Guideline for Clinical Trial Data Submission

(Trial Version)

English Translation by: Hualong Sun, Zhenglong Tian and Zhijun Wei, Richard Hutchison

Disclaimer: The English is for information only and not an official translation and under any dispute the Chinese will prevail.

Center for Drug Evaluation, NMPA

July 2020

Table of Contents

Guideline for Clinical Trial Data Submission

1. Background and Purpose

Clinical trial data are an important component of submissions to regulatory authorities by sponsors and are also a valuable asset for both regulatory authorities and sponsors. It is crucial to standardize the collection, organization, analysis, and presentation of clinical trial data, not only to improve the efficiency and quality of clinical research, but to also speed up the regulatory review process. Furthermore, it is also beneficial for the life-cycle management of drug development, and to promote the exchange and sharing of information in drug development and regulatory activities.

If clinical trial data submitted by sponsors do not follow certain standards, it will take up significant resources for reviewers to get familiar with the data structures and content. In some cases, sponsors or regulatory authorities may need to conduct integrated data analyses from multi-source clinical trials, but non-standardized data will make this task almost impossible.

The submission package associated with clinical trial data usually includes the raw and analysis databases and corresponding data definition files, data reviewers' guides, programming code, and annotated Case Report Form (aCRF). These guidelines set specific requirements for the content and format of submitted clinical trial data to the Center for Drug Evaluation, and aim to give guidance to sponsors to submit data and relevant materials in a standardized approach. It will also help relevant practitioners, such as data managers and statistical programmers, to carry out relevant activities in clinical trials more effectively.

These guidelines are mainly applicable to pivotal clinical trials for the purpose of drug registration, and can be referred to for non-registration trials as well. These guidelines have been established based on the data submission requirements of international regulatory authorities and also the current domestic status quo, which should be followed by sponsors

when preparing corresponding submission packages. Sponsors are encouraged to submit clinical trial data and the associated materials according to the standards developed by the Clinical Data Interchange Standards Consortium (CDISC). It is recognized that developments in clinical trial data standards, awareness and implementation experience may advance rapidly, and therefore, this guideline will be revised as necessary.

2. Submission Components and Definitions

2.1 Study raw database

The study raw database generally contains source data collected directly from the Case Report Form (CRF) and external data sources. It may also contain derived variables such as sequential numbers. However, missing data should not be imputed in the study raw database. To meet submission requirements, collected data may require necessary standardization or coding. For example, mapping dataset names and labels, manipulating dataset structures, mapping variable names and labels, and encoding variable values with the values from standard dictionaries (e.g. Medical Dictionary for Regulatory Activities (MedDRA), WHO Drug Dictionary) where applicable. Study Data Tabulation Model (SDTM)-compliant datasets will be considered the study raw database, if the sponsor submits data according to CDISC standards. It is recommended the sponsor does so.

The study raw database typically contains multiple source raw datasets, which should be organized and named according to the contents within. Study datasets are usually named as a code with two English letters, for example, demographics (dm), adverse event (ae), laboratory test (lb), etc. Please refer to Appendix 1 for details of nomenclature of study raw datasets commonly used in clinical trial data submission.

All submitted datasets should include Study Identifier (STUDYID). Datasets that contain observed results of subjects, such as dm, ae, lb, etc. in Appendix 1, must have Unique Subject Identifier (USUBJID). Subject Identifier (SUBJID) must also be included in the dm dataset. Commonly used identifiers are exemplified as follows:

Study Identifier: Variable name is STUDYID. It is a character variable, and the unique identifier of the study, namely protocol number.

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

Subject Identifier: Variable name is SUBJID. It is a character variable. SUBJID is the identifier of a subject enrolled in a trial.

Time variables, such as Visit Name (VISIT, character type) and Visit Number (VISITNUM, numerical type), should be included in applicable datasets. VISITNUM of scheduled visits should be the assigned values in ascending chronological order and correspond to each VISIT.

2.2 Analysis database

The analysis database is a database derived for statistical analyses and used to produce and support statistical analysis results in clinical study report. Analysis database contains study raw data, as well as data derived from the study raw data following certain specified rules, such as imputation for missing values. Analysis Data Model (ADaM)-compliant datasets will be considered as the analysis database, if the sponsor submits data according to CDISC standards. It is recommended the sponsor does so.

