Study Data Specifications

July 18, 2012

Version 2.0

Study Data Specifications

Revision History

Date	Version	Summary of Changes
2004-07	1.0	Original version
2005-03-18	1.1	Addition of specifications for define.xml and SAS XPORT transport files specifications. Changes in document organization.
2006-03-04	1.2	Update information on annotated ECG waveform data. Delete ecg folder under Specifications for Organizing the Datasets.
2006-11-27	1.3	Addition of specifications for submitting tumor datasets (tumor.xpt) from rodent carcinogenicity studies.
2007-08-01	1.4	Addition of hyperlink to information for 3.1.1 datasets
2009-10-30	1.5	Modified introduction. Additional specifications for submitting data tabulation datasets. Additional specifications for analysis datasets. Revision of maximum file size restrictions. Addition of hyperlink to information for 3.1.2 datasets.
2010-01-04	1.5.1	Modified directory tree structure for organizing datasets and made minor technical corrections.
2011-06-22	1.6	Addition of specifications for general toxicology and carcinogenicity data tabulation datasets
2012-07-20	2.0	Reorganization , minor editorial changes

1 Introduction

These specifications provide useful technical instructions for submitting animal and human study data and related information in electronic format. The specifications are intended to complement other resources located on the FDA Study Data Standards Resources web page.[1] Datasets are views of the study data used by reviewers to conduct specific analyses. They may include both raw and derived data. Because of the inherent variability in the scientific review process across studies and applications, it is impossible to enumerate *a priori* all datasets needed for review. Prior to the submission, sponsors should discuss with the review division the datasets that should be provided, the data elements that should be included in each dataset, and the organization of the data within the datasets. Additionally, not all FDA centers have adopted all aspects of these specifications; sponsors are advised to discuss with the reviewing division data needs prior to preparing data for submission.

2 Dataset Specifications

2.1 File Format

SAS XPORT Transport File format

SAS XPORT transport file format, also called Version 5 SAS transport format, is an open format published by the SAS Institute. The description of this SAS transport file format is in the public domain. Data can be translated to and from this SAS transport format to other commonly used formats without the use of programs from SAS Institute or any specific vendor.

Version

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

Sponsors can find the record layout for SAS XPORT transport files through SAS technical support technical document TS-140. This document and additional information about the SAS Transport file layout can be found on the SAS World Wide Web page at http://www.sas.com/fda-esub.

Transformation of Datasets

SAS XPORT transport files can be converted to various other formats using commercially available off the shelf software.

[1] See http://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm

SAS Transport File Extension

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

Compression of SAS Transport Files

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

2.2 Size (overall file size)

Each dataset is provided in a single transport file. The maximum size of an individual dataset is dependent on many factors. In general, datasets greater than 1 GB in size should be split into smaller datasets, each no larger than 1GB in size. Datasets divided to meet the maximum size restrictions should contain the same variable presentation so they can be easily concatenated.

Datasets which are divided should be clearly named to aid the reviewer in reconstructing the original dataset, e.g., xxx1, xxx2, xxx3, etc. The files that have been divided and need to be concatenated should be noted in the data definition document. This documentation should identify the range of subject numbers (or other criteria used for division) in the label for each of the divided datasets. For further information on file size limitations for files submitted to CDER, contact eData@fda.hhs.gov, for files submitted to CBER, contact CBER.CDISC@fda.hhs.gov

2.3 Sizing of Variables/ Data Elements

Maximum Size of Variables

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

2.4 Sizing of Columns

For all datasets, in order to significantly reduce dataset file sizes, the allotted character column length/size for each column should be the maximum length used. Lengths/sizes of columns should not arbitrarily be set to 200. For example, if USUBJID has a maximum length of 18, the USUBJID’s column size should be set to 18, not 200. An inclusion of a small amount of padding to column width may be acceptable as long as this doesn’t result in significant increases in file size.

2.5 General Considerations for all Datasets

- For an individual study, all dataset names and dataset labels should be unique across both the analysis and tabulation datasets submitted for an individual study. The internal name for an analysis dataset should be the same as the name shown in the data definition file.

