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

STUDY DATA SPECIFICATIONS

These specifications are for submitting animal and human study data in electronic format. Study data includes information from trials submitted to the agency for evaluation and information to understand the data (data definition). The study data includes both raw and derived data.

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. If a dataset is going to be over 100 MB, you should contact the center to discuss if the dataset should be divided. In general, Data Tabulation Datasets provided following the Study Data Tabulation Model (see below) can be larger than 100 MB and do not need to be divided. Datasets divided to meet the maximum size restrictions should contain the same variable presentation so they can be easily merged, joined, and concatenated. Variable names should be limited to 8 characters and 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 located in the Study Data Tabulation Model (SDTM) developed by the Submission Data Standard working group of the Clinical Data Interchange Standard Consortium (CDISC)[2]. This folder is reserved for the datasets conforming to the SDTM standard. The latest release of the SDTM and implementation guides for using the model in clinical trials is available from the CDISC web site. FDA currently accepts SDTM datasets prepared in accordance with the SDTM implementation guide versions listed in the following table. Follow the corresponding hyperlink to view the appropriate SDTM and implementation guide.

Version	Implementation Guide
3.1	www.cdisc.org/models/sds/v3.1/index.html
3.1.1	http://www.cdisc.org/models/sdtm/v1.1/index.html

This SDTM is currently being tested for clinical studies involving biologic products and for animal toxicity studies. The implementation guide for using the model for animal toxicology data is being developed by the Standard for Exchange of Nonclinical Data (SEND) consortium of CDISC. This implementation guide will also be available on the CDISC web site. Updates to the SDTM and implementation guides as a result of this testing will be available through the CDISC web site.

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

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.

[1] Drug products also includes biologic products 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.

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. General information about creating datasets can be found in the SDTM implementation guides referenced in the data tabulation dataset specifications.

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.

Analyses datasets

Definition

Analysis datasets are datasets created to support specific analyses. Programs are scripts used with selected software to produce reported analyses based on these datasets.

Specifications

Each dataset is provided as a SAS Transport (XPORT) file. Programs should be provided as both ASCII text and PDF files and should include sufficient documentation to allow a reviewer to understand the submitted programs. It is not necessary to provide analysis datasets and programs that will enable the reviewer to directly reproduce reported results using agency hardware and software. Currently, there are no other additional specifications for creating analysis datasets.

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.

See *Providing Regulatory Submissions in Electronic Format –NDA* for details on data definition files for other datasets.

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

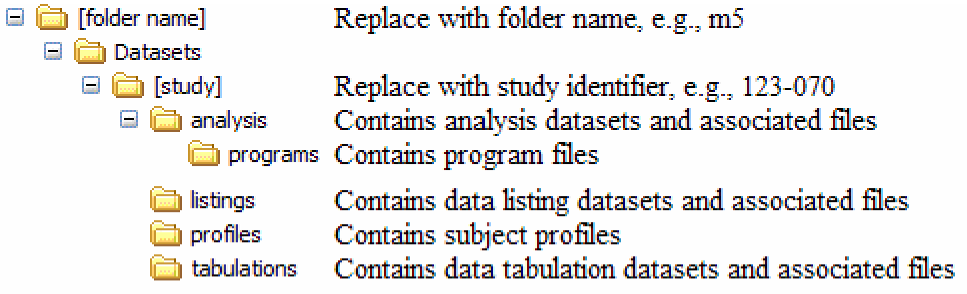
These are raw voltage-versus-time data comprising the electrocardiogram recording, to which have been attached the identification of various intervals or other features.

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/cder/regulatory/ersr/default.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.