#197

Documenting Electronic Data Files and Statistical Analysis Programs

Guidance for Industry

This version of the guidance replaces the version made available December 2015. The document has been revised to clarify existing language, remove recommendations that are no longer applicable, and provide additional details on the README file.

Submit comments on this guidance at any time. Submit electronic comments to https://www.regulations.gov. Submit written comments to the Dockets Management Staff (HFA-305), Food and Drug Administration, 5630 Fishers Lane, Rm. 1061, Rockville, MD 20852. All comments should be identified with the docket number FDA-2009-D-0052.

For further information regarding this document, contact AskCVM@fda.hhs.gov.

Additional copies of this guidance document may be requested from the Policy and Regulations Staff (HFV-6), Center for Veterinary Medicine, Food and Drug Administration, 7500 Standish Place, Rockville MD 20855, and may be viewed on the Internet at either https://www.fda.gov/animal-veterinary or https://www.regulations.gov.

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Veterinary Medicine (CVM)
November 2020

Table of Contents

Contains Nonbinding Recommendations

Documenting Electronic Data Files and Statistical Analysis Programs

Guidance for Industry

This guidance represents the current thinking of the Food and Drug Administration (FDA or Agency) on this topic. It does not establish any rights for any person and is not binding on FDA or the public. You can use an alternative approach if it satisfies the requirements of the applicable statutes and regulations. To discuss an alternative approach, contact the FDA staff responsible for this guidance as listed on the title page.

I. Introduction

This guidance is provided to inform sponsors of recommendations for documenting electronic data files and statistical analyses submitted to the Center for Veterinary Medicine (CVM) to support new animal drug applications. These recommendations are intended to reduce the number of revisions that may be required for CVM to effectively review data submissions. They are also intended to simplify submission preparation for sponsors by providing a suggested documentation framework, including a sample structure on how to describe and organize the information regarding the electronic data files and statistical analysis programs.

The determination of what data and analysis are needed to support a new animal drug application may vary depending on many factors and is outside the scope of this document. You should refer only to those portions of this guidance that are applicable to your particular submission.

In general, FDA's guidance documents do not establish legally enforceable responsibilities. Instead, guidances describe the Agency's current thinking on a topic and should be viewed only as recommendations, unless specific regulatory or statutory requirements are cited. The use of the word *should* in Agency guidances means that something is suggested or recommended, but not required.

II. Background

For new animal drug applications, FDA requires full reports of investigations which have been conducted to show a drug is safe and effective for use [section 512(b)(1)(A) of the Federal Food, Drug, and Cosmetic Act (the FD&C Act)]. Additionally, section 512(n)(1)(E) of the FD&C Act requires that abbreviated applications for the approval of a new animal drug contain information to show that the new animal drug is bioequivalent to the approved new animal drug.

Submissions to CVM in support of new animal drug applications generally include a Final Study Report (FSR)[1] for one or more studies. For each study that includes electronic data files, CVM

[1] For the purpose of this guidance, supporting study reports (e.g., statistical or bioanalytical phase reports) appended to the FSR are considered part of the FSR.

needs information regarding the process for data generation and the statistical analysis conducted to review submissions and verify that there is sufficient quality and detail of evidence to support an animal drug application. An adequately documented submission should include readable electronic data files, a description of how the data were processed, and a description of the statistical analyses employed to support your conclusions.

Your submission should clearly describe the entire process by which the data were collected (e.g., transcribed or electronic data capture), including a record of all changes to the data, starting from the electronic data files created from transcribed case report forms, or from electronic data capture (EDC) systems, to the completed statistical analyses which form the basis for your study's conclusions. To understand the process by which you compiled the data and conducted the statistical analysis, CVM needs to understand the contents of each electronic data file, the computer programs that processed all the electronic data files for analysis, and the programs that implemented the statistical analyses. The information that should be submitted to CVM, together with the FSR and electronic data files, are described in sections III. *Structure and Content of README Files* and IV. *Data and Analysis Programs: Additional Comments*. Section III describes recommendations for how README files should be organized and completed to describe the data files and analyses, and section IV contains additional recommendations regarding statistical analysis programs.

