement the standard.[7] Instances in which a standard allows for more than one implementation should be discussed with the appropriate review division or the data resource team (CBER and CDER products)[8] before data submission. Sponsors and applicants should ensure their data conform to the required standard. Sponsors and applicants should describe their use of study data standards and their conformance validation in a *Reviewer's Data Guide (Data Guide).*[9]

The *Data Guide* 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. The *Data Guide* is recommended as an integral part of a standards-compliant study data submission. The *Data Guide* should be placed in the electronic Common Technical Document (eCTD) in Module 5.

For each study, the *Data Guide* should include, but is not limited to the following:

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 conformance validation rules, versions, and issues

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

[7] For example, the CDISC SDTM Implementation Guide, v. 3.1.2, describes the reference start date (RFSTDTC) as *usually* equivalent to date/time when subject was first exposed to study treatment. The word *usually* indicates other interpretations of the reference start date are possible.

[8] Data Resource Team: cder-eDataTeam@fda.hhs.gov or cber.cdisc@fda.hhs.gov.

[9] A specific template for a Reviewer's Data Guide is not specified: however, an example of a Reviewer's Data Guide can be found at http://www.phusewiki.org/wiki/index.php?title=Study_Data_Reviewer's_Guide.

3. Exchange Format – Electronic Submissions

3.1. Extensible Mark-up Language (XML)

3.1.1. Use

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.[10,11] Its primary purpose is to facilitate the sharing of structured data across different information systems.

One XML use case is CDISC's define.xml file. Define.xml is used to describe the structure and contents of the data collected from clinical and non-clinical studies. The specification for the data definition for datasets provided using SDTM, SEND and ADaM formats is the define.xml file.

3.1.2. Version

XML is defined in the version specification produced by the W3C, as well as several other related specifications; all are free open standards.[12] See the Standards Catalog for the version(s) that are supported by FDA.

3.1.3. File Extension

All XML files should use .xml as the file extension.

3.1.4. Compression

XML files should not be compressed.

3.1.5. File Size

Please see section 3.3.5.1 below.

3.2. Portable Document Format (PDF)

3.2.1. Use

PDF is an open file format used to represent documents in a manner independent of application software, hardware, and operating systems.[13]

PDF use cases, for example, include the submission of the CDISC define.pdf file, the annotated CRF (aCRF / blankcrf), and other documents that align with the International

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

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

[12] Ibid.

[13] Adobe Systems Incorporated, PDF Reference, Sixth edition, version 1, Nov. 2006, p. 33.

Conference on Harmonization (ICH) M2.[14] Detailed FDA PDF specifications are located on FDA's Electronic Common Technical Document (eCTD) Web site.[15]

3.2.2. Version

The Standards Catalog lists the PDF version(s) that are supported by FDA.

3.2.3. File Extension

All PDF files should use .pdf as the file extension.

3.2.4. Compression

PDF files should not be compressed.

3.2.5. File Size

Please see section 3.3.6.1 below.

3.3. SAS© Transport Format (XPORT)

3.3.1. Use

The SAS XPORT file transport format, Version 5, is the file format for the submission of all electronic datasets. The SAS XPORT file transport format, Version 5, is an open file format published by SAS Institute for the exchange of study data.[16] 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.

3.3.2. Version

XPORT transport files can be created by PROC XCOPY in Version 5 of SAS software and by the SAS PROC in Version 6 and higher of the SAS Software. SAS Transport files processed by the CPORT SAS PROC cannot be reviewed, processed, or archived by FDA.

Sponsors can find the record layout for SAS XPORT transport files through SAS technical support technical document TS-140.[17]

3.3.3. File Extension

All SAS XPORT transport files should use .xpt as the file extension.

3.3.4. Compression

SAS transport files should not be compressed. There should be one dataset per transport file.

[14] See http://www.ich.org/products/electronic-standards.html.

[15] Available at http://www.fda.gov/Drugs/DevelopmentApprovalProcess/FormsSubmissionRequirements/ElectronicSubmissions/ucm153574.htm

[16] See http://www.sas.com

[17] Available at http://support.sas.com/techsup/technote/ts140.html.

3.3.5. File Size

3.3.5.1 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 in size. Datasets should be resized to the maximum length used for each character variable prior to splitting. Datasets divided to meet the maximum size restrictions should contain the same variable presentation so they can be easily combined. Split datasets should have matching column widths. This will ensure that the split datasets have matching variable lengths for merging data. Split data should be noted in the define.xml (see section 4.1.9.1) and the *Data Guide*, clearly identifying the method used for the dataset splitting.

