Guideline on the Submission of Clinical Trial Data

(Draft for Public Review)

Center for Drug Evaluation, NMPA
May, 2020

Table of Contents

Guideline on the Submission of Clinical Trial Data

1. Background and Purposes

Clinical trial data is one of the important materials submitted by sponsor to regulatory agencies. It is a valuable resource for both regulatory agencies and sponsor. Standardized collection, organization, analysis, and presentation of clinical trial data play an important role in improving the efficiency and quality of clinical researches and development, shortening review timelines. Also, it is beneficial to the management of entire life cycle of drug development, and further promoting information exchange and sharing between drug development units and regulatory agencies.

If the clinical trial data submitted by sponsor does not follow certain specifications, it will take significant resources for reviewers to get familiarized with and understand the data structure and content. In some cases, sponsor or regulatory agency may need to conduct pooled analyses using multi-sources of clinical trial data. Non-standardized data will make this task almost impossible.

Submission of clinical trial data is usually a package that includes database and its supportive documents, like data definition file, data reviewer's guide, data derivation and analysis programs, and annotated case report forms (aCRFs). This document provides specific requirements for the content and format of clinical trial data submission package, and aims to guide sponsors in the submission of clinical trial data and related materials, and to help professionals such as data managers and statistical programmers to better conduct related tasks for clinical trials.

This guideline is formulated based on the data submission requirements of international regulatory agencies and the current situation in China. Sponsor should prepare the package based on the requirements in this guidance document. Sponsor is encouraged to submit clinical trial data and the associated materials according to Clinical Data Interchange

Standards Consortium (CDISC) standards. With the development and improvement of the understanding and practice of clinical trial data standards, this guidance document will be revised and improved as appropriate.

2. Submission Components of Clinical Trial Data

2.1 Study database

Study database should generally contain source data collected directly from case report forms (CRFs) and external sources as well. It may also contain few derived varialbes such as serial numbers, study days, etc. However, missing data in raw database should not be imputed. To meet the requirements for regulatory submission, collected data may require necessary standardization or coding (e.g., adjusting the name/label of the dataset in the database and name/label of the variable in the dataset, encoding variable values with standard dictionary (e.g. Medical Dictionary for Regulatory Activities (MedDRA)) where applicable.

Study database typically contains multiple raw datasets, which should be organized and named according to its contents. Study dataset is usually named as a two lettercode, for example, demographics dataset (dm), adverse event dataset (ae), laboratory test dataset (lb), etc. Refer to Appendix 1 for details of nomenclature of raw datasets commonly used in clinical trial data submission.

Datasets that contain observed results of subjects must have study identifier, subject unique identifier, and other identifiers to uniquely identify an observation must be included in a dataset (e.g., dm, ae, lb, etc. in Appendix 1). Subject identifier (SUBJID) must be included in dm dataset. Commonly used identifiers are exemplified as follows:

Study identifier: The variable name is STUDYID, character type, unique identifier of the study, usually also regarded as study number.

Subject unique identifier: The variable name is USUBJID, character type. Each subject should be assigned a unique identifier throughout a submission (may include multiple clinical studies). In all datasets (including the raw and analysis datasets), the same subject should have exactly the same unique identifier. When a subject participates in multiple studies, the USUBJID should be consistent across these studies. Following this rule is particularly important for merging datasets from different studies on the same subject (e.g., randomized controlled study and its extension study).

Subject identifier: The variable name is SUBJID, character type. SUBJID is the identifier of a subject enrolled in a trial. If one subject is screened multiple times in a trial, the SUBJID should be different each time.

Time variables such as visit name (VISIT, character type) and visit number (VISITNUM, numerical type) should be included in applicable datasets. VISITNUM should be assigned values in ascending chronological order.

2.2 Analysis database

Analysis database is a database derived for statistical analyses and is used to produce and support statistical analysis results in clinical study report. Analysis database can contain raw data, and also derived variables according to specified rules, such as imputation for missing values.

Analysis database typically includes multiple analysis datasets. Derived and collected data (from raw datasets or other analysis datasets) may be combined into a single dataset when building an analysis dataset. When creating analysis dataset, the following principles should be followed: 1) Analysis dataset must be built up to clearly support the statistical analyses planned for the clinical study. 2) Analysis data set must be traceable; and the specified rules for derived variables should be detailed in the corresponding data definition file. 3) The structure and contents of analysis dataset should facilitate statistical analysis with limited programming efforts, namely analysis ready.

Analysis database should contain all variables required for the planned/intended analysis, including derived variables; and all derived variables should be able to be generated from the study database. Analysis datasets are usually named in the form of "adxxxxxx" and the name should be also consistent with the corresponding raw dataset's name, such as adcm, adae, adlb, etc.