- For unscheduled visits or measurements, numbers are often assigned values between two protocol-scheduled visits. These numbers should be distinct from other visit numbers but retain the chronological order (e.g. two unscheduled visits between visit 3 and visit 4 might be 3.1 and 3.2). The character form of the visit identifier may be UNSCHEDULED or a similar term.

- Variable names and codes should be consistent across studies and where feasible, the NCI CDISC Vocabulary should be used. For example, if glucose is collected in a number of studies, use the CDISC Submission Value "GLUC" for the laboratory test code in all of the studies. The NCI CDISC terminology is available at http://www.cancer.gov/cancertopics/cancerlibrary/terminologyresources/cdisc.

- The format of variables for similar types of data should be consistent within and across studies. For example, all variables that include calendar dates (e.g. birth date, screening visit date, randomization date, date of death) should use the same format for representing the date. The data type may be different between the tabulation and analysis datasets but should be consistent within the tabulation or analysis datasets.

- The key variables (e.g., subject identifier and visit for datasets with multiple records per subject) should appear first in the datasets. Each subject should be identified by a single, unique subject identifier within an entire application (including tabulation, listing and analysis datasets). Subjects enrolled in a primary study and then followed into an extension study should retain their unique identifier from the primary study.

- Each analysis dataset should be described by an internal label which is shown in the data definition file.

2.6 Data Tabulations

2.6.1 Definition

Data tabulations are datasets in which each record is a single observation for a subject.

2.6.2 Standardized (SDTM, SEND)

Data specifications for the Data Tabulation datasets of human drug product clinical and animal toxicology studies[2] are provided by the Study Data Tablulation Model (SDTM[3]). The FDA

[2] Here, "drug product" also includes biologic products (submitted as BLAs) that are reviewed in CDER

Study Data Standards Catalog provides a tabular listing of all the currently supported data standards, with links to important reference materials[4].

While the SDTM provides a valuable representation that may facilitate review, it does not always provide data structured in a way that supports all analyses needed for review. Sponsors should therefore augment the SDTM with analysis datasets as described in the *Analysis Datasets* section.

Standard for the Exchange of Nonclinical Data (SEND) is an implementation for the SDTM standard and the same specifications apply.

2.6.3 Non-standard (legacy)

The submission of non-standardized datasets is not recommended. If provided, each dataset is provided as a SAS Transport (XPORT) file. There are no further specifications for organizing legacy datasets.

2.7 Analysis Datasets

2.7.1 Definition

Analysis datasets are created to support results presented in study reports, the Integrated Summary of Safety (ISS) and the Integrated Summary of Efficacy (ISE) and to support other analyses that enable a thorough regulatory review. Analysis datasets contain both raw and derived data.

2.7.2 General Considerations for Analysis Datasets

- At least one analysis dataset should be labeled in the data definition file as containing the primary efficacy data.

- When a dataset contains multiple records per subject, a variable for relative day of measurement or event and variables for visit should be included. In addition to a protocol-scheduled visit variable, include at least two timing variables; a character variable describing the visit (e.g. WEEK 8) and a corresponding numeric variable (e.g. 8). These two variables are measures of time from randomization.

- Core variables should be listed after the key variables (USUBJID and visit) and included on each analysis dataset. Core variables include study/protocol, center/site, country, treatment

[3] SDTM refers to the Study Data Tabulation Model (SDTM) developed by the Submission Data Standard working group of the Clinical Data Interchange Standard Consortium (CDISC). The model has two implementation guides: human clinical (SDTM IG) and animal nonclinical studies (SEND IG). In this document, SDTM refers to both models unless a specific IG is stated.

[4] See http://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM292505.xls

assignment, sex, age, race, analysis population flags (e.g. ITT, safety) and other important baseline demographic variables.

- .For textual data that have been mapped to numeric codes, provide two variables, one with text and one with numeric codes.

- Dates should be formatted as numeric in the analysis datasets even if dates are in ISO8601 or another character format in the raw data. This formatting will facilitate calculations, such as duration.

- Any submitted programs (scripts) generated by an analysis tool should be provided as ASCII text files or PDF 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 ASCII text program files, e.g. adsl_r.txt or adsl_sas.txt, etc.