Datasets that are split should be clearly named to aid the reviewer in reconstructing the original dataset (e.g., xxx1, xxx2, xxx3). For more detailed instructions on splitting of Laboratory (LB) datasets, please refer to section 4.1.6.3 of this Guide.

3.3.5.2 Column Length

For all datasets, the allotted character length for each column should be set to the maximum length of the variable used. This will significantly reduce file sizes. For example, if USUBJID (see section 4.1.6.2) has a maximum length of 18, the USUBJID's column size should be set to 18, not 200.

3.3.5.3 Variable Length

The length for variable names and descriptive labels and dataset labels should not exceed the maximum length of variables.

Table 1: Maximum Length of Variables

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

3.3.5.4 Special Characters: Variables and Datasets

Variable names, as well as variable and dataset labels should include American Standard Code for Information Interchange (ASCII) codes only. Names and labels should not contain punctuation, dashes, spaces, or other non-alphanumeric symbols. In addition, the variable names should not contain special characters, including:

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

4. Study Data Submission Format – Clinical and Non-Clinical

4.1. Clinical Data Interchange Standards Consortium (CDISC)

CDISC is an open, multidisciplinary, neutral, nonprofit standards developing organization (SDO) that has been working through consensus-based collaborative teams, since its formation in 1997, to develop global data standards for clinical and nonclinical research.[18]

4.1.1. General Considerations

Data format specifications for the tabulation datasets of clinical and non-clinical toxicology studies are provided by SDTM and SEND, respectively. ADaM provides the data format specifications for analysis datasets. As noted above, 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 and SEND with analysis datasets as described below. Currently, ADaM specifications for SEND have not been developed.

When using an FDA-supported data standard, there may be occasional ambiguity resulting in more than one way to implement the standard. 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 these instances, sponsors should discuss their proposed solution with the review division and submit supporting documentation that describes these decisions or solutions in the Data Guide at the time of submission.

4.1.2. 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, all Required, Permissible and Expected variables (for which data were collected or for which derivations are possible) should be submitted. The following are examples of some of the Permissible and Expected variables in SDTM and SEND that should be included:

1. Baseline flags for Laboratory results, Vital Signs, ECG, Pharmacokinetic Concentrations, and Microbiology results.

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

2. EPOCH designators. An extensible code list for EPOCH will be developed. Please follow CDISC guidance for terminology[19].

3. Whenever --DTC, --STDTC or --ENDTC 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 Elements domain, SESTDTC is Required and SEENDTC is Expected; therefore, both SESTDY and SEENDY should be included.

   The variable EPOCH should be included for every clinical 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 the actual intervention the subject experienced during that phase.

4. Study Day variables (--DY, --STDY, and--ENDY).

5. Cause of death or comments in pathology reports.

4.1.3. Dates

Dates in SDTM and SEND domains should conform to the ISO 8601 format. Examples of how to implement dates are included in the SDTM and SEND Implementation Guides.[20]

4.1.4. Naming Conventions

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

4.1.5. SDTM / SEND / ADaM Versions

For each study, sponsors should submit datasets using the same SDTM / SEND / ADaM version. As noted above, the Standards Catalog lists the versions that are supported by FDA.

When integrating (or pooling) data across multiple studies (e.g., Integrated Summaries of Safety (ISS) and Integrated Summaries Efficacy (ISE)), the integrated datasets should be submitted in one standardized FDA-supported version. The sponsor should evaluate the benefits and risks associated with down- or up-versioning of two or more studies with differing original versions (e.g., SDTM versions). Conversions to one standardized version should be described in the Data Guide, including the rationale for the conversion.

[19] See http://www.cancer.gov/cancertopics/terminologyresources/page6.

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

4.1.6. Study Data Tabulation Model (SDTM)

4.1.6.1. Definition

SDTM defines a standard structure for human clinical study data tabulations.

4.1.6.2. 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 (such as with Clinical Data Acquisition Standards Harmonization (CDASH)) allows for a simplified process for creation of SDTM domains.

The SDTM Implementation Guide (SDTMIG) should be followed carefully unless otherwise indicated in this Guide or in the Standards Catalog. However, the conformance criteria listed in the SDTMIG should not be 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. SDTM datasets should not contain imputed data.

<u>USUBJID</u>

Each individual subject should be assigned a single unique identifier across the <u>entire</u> application. An individual subject should have the <u>exact</u> 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 did 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.

<u>SUBJID</u>

The assigned SUBJID should be the exact same identifier that is used to identify subjects in the accompanying study report.