III. Structure and Content of README Files

Data submissions to CVM via eSubmitter[2] typically include data in Statistical Analysis System (SAS) transport XPORT (XPT) or eXtensible Markup Language (XML) format, analysis programming files in XML format, and documents in Portable Document Format (PDF). A manual or instructional resource that contains important information about the electronic data files in the submission should be submitted to CVM in a README file. This section describes the information that should be in the README file for CVM to evaluate the electronic data files. Electronic data files, in the context of this guidance, refer to the XPT and XML files included in a submission. If any of the information described below is contained within the FSR it does not need to be repeated in the README file. Instead, the README file may reference the corresponding location of the information in the FSR.

A. The README File

An overview of electronic data files, documentation, and programming files included in the submission helps CVM assess the submission and the files for review. The README file contains information about the electronic data files and programming files in a submission. The README file explains how the data files are organized and describes the programs used for data file generation and data analysis. An effective README file should quickly orient the user to crucial information needed to understand the electronic data files in a submission.

[2] For current eSubmitter file specifications, see CVM Recommended File Specifications in the CVM eSubmitter Resource Center: https://www.fda.gov/industry/fda-esubmitter/cvm-esubmitter-resource-center.

The README file is typically a PDF file with the filename README.pdf. The README file should not contain data, interpretation of the data, literature, references, notes to file, protocol-associated documents, communication records, personnel records, information not needed to interpret the data provided in the submission, or other information needed for the technical section.

A submission may contain one or more README files depending on the organization and complexity of the submission. README file(s) should be separate from the FSR.

B. Structure of README Files

The README file should include a brief introduction that includes the study number or other identifier, and study descriptor along with general orientation, background, and other information relevant to analyzing and interpreting the data for each study. A description of how the data were processed for analysis could also be included, such as how data were captured and merged to derive the analysis data files or audit trail processes.

The following describes the README file and a sample structure.

1. Electronic Data Files

a. List of Electronic Data Files

In this section, you should provide a table listing the electronic data files submitted in XML or XPT file format with brief descriptions. This table should include the file name, a brief description of contents, name of data collection form(s), if applicable, and information on how the data were collected or any reference to location of information in the FSR that is needed to interpret the data (e.g., reference ranges or scoring definitions). See Example Table 1a. *Listing of XML and XPT Data Files*.

Example Table 1a. Listing of XML and XPT Data Files.

File Name[1]	File contents	Name of Data Collection Form (e.g., Case Report Form)	Comments[2]
ClinObs.xml	Daily clinical observations during hospitalization	Observation Form	Clin Obs were collected via [name of data capture system]
OwnerObs.xpt	Owner observation results for individual animals	Owner Diary	Recorded manually and transcribed
PlateletMan.xpt	Secondary variable: Manual platelet count	Clinical Pathology Form	Reference ranges available in Appendix G of Final Study Report

[1] File name extension included in order to identify the file as XML or XPT.

[2] Information included on how data were collected (paper data collection forms or EDC system) or any reference(s) to location of information in the FSR (e.g., reference ranges or scoring definitions) that are needed to interpret the data file.

b. Electronic Data File Contents

For each data file submitted, you should provide a table that includes the variable names, the abbreviations used in the file, variable label or description, and additional details (e.g., description of coded values, unit of measure, formatting information), if applicable. See Example Table 1b. *Contents of DataFileName.XML*. Results from the CONTENTS procedure in SAS are not sufficient. You may choose to use a standardized internal file format [e.g., Clinical Data Interchange Standards Consortium-Study Data Tabulation Model (SDTM) and Standard Exchange for Nonclinical Data (SEND)][3] to be submitted along with audit trail information if data were collected using EDC (described below).

[3] See Study Data Standards Resources: https://www.fda.gov/industry/fda-resources-data-standards/study-data-standards-resources.

Example Table 1b. Contents of DataFileName.XML (number of observations=, number of variables=).

Variable Name	Variable Label or Description	Additional Details (e.g., descriptions of coded values, units of measure, formatting information)
Animal	Animal Identification	
Dot	Day of treatment	Study day formatted as mm/dd/yyyy
Trt	Treatment	T01=Placebo; T02=Drug
Dscore	Daily Depression Score	0=normal; 1=slight; etc.
Overallscore	Overall Depression Score	Sum of Daily Depression scores (Day 1 to 5)

2. Audit Trail Files

This section only applies to studies that include audit trail files as part of the controls that ensure the authenticity and integrity of records in EDC systems.[4] The audit trail is a portion of the raw data[5] that includes the original recorded data value and any changes to the data value, the identification of individuals that entered or changed data in the EDC system, the date and time the data value was entered or changed, the reason for each change, and the date when the database was locked.[6] If you are submitting audit trails on specified variables, the electronic audit trails should be submitted in an eSubmitter acceptable file format.