The subject level analysis dataset is mandatory (named as adsl) for a submission data package. In this dataset, each subject should have one record that includes, but not limited to, demographics, disease factors, treatment groups, other prognostic factors that may affect treatment response, dates of important events, and population flags.

For some endpoints (e.g., scale scores), a series of derivation processes are needed to get it ready for the final statistical analyses. The intermediate variables/datasets derived to facilitate the creation of the final analysis dataset should also be included in the analysis database for submission, if necessary.

2.3 Data Definition File

Raw and analysis databases submitted must have appropriate data definition file. Data definition file is used to describe the submitted data, and should at least contain the name, label and basic structure of each dataset in the submitted database, and the name, label and type of each variable and derivation process of each derived variable in each dataset.

Data definition file is one of the most important documents for regulatory agencies to accurately understand the submission data. Sponsor should ensure that the code list and sources of each variable are clearly defined and easily searchable. If external dictionaries are used, sponsor needs to specify the dictionary and its version in data definition file. Good traceability between data (e.g., between raw data and CRF, analysis data and raw data) needs to be documented in the file to facilitate regulatory review. Sponsor need to provide details in data definition file, particularly with regard to derived variables. Program codes may need to be provided, if necessary, to assist with review.

Data reviewer's guide is a supplement to data definition file for raw/analysis database, which will help reviewers better understand and use submitted data, so it should be submitted if necessary. Data reviewer's guide provides information in addition to what are presented in data definition file, including but not limited to, instructions on the use of the submitted data, relationships between the study report and the data, certain key information of study documents (e.g. trial protocol, statistical analysis plan, clinical study report), and description/explanation of other special scenarios. Data reviewer's guides are not intended to replace data definition file, but to help reviewers more accurately and efficiently understand and use the submitted database, relevant terminologies, and data definition file.

Data definition file is generally in extensible mark-up language format (XML) or portable document format (PDF) format. Data reviewer's guide should be submitted in PDF.

2.4 Annotated CRF

Annotated CRF is blank CRF with annotations that illustrates the mapping relationship between data units (i.e. field) of collected subject data (electronic or paper) and variables/variable values in submitted study database. Annotated CRF should be submitted in PDF.

In practice, some data fields may be collected on the CRF but not included insubmitted datasets. These data fields should be clearly marked as "NOT SUBMITTED" on the aCRF and reason(s) for not submitting these data should be clarified in data reviewer's guide accordingly.

2.5 Programming Code

Sponsors should submit programming codes, which include, but notlimited to, the derivation process of analysis datasets, generation process of analysis results for the primary and secondary efficacy endpoints, etc. Programming codes submitted in

submission package.should be readable (with comments), understandable, executable, and do not include external program calls, which in particular avoid using large macro programs. Programming codes in submission packages are generally in TXT format.

3. Submission Document Format and Conventions

3.1 Portable document format

Portable Document Format is an open document format that is independent of application software, hardware, and operating system. Any other documents in submission package that follow the requirements of the International Counsel for Harmonization (ICH) Electronic Common Technical Document (eCTD) format can be in PDF format. It is recommended that PDF version1.4 and above to be used in submission. All PDF files should use .pdf as the file extension.

3.2 Extensible mark-up language format

Extensible Mark-up Language is a type of data exchange languages, which is defined by the World Wide Web Consortium (W3C). It can be opened, edited and created by any text editor, and used to transfer and store data. Files in XML format can conveniently exchange information between different systems. All XML files are required to use .xml as the file extension.

3.3 Plain text format

Plain Text Format document (TXT) has characteristics such as simple format, small file size, simple and convenient for storage. It is also a common file format supported by computers and many mobile devices. All TXT files should use .txt as the file extension.

3.4 Data transport file format

Datasets in submission package are usually in transport file format (XPT). One XPT file corresponds to one dataset. XPT file name needs to be consistent with the corresponding dataset name. XPT files should use .xpt as the file extension, for example, ae.xpt for Adverse Event (AE), cm.xpt for Concomitant Medication (CM). SAS Transport File Format version 5 (referred to as XPT V5) or above is recommended as the data submission format. Sponsor should ensure that submitted datasets are free from illegible contents in different operating environments.

3.5 Dataset split

When a dataset in database needs to be split because the file size does not meet submission requirements, detailed rules of splitting and detailed steps of merging it back should be specified in data reviewer's guide to ensure that reviewer can generate the dataset same as what is prior to splitting.

3.6 Dataset name, variable name and length

Specific requirements about the name and length of dataset and variable are as follows:
Dataset name can only contain lowercase letters and numbers and must start with a lowercase letter. The maximum length of a dataset name is 8 bytes.