4.1.6.3. SDTM Domain Specifications

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 Data Guide.

SUPPQUAL

SUPPQUAL represents a series of datasets in SDTM submissions. It is intended to include data variables that are not specified in the SDTM. SUPPQUAL datasets are used for data elements that cannot be allocated to other SDTM domains. Discussion with the review division should occur if the sponsor intends to include important variables (e.g., that support key analyses) in SUPPQUAL datasets. Alternative solutions may include the following: 1. use of a custom domain (see below) or 2. use of both SUPPQUAL and ADaM datasets to include the important variables that support key analyses.

DM Domain (Demography)

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.

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 associate subject data (assessments, events, and interventions) with the study element in which they occurred.

AE Domain (Adverse Events)

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.

When the serious adverse event variable, AESER, is populated with a “Y” the criteria for the assessment of an adverse event as “serious” should be indicated, e.g., death, hospitalization, disability/permanent damage. Frequently, sponsors omit this information, even when it has been collected on the CRF. The criteria that led to the determination should be 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. Sponsors should also check the Standards Catalog to determine FDA support for newly published standards. If it is necessary to create custom domains, sponsors should follow the recommendations in the STDMIG. In addition, sponsors should discuss their proposed implementation approach with the review division prior to submission.

LB Domain (Laboratory)

The size of the LB domain dataset submitted by sponsors is often too large to process (see section 3.3.5.1). 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.

TD Domain (Trial Design)

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

4.1.7. Analysis Data Model (ADaM)

4.1.7.1. Definition

Specifications for analysis datasets for human drug product clinical and analytical studies are provided by the ADaM. Analysis datasets are created and used to support the results in clinical study reports, Integrated Summaries of Safety, and Integrated Summaries of Efficacy, as well as other analyses required for a thorough regulatory review. Analysis datasets can contain imputed data or data derived from tabulation datasets (e.g., SDTM). For standardized analysis datasets, sponsors should refer to the published ADaM Implementation Guide (ADaMIG). Sponsors should have a significant discussion with reviewers to appropriately determine which analysis datasets and associated content should be submitted to support application review.

4.1.7.2. General Considerations

The Standards Catalog provides a listing of all the currently supported data standards with links to reference materials. Although ADaM generally facilitates FDA review, it does not always provide data structured in a way that supports all of the analyses that should be submitted for review. Therefore, sponsors should 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. 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. Sponsors who provide the software programs used to create ADaM datasets help reviewers to better understand how the datasets were created (see section 4.1.7.8). Each analysis dataset that is shown in the define.xml file should be described.

4.1.7.3. 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 data. Further, primary and secondary variables and their derivations (as applicable) should be provided, as well as documented in the define file and Data Guide.

In addition, 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 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.

4.1.7.4. Timing Variables

When an analysis dataset contains multiple records per subject (i.e., repeated measures data), a variable for relative day of measurement or event, along with timing variables for visit, should be included. 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.” The “time on study” 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 calendar time on study, and so it is not sufficient for plotting data where the x-axis measures real time.

4.1.7.5. 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 to avoid the merging of datasets to perform analyses. Core variables include study/protocol, center/site, region, country, treatment assignment, sex, age, race, analysis population flags (e.g., Intent to Treat (ITT), safety), and other important baseline demographic variables. All variables that contain coded data should be accompanied by a variable that provides the decoded information.

4.1.7.6. Dates

Dates should be formatted as numeric in the analysis datasets.

4.1.7.7. 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.”

4.1.7.8. Software Programs

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 programs created by the analysis tool use a file extension other than .txt, the file name should include the native file extension generated by the analysis tool for the ASCII text program files (e.g., adsl_r.txt or adsl_sas.txt).

4.1.7.9. ADaM Domain Specifications

4.1.7.9.1. Analysis Data Subject Level (ADSL)

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, the sponsor should include multiple additional variables representing various important baseline subject characteristics / covariates presented in the study protocol. Some examples include, but are not limited to, disease severity scores such as Acute Physiology and Chronic Health Evaluation (APACHE) scores,[21] baseline organ function measurements such as calculated creatinine clearance or Forced Expiratory Volume in 1 second (FEV1); range categories for continuous variables; and numeric date variables in non-International Standards Organization (ISO) formats, such as SAS or Oracle.

4.1.7.9.2. Imputed Data

When data imputation is utilized, sponsors should submit imputed data in an analysis dataset, and the relevant supporting documentation (e.g., Study Data Reviewer's Guide, define.xml) explaining the imputation methods. SDTM datasets should not contain imputed data.