2.7.3 Standardized

2.7.3.1 ADaM – Clinical

Specifications for Analysis datasets for human drug product clinical and analytical studies are provided by the ADaM (http://www.cdisc.org/adam). The FDA Study Data Standards Catalog provides a tabular listing of all the currently supported data standards with links to important reference materials.

While the ADaM provides a valuable representation that may facilitate review, it does not always provide data structured in a way that supports all analyses needed for review. Sponsors should therefore augment their ADaM analysis datasets with analysis datasets based on requirements specified by the review division.

2.7.4 Non-standard (legacy)

FDA review divisions may request non-standard analysis datasets to facilitate regulatory review. Some requested datasets may be formatted to allow for specific analyses by FDA review divisions. An example of a non-standard analysis dataset is a tumor dataset that is described in the following section.

2.7.4.1 Tumor.xpt – Carcinogenity Studies

CDER statisticians perform analyses on the tumor data from each rodent carcinogenicity study and they need the tumor data to be provided as an electronic analysis dataset. The Table below specifies the recommended data elements to be included in the analysis dataset. The dataset containing this information should be named tumor.xpt to aid in identification.

Tumor Dataset For Statistical Analysis[1,2] (tumor.xpt)

Variable	Label	Type	Codes	Comments
STUDYNUM	Study number	char		[3]
ANIMLNUM	Animal number	char		[1,3]
SPECIES	Animal species	char	M=mouse R=rat	
SEX	Sex	char	M=male F=female	
DOSEGP	Dose group	num	Use 0, 1, 2, 3,4,... in ascending order from control. Provide the dosing for each group.	
DTHSACTM	Time in days to death or sacrifice	num		
DTHSACST	Death or sacrifice status	num	1 = Natural death or moribund sacrifice 2 = Terminal sacrifice 3 = Planned intermittent sacrifice 4= Accidental death	
ANIMLEXM	Animal microscopic examination code	num	0= No tissues were examined 1 = At least one tissue was examined	
TUMORCOD	Tumor type code	char		[3,4]
TUMORNAM	Tumor name	char		[3,4]
ORGANCOD	Organ/tissue code	char		[3,5]
ORGANNAM	Organ/tissue name	char		[3,5]
DETECTTM	Time in days of detection of tumor	num		
MALIGNST	Malignancy status	num	1 = Malignant 2= Benign 3 = Undetermined	[4]
DEATHCAU	Cause of death	num	1 = Tumor caused death 2= Tumor did not cause death 3 = Undetermined	[4]
ORGANEXM	Organ/Tissue microscopic examination code	num	1 = Organ/Tissue was examined and was usable 2= Organ/Tissue was examined but was not usable (e.g., autolyzed tissue) 3 = Organ/Tissue was not examined	

Each animal in the study should have at least one record even if it does not have a tumor.

[1] Additional variables, as appropriate, can be added to the bottom of this dataset.

[2] ANIMLNUM limited to no more than 12 characters; ORGANCOD and TUMORCOD limited to no more

than 8 characters; ORGANNAM and TUMORNAM limited to no more than 30 characters.

3. A missing value should be given for the variable MALIGNST, DEATHCAU, TUMOR and TUMORCOD when the organ is unusable or not examined.
4. Do not include a record for an organ that was useable and no tumor was found on examination. A record should be included for organs with a tumor, organs found unusable, and organs not examined.

3 Dataset Documentation

3.1 Data Definition/Metadata

3.1.1 Definition

The data definition file describes the format and content of the submitted datasets.

3.1.2 Specification

3.1.2.1 Standardized

- The specification for the data definitions for datasets provided using CDISC is included in the Case Report Tabulation Data Definition Specification (define.xml) developed by the CDISC define.xml Team. The latest release of the define.xml is available from the CDISC web site (http://www.cdisc.org/define-xml). Include a reference to the style sheet as defined in the specification and place the corresponding style sheet in the same folder as the define.xml file.

- The internal dataset label should clearly describe the contents of the dataset. For example, the label for an efficacy dataset might be "TIME TO RELAPSE (EFFICACY)".

3.1.2.2 Non-standard (Legacy)

- The data definition tables should be provided as a single PDF file named define.pdf and placed in the appropriate study, specific analysis type or integrated summary folder in the datasets folder. The Title portion of the Document Information field of each data definition file should include the appropriate study report number, specific analysis type or integrated summary name and data definitions. For example, the data definition file for study 2001 would be identified as: study 2001, data definitions. This file is considered part of the comprehensive table of contents.

