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
2009-11-03	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.

STUDY DATA SPECIFICATIONS

These specifications are for submitting animal and human study datasets in electronic format. Datasets are views of the study data used by reviewers to conduct specific analyses of the study data. They may include both raw and derived data. Because of the unpredictability of the scientific review process, it is impossible to enumerate a priori all datasets needed for review. Prior to the submission, you 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 file. 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.

SAS XPORT TRANSPORT FILE FORMAT

SAS XPORT transport 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.

You 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.

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.

CONTENT OF DATASETS AND SIZE OF DATASETS

Each dataset is provided in a single transport file. The maximum size of an individual dataset is dependent on many factors. In general, datasets other than SDTM datasets, should be less than 400 MB; SDTM datasets should not be divided. . 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 patient 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 cder-edata@fda.hhs.gov .

Variable names should be limited to 8 characters, and the accompanying descriptive name in the label header can be up to 40 characters.

SPECIFICATIONS FOR SPECIFIC DATASETS AND DOCUMENTATION

Study data are provided using different presentations: Data Tabulation Datasets, Data Listing Datasets, Subject Profiles, and Analysis Datasets.

Data tabulation datasets

Definition

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

Specifications

Specifications for the Data Tabulation datasets of human drug product clinical studies[1], are provided by the Study Data Tabulation Model (SDTM) developed by the Submission Data Standard working group of the Clinical Data Interchange Standard Consortium (CDISC) [2]. FDA centers and reviewing divisions differ in their use of SDTM data. CDER currently accepts SDTM datasets prepared in accordance with the SDTM implementation guide versions listed in the following table. CBER is currently testing SDTM for clinical studies biologic products and for animal toxicity studies. Individual centers or reviewing divisions may specify the version of SDTM needed for review (see

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

[2] CDISC, www.cdisc.org, is an open, multidisciplinary, not-for-profit organization committed to the development of worldwide industry standards to support the electronic acquisition, exchange, submission and archiving of clinical trials data and metadata for medical and biopharmaceutical product development.

below). Follow the corresponding hyperlink to view the appropriate SDTM and implementation guide.

Each SDTM dataset is provided as a SAS Transport (XPORT) file.

Version	Implementation Guide	Support Begins	Support Ends
3.1	http://www.cdisc.org/models/sds/v3.1/index.html	2004-07-01	2010-3-31
3.1.1	http://www.cdisc.org/content1605	2007-08-01	
3.1.2	http://www.cdisc.org/content1055	2009-10-30	

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 SDTM with analysis data sets as described in the *Analysis datasets* section.

Currently, CDER statisticians perform analyses on the tumor data from each rodent carcinogenicity study, and they need this information provided as an electronic dataset. See Appendix 1 on data elements for the dataset recommended by these statistical reviewers (tumor.xpt). This information will be needed until testing is completed on SDTM for animal toxicity studies.

Data Listings

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.

Specifications

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

Subject profiles

Definition

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

Specifications

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.

Analysis datasets

Definitions

Analysis datasets are datasets created to support results presented in study reports, ISS, ISE and other analyses that enable a thorough regulatory review. Analysis datasets contain both raw and derived data.

Specifications

Each dataset is provided as a SAS Transport (XPORT) file. Prior to submission, sponsors should contact the appropriate center's reviewing division to determine the division's analysis dataset needs. CDISC/ADaM standards for analysis datasets (http://www.cdisc.org/adam) may be used if acceptable to the review division.

Any requested programs (scripts) generated by an analysis tool should be provided as ASCII text 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 for ASCII text files other than .txt, the file name should include the native file extension generated by the analysis tool for ASCII text program files, e.g. myRcode.r, mySAScode.sas, etc. If the analysis tool does not save programs in ASCII format, a PDF rendition of the program file should be provided in addition to the program file.

General Considerations for Datasets

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

- Each analysis dataset should be described by an internal label (up to 40 characters) which is shown in the data definition file. This label should clearly describe the contents of the dataset. For example, the label for an efficacy dataset might be "TIME TO RELAPSE (EFFICACY)". At least one analysis dataset should be labeled in the data definition file as containing the primary efficacy data.