4.1.8. Standard for Exchange of Nonclinical Data (SEND)

4.1.8.1. Definition

The SEND provides the organization, structure, and format of standard non-clinical (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.

4.1.8.2. 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. The SENDIG should be followed carefully. The SENDIG is based upon and should be used in close concert with the CDISC SDTM. The SDTM describes the general conceptual model for representing study data for electronic data interchange and should be read prior to reading the SENDIG. If there is uncertainty regarding implementation, the sponsor should discuss the issue with the review division.

The ideal time to implement SEND standards for representation of nonclinical (animal toxicology studies) tabulation data is prior to the conduct of the study. It is very important

[21] 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.

that the results presented in the accompanying study report be traceable back to the original data as represented in the SEND dataset.

4.1.8.3. SEND Domain Specification

SUPPQUAL Datasets

SUPPQUAL represents a series of datasets in SEND submissions. SUPPQUAL is intended to include data variables that are not specified in SEND. SUPPQUAL datasets are often used for data elements that the sponsor is not sure how to allocate. Discussion should occur if the sponsor intends to include important variables, e.g., that support key analyses, in the SUPPQUAL datasets.

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.

Custom Domains

The SEND Implementation Guide allows 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 Web site 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.

4.1.9. Dataset Documentation

4.1.9.1. Define File

The data definition file, define.xml, describes the format and content 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 properly. 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.

In addition, 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 (SDTM versions greater than 3.1.2).

4.1.9.2. Specification

The data definition specification for submitted datasets should be included in the define file XML format.[22] 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).”

4.1.10. Annotated Case Report Form (aCRF)

4.1.10.1. Definition

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.

4.1.10.2. Specifications

The aCRF should include treatment assignment forms 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. The sponsor should write “not entered in database” or “not entered in the tabulation dataset” for all items where this applies. The aCRF should be provided as a PDF file named “aCRF.pdf.”

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 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 /

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

lower case, or the use of hyphens. 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 documented in the define.xml file and the *Data Guide*.

6.1.1. Use of Controlled Terminologies

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. The use of standard terminology for adverse events is perhaps the earliest example of using data standards for study data. *Myocardial infarction* and *heart attack* are both synonyms for the same clinical concept, and as such should be mapped to the same term in a standard dictionary. This standardization facilitates an efficient analysis of events that are coded to the standard term. FDA expects sponsors to provide, in the electronic study data submission, the actual verbatim terms that were collected on the case report form as well as the coded term, so that review staff can evaluate the standardization process. Further, controlled terminology is particularly useful when applied across studies, facilitating appropriate integrated analyses that are stratified by study and related cross-study analyses (e.g., when greater power is needed to detect important safety signals). Cross-study comparisons and multi-study pooled analyses frequently provide critical information for regulatory decisions, such as statistical results that support effectiveness,[23] and important information on exposure-response relationships[24] and population pharmacokinetics.[25]

6.1.2 Maintenance of Controlled Terminologies

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* for any organization. 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. 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 *Data Guide*, and use them consistently throughout the application.

[23] 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.

[24] 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.

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

If a sponsor identifies an entire information domain[26] for which FDA has not accepted a specific standard terminology, the sponsor may select a standard terminology to use, if one exists. FDA recommends that sponsors include this selection in the *Standardization Plan* or in an update to the existing plan, and reference it in the *Data Guide*. 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.

A frequent question is how to handle multiple versions of a dictionary within a single application (e.g., different versions of MedDRA). FDA recognizes that studies are completed at different times, during which different versions of a dictionary may be the most current. In most cases, FDA expects sponsors and applicants to use the most current version of an FDA-supported dictionary available at the time of coding.[27] 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 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, in the *Standardization Plan*, the impact, if any, of the older version on the study results. 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 which versions are used for individual studies, important pooled analyses of coded terms across multiple studies (e.g., a pooled adverse event analysis of all pivotal trials in an integrated summary of safety) should be conducted using a single version of a dictionary. This is the only way to ensure a truly consistent and coherent comparison of clinical and scientific concepts across multiple studies. Sponsors should clearly reference in the *Data Guide* which terminologies and versions are used for every study within a submission.

6.2. CDISC

6.2.1. Controlled Terminology

6.2.1.1. General Considerations

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.[28] 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 Data Guide.

[26] 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).

[27] If a new version of a dictionary is released in the middle of the coding process for a given study, then either the version that was most current when the coding process began or the newer version may be used.