The analysis database typically includes multiple analysis datasets. Derived and collected data (from study raw datasets or other analysis datasets) may be combined into a single dataset when building an analysis dataset. When creating analysis datasets, the following

principles should be followed: 1) Analysis datasets, used to support statistical analyses, must have clear contents and sources. 2) Analysis datasets must be traceable, and the specific rules for derived variables should be detailed in the corresponding data definition file. 3) The structure and contents of the analysis datasets should facilitate statistical analysis with limited programming effort.

The analysis database should contain all variables required for the planned/intended analysis, including derived variables. All derived variables should be able to be generated from the study raw database and other supportive data documents. Analysis datasets are usually named using the convention "adxxxxxx", and the name should be consistent with the corresponding raw dataset, such as adcm, adae, adlb, etc.

The Subject Level Analysis Dataset is mandatory (named as adsl) for a submission package. In this dataset, there should be only one record per each subject, which should include demographics, important baseline features/stratification factors, treatment groups, prognostic factors, dates of important events, analysis population flags and other information as required.

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

2.3 Data Definition Files

The submitted study raw and analysis datasets need to have corresponding data definition file. Data definition file is the document used to describe submitted datasets and should at least include the name, label, description of datasets' basic structure and the name, label, type, origin/derivation process of the variables in the datasets.

Data definition file is one of the most critical documents which help regulatory authority comprehend the content of the submitted datasets accurately. Sponsors need to ensure that

the definition of the code list and the origin of all variables are clear and easy to be searched. If using an external dictionary, sponsors must specify the dictionary and the version they used in the data definition file.

Sponsors must provide the details, especially the description of the derived variables, in the data definition file. Moreover, program code snippets might be included for further clarification and better understanding if necessary. It is essential to establish a good traceability between data through the data definition file (e.g. the relationship between study raw dataset and CRF; the relationship between analysis and study raw datasets), which would facilitate the regulatory review.

In General, data definition file is in Extensible Mark-up Language, XML or a Portable Document Format, PDF format. The Extensible Stylesheet Language, XSL, must be submitted as well when data definition file in XML format is submitted.

2.4 Data Reviewer's Guide

To help reviewers better comprehend and utilize the submitted data, it is highly recommended that sponsors submit a data reviewer's guide for the study raw database and a separate reviewer's guide for the analysis database. The data reviewer's guides provides a supplement to the data definition file and the content includes but is not limited to the circumstances as below: user instruction of the clinical data, the relationship between clinical study reports and data, the critical information of the study documents (e.g. protocol, statistical analysis plan, clinical study report), instructions for submitted program code and encoding of the datasets (e.g. urf-8, euc-cn, etc.). The data reviewer's guides are not intended to replace the data definition files, but to facilitate the regulatory authority to comprehend and utilize the submitted datasets, terminology, program code, data definition files, etc., with accuracy and efficiency. The data reviewer's guides should be generated in a PDF file.

2.5 Annotated Case Report Form

An aCRF is based on a blank CRF with annotations that illustrate the mapping between data units (i.e. fields) which are collected at the subject-level (from electronic or paper CRF) and variables/variable values in the submitted datasets. The aCRF should be provided as a PDF with the file name ‘acrf.pdf.’

In practice, data not recorded in the submitted datasets may be collected through CRF, which should be annotated as “NOT SUBMITTED” in the aCRF with the corresponding reason clarified in the data reviewer’s guide.

2.6 Program code

Program codes which need to be submitted by sponsors includes, but not limited to the derivation of the derived variables in analysis dataset, the generation of analysis result for efficacy endpoints etc. Program codes should be understandable and readable. It is highly recommended that sponsors must provide adequate comments and avoid calls from external program and macros. In general, program codes are submitted with file in TXT format.

3. Clinical Trial Data Relative Document Format

3.1 Portable document format

Portable Document Format (PDF) is an open document format in a manner independent of application software, hardware, and operating systems. By following the requirements of the International Council for Harmonization (ICH) Electronic Common Technical Document (eCTD), other documents can be submitted in PDF format in submission package. PDF version1.4 or above is recommended for document submission. All PDF files should use .pdf as file extension.