- For datasets not prepared using CDISC specifications, sponsors should include a define.pdf to describe the datasets for each study, specific data analysis (e.g., population PK), and integrated summaries. For the datasets to be useable, the definitions of the variables should be provided. Sponsors should document all of the variables in the datasets in data definition tables. There should be one set of data definition tables for each study, specific data analysis (e.g., population PK) and integrated summary. The first table should include a listing of all datasets provided for the study with a description of the dataset and the location of the dataset file. Provide a hypertext link from the description of the dataset to the appropriate data definition table. Provide a hypertext link from the location listing of the file to the SAS transport file. The reviewer can use the first hypertext link to view the data definition table and the second to open the SAS transport dataset file. For clinical study data, sponsors should also provide a link to the appropriate annotated case report form file (blankcrf.pdf).

- In the following table, the dataset for AE is described as adverse events, and the dataset file is located in listings folder for study 1234

Datasets for Study 1234		
Dataset	**Description of Dataset**	**Location**
AE	Adverse Events	m5/datasets/study1234/listings/ae.xpt
…	…	…

- Subsequent pages should contain a table for each dataset that includes an organized listing of all variable names used in the dataset, a descriptive variable label, data types, codes (and decodes), and comments. The comments field is for further description of the variables. For derived variables, the method for calculating the variable should be included in the comments field. For raw variables in clinical datasets, the location of the variable on the annotated CRF should be provided as well as the CRF field name if different from the variable name in the dataset. Providing a hypertext link from each raw data variable in the data definition table to the appropriate location of the blankcrf.pdf also helps the review process. An example of part of a data definition table for the demographics dataset for study 1234 is provided below.

Study 1234 – Demographics Dataset Variables			
Label	**Type**	**Codes**	**Comments**[1]
Unique subject ID number	char		Demographics page 3

Sex of subject	char	f = female
		m = male
			Demographics page 3

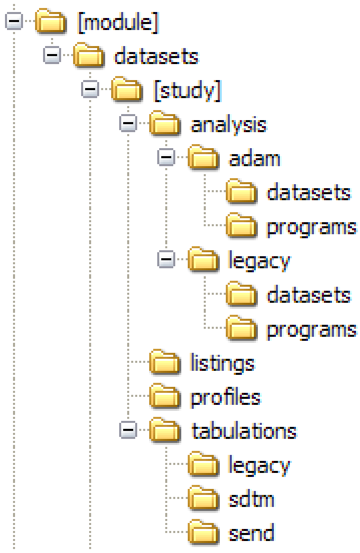
Birth date	date		Demographics page 3
Duration of Treatment	num		Derived STOP DATE – START DATE
Assigned treatment group	num	0= placebo 5= 5mg/day	

[1] Use footnotes for longer comments

4 Organization within a submission

4.1 eCTD

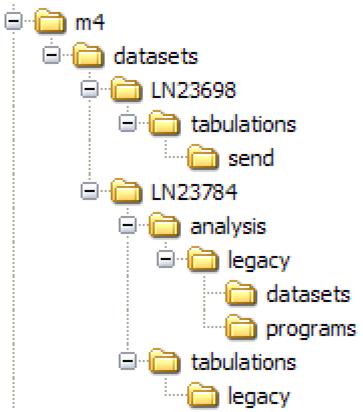
The specifications for organizing study datasets and their associated files in folders are summarized in the following figure and accompanying table. No additional subfolders are needed; unused folders do not need to be supplied.