[28] See http://www.cancer.gov/cancertopics/terminologyresources/page6.

6.2.1.2. Extending Code lists

"Extensible" should not be interpreted to mean that sponsors may substitute their own nonstandard codes in place of existing equivalent standardized codes.

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 and other MedDRA terms.

Generally, the studies included in an application were conducted over many years and may have used a different MedDRA version for each study. However, to avoid the potential for 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 events across studies, including the use of Standardized MedDRA Queries.[29] In addition, sponsors should use the MedDRA-specified mapping of terms. The SDTM variables for the different hierarchy levels should represent MedDRA-specified primary mapped terms.

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. FDA is currently exploring the use of codes for particular variables (e.g., Unique Ingredient Identifier (UNII) codes for concomitant medications, MedDRA codes for adverse events, and NCI EVS concept codes for pathology findings in SEND).

6.4. Medications

6.4.1. FDA Unique Ingredient Identifier

6.4.1.1. General Considerations

The FDA Unique Ingredient Identifier (UNII)[30] 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), active comparators (TSPARM=COMPTRT), and any protocol-specified background treatments (TSPARM=CURTRT).

[29] See http://www.meddra.org/standardised-meddra-queries.

[30] See http://www.fda.gov/ForIndustry/DataStandards/SubstanceRegistrationSystem-UniqueIngredientIdentifierUNII/

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, 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.[31] 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.

6.5. 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)[32] 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).[33] The established pharmacologic class is generally the MOA, PE, a term or CS phrase 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.[34] If the established pharmacologic class is not available for an active moiety, then the sponsor should discuss the 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.

[31] The Substance Registration System can be accessed at http://fdasis.nlm.nih.gov/srs

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

[33] 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.

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

6.6. Indication

6.6.1. SNOMED CT

6.6.2. General Considerations

The International Health Terminology Standards Organization's (IHTSDO) Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT)[35] 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 harmonize with Indication information in Structured Product Labeling (SPL). A reviewer should be able to take the Indication term from product labeling and be able to search for clinical or nonclinical studies of that indication without having to translate.

FDA uses a publicly available subset of SNOMED CT for Indication information (i.e., the Veterans Administration/Kaiser Permanente (VA/KP) Problem List).[36] If an appropriate term cannot be found in this list, then sponsors should select a term from the full SNOMED CT dictionary.

7. General Electronic Submission Format

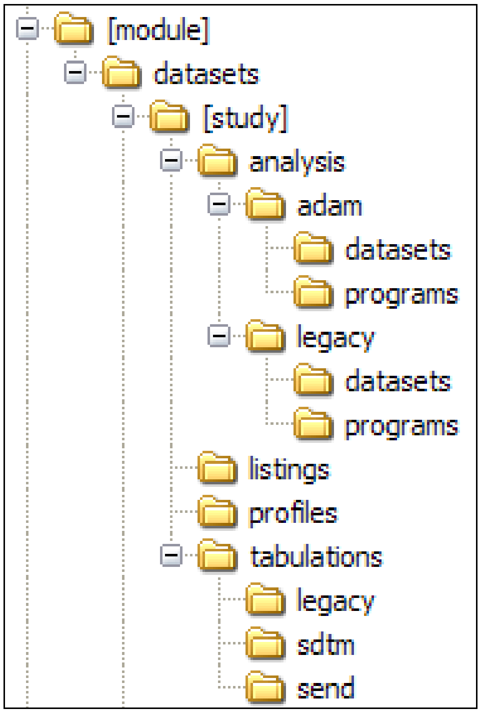
The specifications for organizing study datasets and their associated files within the Electronic Common Technical Document (eCTD) [37] format are summarized in figure 1 and table 2. Unused eCTD folders do not need to be supplied. No additional subfolders are needed; however, if a sponsor has split a file that exceeds file size limits, sponsors should submit the smaller split files in a separate sub-folder/"SPLIT" that is clearly documented in addition to the larger non-split standard LB domain file in the datasets directory. All datasets should be referenced in the eCTD XML backbone.

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

[36] See http://www.fda.gov/ForIndustry/DataStandards/StructuredProductLabeling/ucm163377.htm.

[37] See http://www.ich.org/products/ctd.html.

