#197

Guidance for Industry

Documenting Statistical Analysis Programs and Data Files

DRAFT GUIDANCE

This guidance document is being distributed for comments only.

(This version of the guidance replaces the version that was made available in March 16, 2009. This guidance document has been revised to correct the contact information in regard to this document.)

The purpose of this document is to recommend a uniform system for documenting statistical analysis programs and data files and to provide a list of items that should be included in a data submission.

Submit written comments to the Division of Dockets Management (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 listed in the notice of availability that publishes in the *Federal Register*. Submit electronic comments to http://www.regulations.gov.

For questions regarding this guidance, contact: Anna Nevius, Center for Veterinary Medicine (HFV-105), Food and Drug Administration, 7500 Standish Pl., Rockville, MD 20855, 240-276-8170; e-mail: anna.nevius@fda.hhs.gov.

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Veterinary Medicine (CVM)
April 27, 2009

TABLE OF CONTENTS

DOCUMENTING STATISTICAL ANALYSIS PROGRAMS AND DATA FILES[1]

This draft guidance, when finalized, will represent the Food and Drug Administration's (FDA) current thinking on this topic.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, call the number listed on the title page of this guidance.

I. INTRODUCTION

This guidance is provided to inform study statisticians of recommendations for documenting statistical analyses and data files submitted to CVM for the evaluation of safety and effectiveness in new animal drug applications. Sections III – XIII of this guidance document describe our recommendations. These recommendations are intended to reduce the number of revisions that may be required for CVM to effectively review statistical analyses and to simplify submission preparation by providing a uniform documentation system. Our recommendations are intended to encompass the most complex data submissions to CVM. We understand that not all submissions are of this complexity; you should refer only to those portions of this guidance applicable to your particular submission.

FDA's guidance documents, including this guidance, 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

An adequately documented statistical analysis should include submitted data that is readable, describe how the data is processed, and describe the statistical analyses employed to support your conclusions.

Our main concern is that you provide the information we need to understand and access your analyses. The complexity of the documentation will depend to a certain extent upon the programming style you employ. For example, if you provide a single program that specifies locations of data libraries, sets file locations for output and log listings, and calls programs in

[1] This draft guidance document was prepared by the Office of New Animal Drug Evaluation in the Center for Veterinary Medicine, FDA.

the appropriate logical sequence, your documentation may refer to that single program. If, on the other hand, you employ multiple programs, your documentation should detail which programs specify locations of data libraries and which programs specify file locations for output and log listings. In addition, if you employ multiple programs, your document should detail the sequence in which the programs need to be called. Similarly, if you provide a single transcribed data file which contains merged data files describing treatment, inclusion data, and dosing data for each animal, you will not need to document how those merges were performed.

Your documentation should clearly describe the entire process by which the analysis was generated, from the transcribed data files, which contain data transcribed from case report forms for each individual animal in the study, to the final statistical analyses, which form the basis for your study's conclusions. For example, to analyze an effectiveness study, you may begin with transcribed data files, which record for each animal: (i) acceptance according to study inclusion criteria; (ii) subsequent randomization into treatment; (iii) administration of specified treatment; (iv) retention of eligibility during the study; and (v) treatment success or failure at study conclusion. Before any of the data for that animal can be analyzed, the entire array of data from case report forms should be compiled into one or more modified data files for analysis. To understand the process by which you compiled the data and conducted the statistical analysis, CVM needs to understand the contents of each data file, the computer programs that compile all the data files together into data files for analysis, and the computer programs that calculate necessary statistics on these analysis data files.

Except for comments within programs, you should provide your documentation in a separate 'readme' file.

III. INTRODUCTORY MATERIAL

A. Study Number/ Drug/ Sponsor Point of Contact for Statistical Issues

Your documentation should first include the study name and number, the new animal drug being investigated, and the name of a person CVM should contact to discuss any questions about the documentation.

B. Directory and Folder Structure

You should explain where all directories and folders containing data or programs may be found in your submission and briefly describe the purpose of each directory and folder.

C. Software Employed to Generate, Read, and Analyze Data Files

You should provide the name and version number of any software we need to generate, read, or analyze your data files.

Data files should be provided on media readable on a personal computer. For example we are able to accept data in SAS System XPORT transport format[2] (Version 5 SAS transport format), an open format published by the SAS Institute. If you wish to provide data files or output which can neither be read by the operating system nor read as SAS XPORT transport files, you should contact CVM.

D. Overview of Data and Analysis Flow

1. Sequence of Changes in Data Files

You should provide a step-by-step overview of modifications in transcribed data files made to generate the final analysis data files including, for example, merges of data files, data exclusions, and data transformations, and name the program used to generate each modification.

2. Program Sequence

You should provide a general overview of the program flow used in your submission. Starting with the first program called, you should document all calls to other programs, custom functions, and macros, in sequence, including both interactive calls and calls made within programs.

3. Table and Graph Mappings

You should name the program used to generate each analysis, table, and graph provided in your submission.

IV. DATA FILES AND VARIABLES

The data files directly employed in statistical analyses may have been derived from one or more transcribed data files. You should document all such derived data files external to the programming code. You should include the name and purpose of each derived data file and describe each field (column) in each derived data file, including the field's name, purpose, and, where applicable, labels as well as units of measurement. You should include definitions of codes used for categorical variables and, if categorical variables in data files are not formatted

[2] The description of this format is in the public domain, and data can be translated to and from this format to other commonly used formats without the use of programs from SAS Institute or any specific vendor. Information concerning the SAS System XPORT transport format may be currently found on the internet at http://www.sas.com/govedu/fda/faq.html.

by the programs provided, you should provide instructions for including the formats so that data files can be read.

You should explain any alterations of data on individual subjects, e.g., corrections or changes in inclusion status or measured data, and describe any data manipulations used to create the data files such as merges, calculation of derived variables, recoding of categorical variables, and data transformations.

A. Transcribed and Modified Data Files

1. Name and Purpose of Each Data File

You should describe the purpose of each data file.

2. Variables Definitions

External documentation for each field in each data file should include the field name, definition, and, where applicable, labels, formats, and units of measurement.

B. Modifications of Transcribed Data

You should describe, and explain the reason for, any modifications of data on individual subjects, e.g., corrections or changes in inclusion status, changes in data values, calculation of derived variables, recoding of categorical variables, and data transformations.

V. INSTRUCTIONS FOR RUNNING PROGRAMS

A. How to Set up Pathnames, Libraries, Custom Styles and Formats

For each program, you should document all directories and files referenced to access or store data, including directory and file names, locations, and aliases. You should also document the location of programs defining custom styles or formats or, if such styles or formats are predefined, you should provide instructions for their installation.

B. Required Program Sequence

You should describe the sequence of program calls needed for CVM to run your programs.

C. Diversion and Resetting of Standard Output

You should name any programs that divert log files and analysis output away from default, e.g., from windows to stored files, and name any programs that divert such listings back to default. When diversion is to stored files, you should specify the name of each file and the directory in which it will be stored.

D. Software Termination Statements in Programs

You should describe any statements employed in your programs that terminate software sessions, including the program name and where in each program they occur.

VI. *PROGRAM COMMENTS*

You should describe the general purpose of the program at the beginning. Within the program, you should include sufficient comments to explain complex sections.

VII. *RANDOMIZATION*

A. Allocation Tables

You should provide the allocation tables used for any randomization, e.g., experimental units to treatments, order of necropsy.

B. Process Used to Generate Allocation Tables

You should describe the process used to generate your allocation tables, including any programs used.

VIII. *LOG AND OUTPUT FILES*

In your submission, you should provide in stored files all log and output listings generated by your programs.