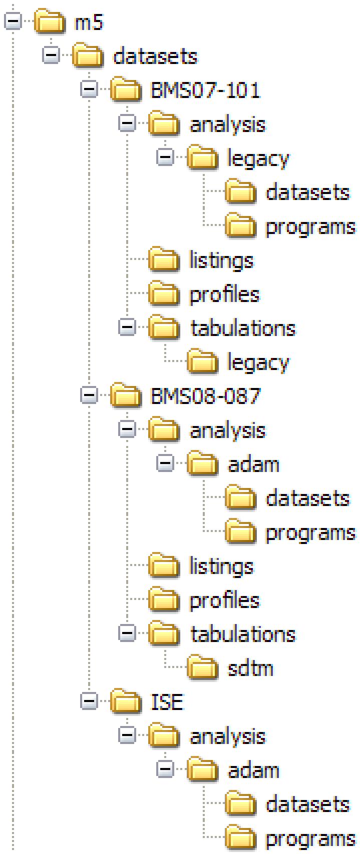
Folder	**Description**
[module name]	Name the folder according to the CTD module which the datasets apply to. Use m4 for non-clinical data and m5 for clinical data.
datasets	The folder under which all of the study data being submitted for the module specified is organized.
[study]	Name the folder according 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 programs will be organized **Note:** The analysis datasets and programs are to be placed in specific folders based on their format.
adam	The sub-folder under which ADaM formatted datasets and programs are organized
datasets	The sub-folder in which the analysis datasets are organized
programs	The sub-folder in which the analysis programs are organized
legacy	The sub-folder under which legacy formatted datasets and programs are organized
datasets	The sub-folder in which the analysis datasets are organized
programs	The sub-folder in which the analysis programs are organized
listings	The folder in which miscellaneous datasets that don't qualify as analysis, profile, or tabulation datasets are organized
profiles	The folder in which patient profiles are organized.

Folder	Description
tabulations	The folder under which tabulation datasets will be organized. **Note:** The tabulations datasets are to be placed in specific folders based on their format.
legacy	The sub-folder in which tabulations not formatted according to an identified standard format are organized, e.g., non-SDTM datasets.
sdtm	The sub-folder in which tabulations formatted according to the SDTM IG standard are organized. Should only be used in m5 for clinical data
send	The sub-folder in which tabulations formatted according to the SEND IG standard are organized. Should only be used in m4 for animal data

The following example shows the folder structure for a submission containing 2 individual nonclinical studies and 2 clinical studies and an ISE. SEND IG tabulations have been submitted for LN23698. Legacy tabulations and analysis data (such as a tumor.xpt) have been submitted for LN23784. Legacy clinical study data tabulations have been submitted for BMS07-101. SDTM IG and ADaM data has been submitted for BMS08-087.



```
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
```

5 Other Types of Data

5.1 Annotated ECG Waveform Data

5.1.1 Definition

These are raw voltage-versus-time data comprising the electrocardiogram recording, to which have been attached the identification of various intervals or other features.

5.1.2 Specification

See the HL7 normative standard for creating the annotated ECG waveform data files. This information may be found on the HL7 web site www.hl7.org. More information may be found at:

http://www.fda.gov/Drugs/DevelopmentApprovalProcess/HowDrugsareDevelopedandApproved/ApprovalApplications/AbbreviatedNewDrugApplicationANDAGenerics/ucm154163.htm#ECG

5.2 Data Listings

5.2.1 Definition

Data listings are datasets in which each record is a series of observations collected for each subject during a study or for each subject for each visit during the study organized by domain.

5.2.2 Specification

Each dataset is provided as a SAS Transport (XPORT) file. Currently, there are no further specifications for organizing data listing datasets.

5.3 Subject Profiles

5.3.1 Definition

Subject profiles are displays of study data of various modalities collected for an individual subject and organized by time.

5.3.2 Specification

Each individual patient's complete patient profile is in a single PDF file. Including the patient ID in the file name will help identify the file. Alternatively, all patient profiles for an entire study may be in one file if the size of each individual patient profile is small and there are not a large number of patient profiles needed for the study. If you do the latter, bookmark the PDF file using the subject's ID. Including the study number in the file name will help identify the file.

5.4 Annotated Case Report Form

5.4.1 Definition

This is a blank case report form with annotations that document the location of the data with the corresponding names of the datasets and the names of those variables included in the submitted datasets.

5.4.2 Specification

The annotated CRF is a blank CRF that includes treatment assignment forms and maps each item on the CRF to the corresponding variables in the database. The annotated CRF should provide the variable names and coding for each CRF item included in the data tabulation datasets. All of the pages and each item in the CRF should be included. The sponsor should write *not entered in database* in all items where this applies. The annotated CRF should be provided as a PDF file. Name the file *blankcrf.pdf*.