Figure 1: eCTD Module File Structure



```
[module]
    datasets
        [study]
            analysis
                adam
                    datasets
                    programs
                legacy
                    datasets
                    programs
            listings
            profiles
            tabulations
                legacy
                sdtm
                send
```

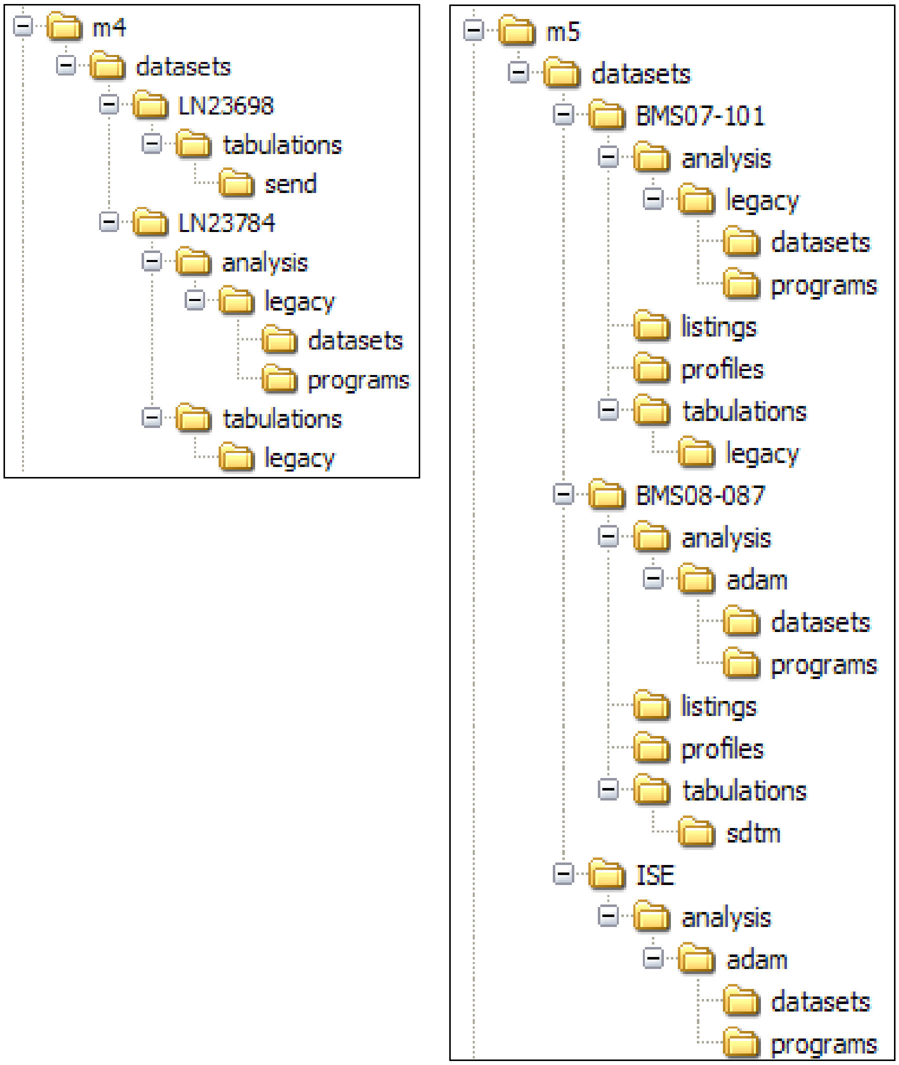
Table 2: eCTD Study Dataset and File Folder Structure

Folder	Description
[module name]	The folder should be named according to which CTD module the datasets apply. M4 is used for nonclinical data and m5 is for clinical data.
datasets	The folder under which all of the study data being submitted for the module specified should be organized.
[study]	The folder should be named according to the study identifier or analysis performed for which the data is being supplied (e.g., BMS09-101, ISS, ISE).
analysis	The folder under which analysis datasets and software programs should be organized **Note:** The analysis datasets and software programs should be placed in specific folders based on their format.
adam	The sub-folder under which ADaM formatted datasets and corresponding software programs should be organized.
datasets	The sub-folder in which the ADaM datasets should be organized.
programs	The sub-folder in which the ADaM dataset software programs should be organized.
legacy	The sub-folder under which legacy formatted datasets and corresponding software programs should be organized.
datasets	The sub-folder in which the legacy analysis datasets should be organized.
programs	The sub-folder in which the legacy analysis dataset software programs should be organized.
listings	The folder in which miscellaneous datasets that don't qualify as analysis, profile, or tabulation datasets should be organized (as needed).
profiles	The folder in which patient profiles should be organized (as needed).
tabulations	The folder under which tabulation datasets should be organized. **Note:** The tabulations datasets should be placed in specific folders based on their format.
legacy	The sub-folder in which tabulation datasets not formatted according to an identified standard format should be organized (e.g., non-SDTM datasets).
sdtm	The sub-folder in which tabulation datasets formatted according to the SDTM IG standard should be organized. Should only be used in m5 for clinical data
send	The sub-folder in which tabulation datasets formatted according to the SEND IG standard should be organized. Should only be used in m4 for animal data.