The variable name can only contain upper case characters and numbers, and must start with a letter. The maximum length of a variable name is 8 bytes.

The length of each character variable should be set to the maximum actual value length of the variable across all datasets of the same study, to effectively control the size of the file. Variable length should be set not to exceed 200 bytes; variable splitting may be needed. When splitting, bytes cannot be truncated, and efforts should be made to maintain the integrity of statement in each splitting variable.

3.7 Dataset labels and variable labels

For ease of review, dataset labels and variable labels should be in Chinese and should not exceed 40 bytes in length. If necessary, labels can contain English letters, underlines, or numbers, but cannot start with numbers. In addition, labels cannot include the following cases:

- Unpaired half-width/full-width single/double quotation marks
- Unpaired half-width or full-width brackets
- Special characters

4. Other Considerations

4.1 Traceability of trial data

An important part of regulatory review is accurate understanding of the source of data, that is, the traceability of data. Traceability enables reviewers to understand the relationship between statistical analysis results (table, listing and figures in study report), analysis data, and raw data.

The traceability of data ensures that reviewers are able to accurately:

- understand the construction of analysis datasets
- identify records used for derived variables and the corresponding algorithms
- understand the algorithm/model of corresponding statistical results
- establish technique used to link raw data to corresponding table(s)

When submitting study database, sponsor should ensure that regulatory reviewers can use the study database to derive the analysis database that is consistent with what the sponsor submitted, and that analysis database can directly reproduce statistical analysis results that

are also consistent with what the sponsor submitted. Traceability can be supplemented by providing a detailed data flowchart from the collectionto the submission.

4.2 Data files under eCTD

When it comes to registration using eCTD, all documents, trial data and associated supportive documents should be organized according to the specified folder structure. All submitted files should be in the correct folder and tagged using the appropriate Study Tagging File (STF). Refer to Appendix 2 and Appendix 3 for more information regarding STF and folder structure.

4.3 Foreign language database

When it comes to registration using foreign language database, dataset label, variable label, adverse events terms, generic name of concomitant medications, medical history, and names of the clinical endpoints in normalized datasets (corresponding to variables in the horizontal structured datasets) should be in Chinese. CRFs, aCRFs, data definition files, and data reviewer's guides should also be submitted in Chinese. Chinese translations in database should be consistent with all other documents in submission package.

4.4 Communication with regulatory agency

Based on specific characteristics and complexity of clinical trial data, sponsor may, if necessary, communicate with regulatory authority at Pre-NDA meetings regarding the clinical trial database and relevant materials to facilitate timely and accurate understanding of the clinical trial data submitted by the sponsor.

References

1. CFDA: Technical Guide for Data Management in Clinical Trials, July 2016
2. FDA: Study Data Technical Conformance Guide, Oct 2019