3.2 Extensible mark-up language format

Extensible Mark-up Language (XML) is a kind of data exchange language 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. Data can be exchanged between different systems in the files with XML format. All XML files should use .xml as file extension.

3.3 Plain text format

Plain Text Format document (TXT) have simple format, small size, and are easy to store. TXT is also a common format supported by computers and mobile terminals. All TXT files should use .txt as file extension.

3.4 Data transport file format

Submitted datasets should be in SAS Transport Format (XPT). One XPT file corresponds to one dataset and the XPT file name must be consistent with the dataset name with. xpt as the file extension. For example, ae.xpt for adverse events, cm.xpt for concomitant medications, etc. SAS Transport File Format version 5 (referred to XPT V5) or above is recommended as the data submission format. Sponsor should clarify the encoding (e.g. utf-8, euc-cn, etc.) of submitted datasets to avoid transcoding issues.

3.5 Dataset splitting

Split datasets can be submitted when a dataset in database must be split because the file size does not meet the submission requirements. It is allowed to submit only the split datasets. The instruction of splitting and merging datasets back in detail should be specified in the data reviewer's guide(s) to ensure that reviewers can generate the original dataset (before splitting).

3.6 Dataset name, variable name and length

Specific requirements about the naming of dataset and variable are as the following:

Dataset name can only contain English alphabets in lower case and numbers. Besides, the prefix must be an English alphabet in lower case. The maximum length of a dataset name is 8 characters.

The variable name can only contain English alphabets in upper case, underscore and numbers. In addition, the prefix must be an English alphabet. The maximum length of a variable name is 8 characters.

To maintain size of datasets, the length of each character variable should be assigned with the maximum length of actual value and aligned between all datasets which contain the same variable.

3.7 Dataset labels and variable labels

To facilitate the review of datasets, dataset labels and variable labels should use Simplified Chinese and should not exceed 40 bytes when possible. Labels can contain English alphabets, underscores, or numbers, but cannot use a number as prefix. In addition, labels cannot include the following situations:

- Unpaired halfwidth/fullwidth single/double quotation marks
- Unpaired halfwidth or fullwidth brackets
- Special characters (e.g. ‘>’, ‘<’)

4. Other Considerations

4.1 Traceability of trial data

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

When submitting study raw database, sponsor should ensure that regulatory authorities can use them 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 consistent with what the sponsor submitted. Traceability can be supplemented by providing a detailed data flowchart from the collection phase to the submission phase.

4.2 Electronic submission format under eCTD

Study datasets and their supplemental files should be organized into a specific file directory structure when submitted in the eCTD format. 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

The submission of clinical trial data and associated documents should primarily be in Simplified Chinese, and the translation of a foreign language to Simplified Chinese should be consistent, for example, the names of adverse events in the analysis dataset and the names of adverse events in the clinical summary report should be the same. In order to

ensure the efficiency of review, the minimum requirements for the submission package related to clinical trial data that need to be translated into Simplified Chinese are as follows:

- The following in the submitted databases should be in Simplified Chinese: dataset labels and variable labels; adverse event terms, concomitant medication terms, and medical history terms that appear in submitted reports, such as the Clinical Study Report (CSR);
- The following in the data definition files should at least be in Simplified Chinese: descriptions/labels of the datasets; descriptions/labels and the derivation process of variables in the dataset; code lists/code values of efficacy variables;
- The following in the annotated Case Report Form (aCRF) should at least be in Simplified Chinese: descriptions of the leading questions to collect data; code lists/code values related to efficacy endpoints;
- Reviewer's Guides should be in Simplified Chinese.

4.4 Communication with regulatory authority

Based on the characteristics and complexity of the specific clinical trial data, the sponsor may, if necessary, communicate with regulatory authority at Pre-NDA meetings regarding the clinical trial database and relevant submission materials in accordance with the management of drug development and technical review communication, so as to facilitate the reviewers to quickly and accurately understand 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, Mar 2020
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 Study raw Datasets