Figure 2 shows the folder structure for a submission containing two individual nonclinical studies and two clinical studies and an ISE. SEND tabulations have been submitted for study

LN23698. Legacy tabulations and legacy analysis data have been submitted for study LN23784. Legacy clinical study data tabulations and legacy analysis data have been submitted for study BMS07-101. SDTM and ADaM data have been submitted for study BMS08-087. ADaM datasets have been submitted for the ISE.

Figure 2: Example of the eCTD M4 and M5 Folder Structure



```
m4
    datasets
        LN23698
            tabulations
                send
        LN23784
            analysis
                legacy
                    datasets
                    programs
            tabulations
                legacy
```

```
m5
    datasets
        BMS07-101
            analysis
                legacy
                    datasets
                    programs
            listings
            profiles
            tabulations
                legacy
        BMS08-087
            analysis
                adam
                    datasets
                    programs
            listings
            profiles
            tabulations
                sdtm
        ISE
            analysis
                adam
                    datasets
                    programs
```

8. Data Fitness

Data fitness for regulatory review is characterized by the ability to use the study data to effectively, efficiently, and consistently assess the benefits and risks of a medical product. Data that comply with FDA-supported standards and terminologies allow the use of validation processes and tools to ensure that the data can be processed and properly used for review. Data fitness recommendations focus on issues with respect to conformance to supported data standards, use of validation rules to check conformance, and understanding the importance of data traceability in the conduct of a regulatory review.

8.1. Common Data Standards Conformance Errors

To facilitate effective data standards implementations, periodically and as needed, FDA may post to the Standards Web page a list of the most common data standards conformance errors (e.g., define.xml doesn't validate, invalid date format) observed in regulatory submissions.

8.2. Study Data Validation Rules

8.2.1. 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 can 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 quality checks (e.g., completeness, reasonableness). 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, and 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.

8.2.2. Type of Data Validation Rules

FDA generally recognizes two types of validation rules:

> **Conformance validation rules** help ensure that the data conform to the data standards. For example, a conformance validation rule for Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model (SDTM) data would check that the value in the DOMAIN column of all datasets matches the name of the domain.
>
> **Quality checks** that help 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.

8.2.3. Support on Data Validation Rules
The Standards Web page provides links to the validation rules needed to ensure data compliance with CDISC standards, such as SDTM, SEND, ADaM, and define.xml.

Sponsors should validate their study data before submission using the published validation rules and either correct any validation errors or explain in the *Data Guide* 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.

FDA will conduct data validation on the submitted datasets and will use the results to inform review staff of potential problems in using the data, and to assess the usefulness of the rules. If applicable, FDA will report 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, SDTM 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 SDTM data. If the reviewer is unable to trace study data from the data collection of individual subjects participating in a study to the analysis of the overall study data, this may compromise the regulatory review of a submission.

8.3.2. Standardization of Previously Collected Non-Standardized Data
FDA recognizes that for a period of time some study data (i.e., legacy data) may not conform to FDA-supported study data standards. Although FDA recognizes that legacy data are not always easily amenable to retrospective conversion to the FDA-supported standard, the submission of converted data during this period of time may facilitate efficient review of the study data.

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 *Data Guide* as to why certain data elements could not be fully standardized or were otherwise not included in the standardized data submission. Further, sponsors and applicants should consider the submission of legacy data in addition to the 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 *Data Guide*. The tabular list of studies in a submission should indicate which studies contained previously collected non-standard data that were subsequently converted to standard format.

Legacy data conversion to standardized study data processes that do not account for traceability have been associated with major issues and have created difficulty for reviewers conducting the review. FDA realizes that data standards implementation may require time and include changes in processes, technologies, and training. During this transition, the sponsor's submission of data converted to SDTM and ADaM from their legacy format should remain a viable option. FDA encourages sponsors to prospectively design their studies to collect data in standardized formats using CDASH.

8.3.3. Common Issues Involving Traceability

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 is adequate to support review.