3. PMDA: Revision of Technical Conformance Guide on Electronic Study Data Submissions, Jan 2019

4. CDISC: Study Data Tabulation Model Implementation Guide, Nov 2018

5. CDISC: Analysis Data Model Implementation Guide, Oct 2019

Appendix 1: Commonly Used Raw Datasets

Table 1 Common raw data sets and nomenclature

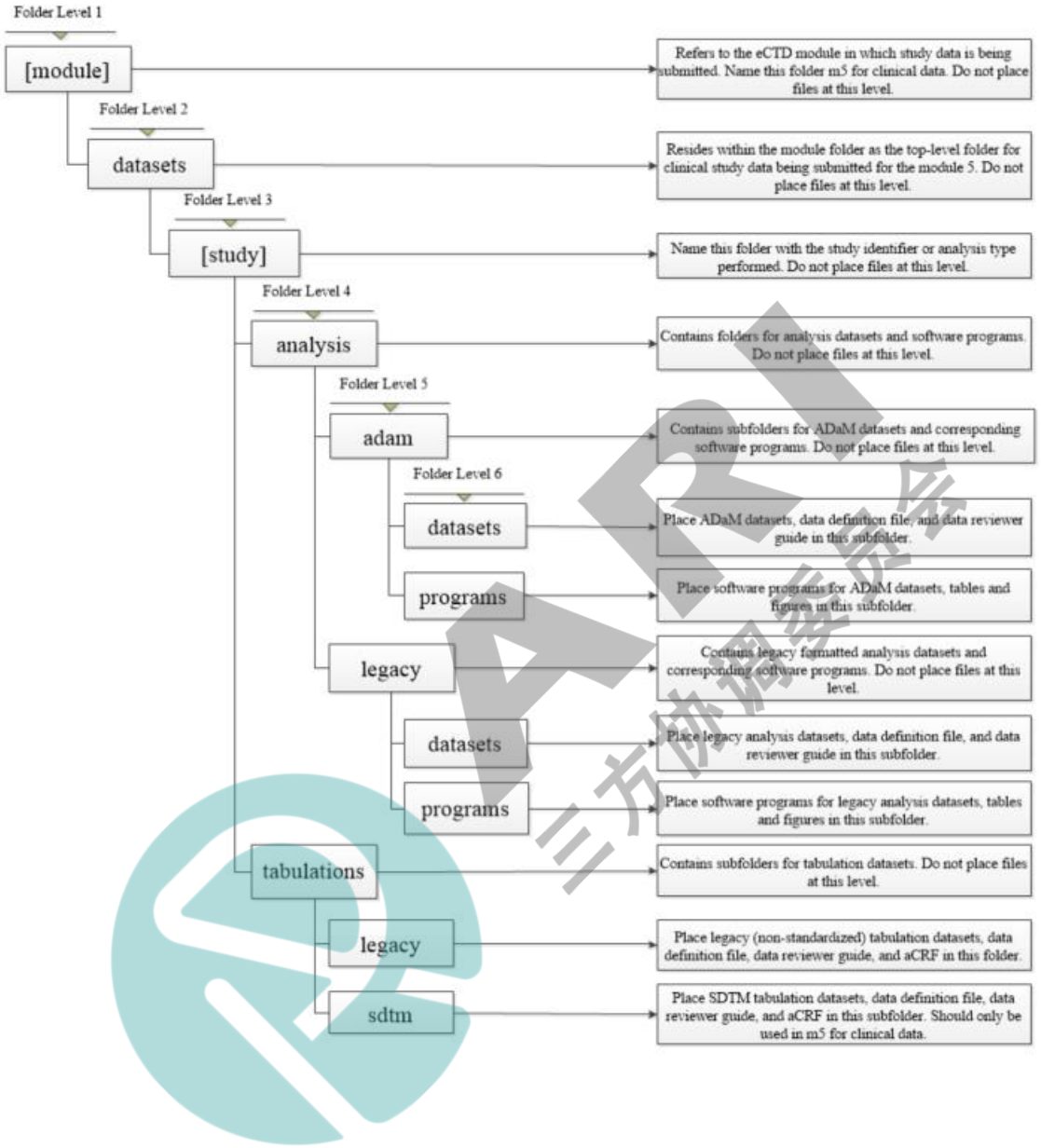
Datasets	Naming	Submission Requirements
Demography	dm	Must be submitted
Medical History	mh	If applicable
Adverse Events	ae	If applicable
Prior and Concomitant Medications	cm	If applicable
Exposure	ex	If applicable
Subject Disposition	ds	If applicable
Questionnaire	qs	If applicable
Protocol Violation	dv	If applicable
Laboratory Tests	lb	If applicable
ECG	eg	If applicable
Vital Signs	vs	If applicable
Clinical Events	ce	If applicable
Physical Examination	pe	If applicable
Disease Response	rs	If applicable

283 **Appendix 2: STF**

Name attribute values for the file-tag element	Description
data-tabulation-dataset-legacy	Study database (non-CDISC standard)
data-tabulation-dataset-sdtm	Study database (CDISC standard)
data-tabulation-data-definition	Study database data define file and data reviewer's guide
analysis-dataset-adam	Analysis database (CDISC standard)
analysis-dataset-legacy	Analysis database (non-CDISC standard)
analysis-data-definition	Analysis database data define file and data reviewer's guide
annotated-crf	Annotated CRF
analysis-program	data derivation and analysis programs

284

Appendix 3: Folder structure



Appendix 4: Glossary

Code List:

Code list for a variable is a list of allowable values that this variable may have. It includes standard codes, industry commonly used codes, and sponsor custom-defined codes.

Case Report Form (CRF) :

A printed, optical, or electronic document designed to record all of the protocol required information to be reported to the sponsor on each trial subject.

Electronic Common Technical Document (eCTD) :

Electronic registration documents submitted for drug registration and review. Organize, transmit, and present the CTD-compliant drug submissions electronically in extensible mark-up language format.

Data Definition File:

Data definition file is used to describe the submitted data, andshould at least contain the name, label and basic structure of each dataset in the submitted database, and the name, label and type of each variable and derivation process of each derived variable in each dataset.

Data Reviewer's Guide:

Data reviewer’s guide is a supplement to data definition file. It includes, but not limited to, instructions on the use of the submitted data, relationships between the study report and the data, certain key information of study documents, and description/explanation of other special scenarios.

Annotated Case Report Form (aCRF):

Annotated CRF is blank CRF with annotations that illustrates the mapping relationship between data units (i.e. field) of collected subject data (electronic or paper) and variables/variable values in submitted study database.