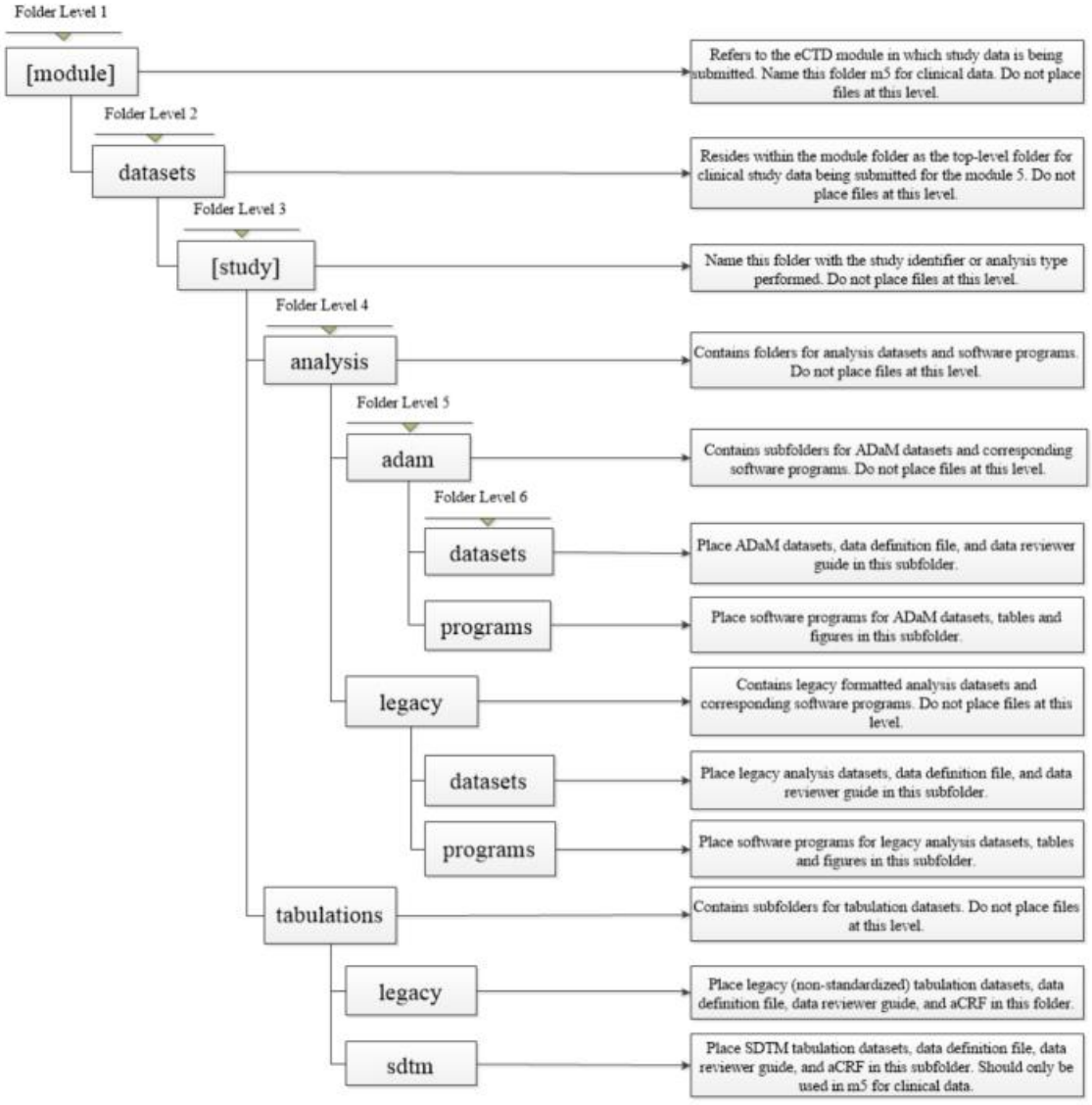
Table 1 Common study raw data sets and naming convention

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 Deviation	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

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's 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

Appendix 3: Folder structure



Appendix 4: Glossary

Code List:

A code list includes the possible values of variables and the corresponding standard codes, industry commonly used codes, or customized codes by the sponsor.

Case Report Form (CRF):

A printed, 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):

The eCTD is electronic registration document 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:

The 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 Reviewer's Guide:

The data reviewer’s guide provides a supplement to the data description file. It helps the regulatory authority comprehend and utilize the submitted datasets, relevant terminologies, program codes and the data definition file with accuracy and efficiency.

Annotated Case Report Form (aCRF):

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