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

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

1.	No traceable path from legacy analysis data to SDTM.
2.	No ability to confirm analysis variable imputation or derived variables.
3.	Unable to replicate tables, listings and and figures (TLFs) and legacy analysis datasets using SDTM datasets.
4.	No ability to confirm derivation of intermediate datasets or custom domains.
5.	No ability to determine location of collected CRF variables in the converted SDTM data.
6.	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 data are independently converted to SDTM and ADaM formats rather than ADaM datasets being created from the SDTM dataset.

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

Issues
1. No traceable path from legacy to SDTM to ADaM and to Study Report.
2. No explanation or source for analysis imputed or derived variables.
3. No traceable path to ISS and ISE / pooled data.
4. TLFs do not match datasets (analysis datasets or SDTM datasets (when used)).
5. No traceable path to intermediate datasets or custom domains.
6. No explanation or source for imputed or derived variables in datasets.

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

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

1. May not be able to replicate the results in the TLFs and CSR using the ADaM or the SDTM datasets.

8.3.4. Conversion of Legacy Study Data

Sponsors should evaluate the risk involved in converting legacy study data to standardized data (e.g., CDISC SDTM, SEND, and ADaM). Sponsors should provide the explanation and rationale for the study data conversion in the *Data Guide*. 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 used for the conversion.
   - The report should present the results of the conversions, issues encountered and resolved, and outstanding issues.

2. Provide an annotated CRF that maps the legacy data elements.
   - Sponsors should provide two CRF annotations, one based on the original legacy data, and the other based on the converted data. 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 *Data Guide*. 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. Sponsors should consider the submission of the legacy data (i.e., legacy CRF/aCRF, legacy tabulation data, and legacy analysis data) in addition to the converted.

Appendix A: 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.[38] For example, the number *44* itself is meaningless without an association with *Hematocrit*. 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 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.
Position = lying

[38] 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.

When presented as a series of unrelated data elements, they cannot reliably be interpreted. Once the relationships are captured, as shown simply below using arrows:

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

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.

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 informatics community. In August 2006, the President signed an Executive Order mandating that the Federal Government use interoperable data standards for health information exchange.[39] Although this order was directed at Federal agencies that administer health care programs (and therefore not FDA), it is relevant to this guidance because it defined interoperability:

"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.[40] 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.

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

[39] See http://www.cga.ct.gov/2006/rpt/2006-R-0603.htm.
[40] See Coming to Terms: Scoping Interoperability for Health Care http://www.hln.com/assets/pdf/Coming-to-Terms-February-2007.pdf.

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 both parties.

Glossary of Acronyms

The following list of acronyms are used in this Guide:

aCRF:	Annotated Case Report Form
ANDA:	Abbreviated New Drug Application
ADaM:	CDISC Analysis Dataset Model
ADSL:	CDISC Analysis Data Subject Level
ASCII:	American Standard Code for Information Interchange
AVISIT:	CDISC ADaM Visit name
AVISITN:	CDISC ADaM Visit number
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
COMPTRT:	CDISC Comparative Treatment
CURTRT:	CDISC Current Treatment
Domain:	A collection of observations with a topic-specific commonality (CDISC)
DY:	CDISC Study Day of Visit
DTC:	CDISC Date / Time of Collection
eCTD:	Electronic Common Technical Document
ENDY:	CDISC Study Day of End of Observation
EPOCH:	CDISC Trial Epoch
ICH:	International Conference on Harmonisation
IND:	Investigational New Drug
ISO:	International Organization for Standardization
ISO 860:	ISO character representation of dates, date/times, intervals, and durations of time.
ISS:	Integrated Summary of Safety
ISE:	Integrated Summary of Efficacy
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
PCLAS:	CDISC Pharmacologic Class
PDF:	Portable Document Format
PE:	Physiologic Effect
SDTM:	Study Data Tabulation Model
SEENDY:	CDISC Study Day of End of Observation (Subject Elements Domain)
SEND:	Standard for Exchange of Non-Clinical Data
SESTDTC:	CDISC Start Date/Time of Element (Subject Elements Domain)
SESTDY:	CDISC Study Day of Start of Observation (Subject Elements Domain)
SNOMED:	Systematized Nomenclature of Medicine
STDY:	CDISC Study Day of Start of Observation
SUBJID:	Study Identifier

SUPPQUAL:	CDISC Supplemental Qualifier dataset
TDIGRP:	CDISC Diagnosis Group
TLF:	Tables, Listing and Figures
TSPARM:	CDISC Trial Summary Parameter Test Name
UNII:	FDA Unique Ingredient Identifier
USUBJID:	CDISC Unique Subject Identifier
XML:	eXtensible Markup Language
XPORT:	SAS Transport Version 5