CVM prefers to receive one copy of the EDC study database that includes the final variable observations (e.g., the electronic data files (see footnote 2) described in section III.B.1 *Electronic Data Files* above) and associated audit trail information, described in this section, together in one electronic data file or set of data files. If you submit these combined data files, you should also submit the programs needed to create data files that contain the final observations suitable for review and statistical analysis. If you are unable to submit one copy of the EDC study database with associated audit trail information, submission of separate data files for analysis and electronic audit trail files is acceptable.

[4] Electronic Data Capture (EDC) or Electronically Captured Data (ECD) refers to data that was electronically captured at first observation (e.g., web-based software or analytical instrument).

[5] As defined in 21 CFR 58.3(k) and in FDA Guidance for Industry #85 (VICH GL9), "Good Clinical Practice" (https://www.fda.gov/regulatory-information/search-fda-guidance-documents/cvm-gfi-85-vich-gl9-good-clinical-practice).

[6] See Guidance for Industry, "Computerized Systems Used in Clinical Investigations," dated May 2007. (https://www.fda.gov/regulatory-information/search-fda-guidance-documents/computerized-systems-used-clinical-investigations)

a. Audit Trail File Listing

If you are submitting the electronic audit trail separately from the data files for review and statistical analysis, you should provide a table listing the audit trail files submitted. This table should include the file name, the description of the file including EDC system name, and any information necessary for review. See Example Table 2a. *Listing of Audit Trail Files*.

Example Table 2a. Listing of Audit Trail Files.

File Name	Description	Comments
Site_A_Audit.xml	Audit trail of Site A	[name of EDC system]
Site_B_Audit.xml	Audit trail of Site B	[name of EDC system]

b. Audit Trail File Contents

For each audit trail file submitted, you should provide a table that includes the variable names, variable label or description (e.g., description of coded values, unit of measure, formatting information), and other information necessary for review. See Example Table 2b. *Contents of AuditTrailFile.XML*. Each audit trail file submitted should include every original and updated value, operator identification and date and time stamps for each data entry and any change, and reason for each change. Additional columns may be added as needed.

Example Table 2b. Contents of AuditTrailFile.XML (number observations=, number of variables=).

Variable Name	Description
animalid	Animal identification number
formname	Name of the data collection form
entryfield	Name of entry field from data collection form
entrytype	Type of entry field (e.g., numeric, text, radio button, or checkbox)
entry	Data entry [include explanation if applicable (e.g., "on" = radio button was selected, "off" = radio button not selected; if blank, no data entered)]
operatorid	Individual entering data
entrydate	Date/time stamp when the data were entered
reason	Reason for change, if applicable
lockdate	Date and time when the database was locked

3. Data Analysis Programs

a. Program File Listing

In this section, you should provide a table listing the programs used to perform randomization, process the data, generate summaries, and perform the statistical analysis. This table should include the file name, the purpose of the program, the electronic data files accessed and generated by each program, a list of any results (tables/graphs) generated, and software used, if applicable. See Example Table 3. *Listing of Program Files*.

Example Table 3. Listing of Program Files.

Program File Name	Purpose	Data File(s) Used in Program (Input)	Data File(s) Created (Output)	Results or Tables/ Graphs Created	Comments
setup.xml	Sets up needed libraries and data formats. Creates all temporary data files used in the analysis	Effect_data.xpt AE_all.xpt Clinobs_all.xpt	Effect.sas7bdat AE.sas7bdat ClinObs.sas7bdat	None	Run this program first before running any analysis programs
s_clinobs.xml	Summarizes clinical observations	ClinObs.sas7bdat	None	Summary of clinical observa-tions Profile plots	Table X provided in Final Study Report
effect_prog.xml	Summarizes effectiveness variables and determine success/ failure	Effect.sas7bdat	Success.sas7bdat	None	Details of success/ failure evaluation. Table Y provided in Final Study Report.
prim_anlys.xml	Conducts primary analysis	Success.sas7bdat	None	Primary analysis results	

b. Overview of Data and Analysis Flow

You should provide a general overview of the data and analysis process used in the

study and submission. For example, you should describe whether data files were directly downloaded from data management systems and whether any processing (merging, validation, editing) were performed prior to analysis.