- The key variables (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.

- When a dataset contains multiple records per patient, 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.

- 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.

- Core variables should be listed after the key variables and included on each analysis dataset. Core variables include study/protocol, center/site, country, treatment assignment, sex, age, race, analysis population flags (e.g. ITT, safety) and other important baseline demographic variables.

- The variable names and codes should be consistent across studies. For example, if glucose is collected in a number of studies, use the same laboratory test name (e.g. GLUCOSE) to describe this test in all of the studies. When feasible, the currently approved NCI/Janus terminology codes should be used (http://evs.nci.nih.gov/ftp1/CDISC/SDTM/).

- 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.

- .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 the calculation of duration.

SPECIFICATIONS FOR DATASETS DOCUMENTATION

Dataset documentation includes data definitions and annotated case report forms.

Data definition file

Definition

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

Specifications

The specification for the data definitions for datasets provided using the CDISC SDTM is included in the Case Report Tabulation Data Definition Specification (define.xml) developed by the CDISC define.xml Team. The latest release of the Case Report Tabulation Data Definition Specification is available from the CDISC web site (http://www.cdisc.org/models/def/v1.0/index.html). 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.

For datasets not prepared using the CDISC SDTM specifications, consult Appendix 2 for information concerning the preparation of a define.pdf data definition file.

Annotated case report form

Definition

This is a blank case report form 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.

Specifications

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*.

SPECIFICATIONS FOR OTHER TYPES OF STUDY DATA

Annotated ECG waveform data

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.

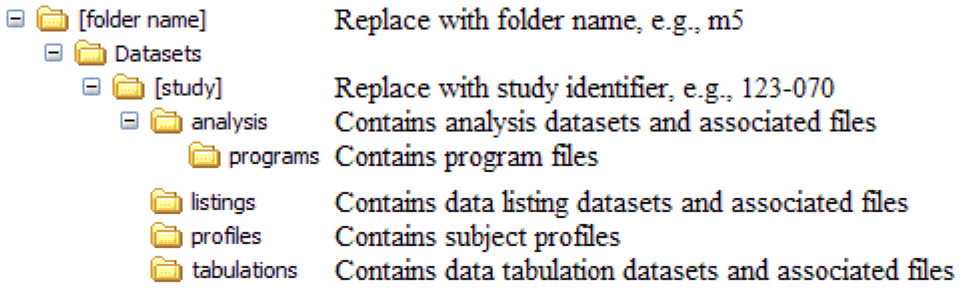
Specifications

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/FormsSubmissionRequirements/ElectronicSubmissions/ucm085324.htm#ECG.

SPECIFICATIONS FOR ORGANIZING THE DATASETS

The specifications for organizing study datasets and their associated files in folders are summarized in the following figure. No additional subfolders are needed.



[folder name]	Replace with folder name, e.g., m5
Datasets	
[study]	Replace with study identifier, e.g., 123-070
analysis	Contains analysis datasets and associated files
programs	Contains program files
listings	Contains data listing datasets and associated files
profiles	Contains subject profiles
tabulations	Contains data tabulation datasets and associated files

APPENDIX 1

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	

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

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

[3] ANIMLNUM limit to no more than 12 characters; ORGANCOD and TUMORCOD limited to no more than 8 characters; ORGAN and TUMOR should be as concise as possible.

[4] A missing value should be given for the variable MALIGNST, DEATHCAU, TUMOR and TUMORCOD when the organ is unusable or not examined.

[5] 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.

APPENDIX 2

You 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. You 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. You 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 (up to 8 characters) used in the dataset, a descriptive variable label (up to 32 characters), 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, 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				
Variable	**Label**	**Type**	**Codes**	**Comments**[1]
USUBJID	Unique patient ID number	char		Demographics page 3

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

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

[1] Use footnotes for longer comments

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.