For electronic databases that do not store or directly export data in XPT or XML formats, conversion from another file format (e.g., XLSX, CSV, or SAS7bdat) will be needed in order to submit to CVM via eSubmitter (see footnote 2). If conversion is necessary, CVM expects that electronic data files will be converted into XPT or XML format without modification (e.g., no column changes or calculations). The Appendix has examples of SAS programs that convert data files to acceptable formats. The program used to make the conversion should be provided to CVM in XML format. If a system limitation precludes the creation of XPT or XML formatted files, you should contact CVM. Additionally, CVM encourages you to evaluate/analyze the data using the same file that is submitted to CVM. For example, if the study data are submitted to CVM in file "ALLDATA.XML," the programs submitted should use this file for analysis.

Submitted data files should be unmodified after export from the database. If a data file was modified for analysis (e.g., addition or edit of variable names, variables or animals excluded, or coded information reformatted as text), a description of the changes and the computer program used to create the changes should be included, but the modified data files should not be submitted. The logic or coding that was used in creating the modified data files should be documented in the computer program and clearly detailed in the submission so that CVM may recreate and verify the modifications.

c. Instructions for Running Programs

You should describe the sequence of program calls needed for CVM to run your programs. Starting with the first program to be run, you should describe calls to other programs and custom functions.

So that CVM can understand and verify your analysis process, for each program you should document all directories and files referenced to access or store data, including directory and file names, locations, and aliases, if used. Describe programs defining custom styles or formats or, if such styles or formats are predefined, you should provide instructions for their installation. If the programs were designed to call other programs or access data in specific folders or directory structures, you should describe this structure so that CVM can verify your analysis process.

d. Randomization Programs

Electronic data files of all randomization schedules used (e.g., animal to treatment, procedure order) should be included in the submission and listed in the README file. These files should contain all variables specified in restrictions to

randomization, for example, any stratification variables (e.g., weight, sex) or if treatments were balanced by one or more factors (e.g., room, cohort). If available, programs used to generate random assignments, including the randomization seed(s), should be included in the submission and listed in the README file.

Detailed descriptions of all randomization processes should be included in the FSR.

IV. Data and Analysis Programs: Additional Comments

Submission of a separate document (e.g., Statistical Report) as an appendix in the FSR that provides details on the statistical analysis as well as additional analysis results and summaries is acceptable. As is the case with all README file content described in this document, if the FSR contains the information regarding program files and program execution as described in section III.B.3. *Data Analysis Programs* above, then the information does not have to be repeated in the README file. Instead, the README file can refer to the corresponding section(s) of the FSR.

CVM's verification of the analysis process will be facilitated by submission of well-documented analysis programs. You should describe the general purpose of each program in the beginning of the statistical program. Within the program, you should include sufficient comments to explain complex sections, for example, if a program will call certain macros or subroutines.

You should not submit electronic files generated by the statistical analysis, such as tables of descriptive summaries or least squares means, in data format (e.g., XML or XPT). Documents containing these summaries may be included or appended to the FSR as appropriate. For example, a listing of subject success/failure outcomes should be included in the FSR, but the secondary electronic data file derived from the initially collected data should not be submitted.

Descriptions of derived variables and corresponding formulas or complex transformations should be described in the FSR or contributing scientist report and documented in the analysis program. If values were computed, derived, or transformed from other variables, the equation(s) for each variable, descriptions of calculations, and a table of the calculated values should be in the FSR. recFormulas for standard conversions (e.g., lb to kg; mg to kg) do not need to be provided.

Additionally, you should not submit outputs from running the analysis, for example, the outputs and log files produced when running SAS.

Appendix

Examples of SAS codes that convert data files to acceptable formats

1. SAS code to generate XML files

   libname in '*file location*';
   libname out xml '*file location\filename1.xml*';
   data out.*dataset1*; set in.*filename*; run;

2. SAS code to generate XPT files

   libname in '*file location*';
   libname out xport '*file location\filename1.xpt*';
   data out.*filename1*; set in.*filename*; run;