Analysis Data Model: Version 2.0

Prepared by the
CDISC Analysis Dataset Modeling Team
(ADaM)

Notes to Readers

- This Model incorporates aspects of the previous General Considerations document version 1.0 and Subject Level Characteristics Model version 0.2, replacing these 2 documents. In order to avoid confusion with earlier documents, the version of this new document will begin with 2.0.
- A library of on-line satellite documents will be created as additional statistical analysis models are developed or other analysis topics and concepts are addressed. These documents will be developed using the concepts and standards presented here. A guide listing all of the documents and providing links to them will be maintained.

Revision History

Date	Version	Description
2/15/2006	v 2.0	Reformatted from General Considerations v1.0, incorporating Subject-level model, emphasizing requirements and naming and content rules and guidelines.
5/31/2006	V2.0	Incorporate comments from public review
8/11/2006	V2.0	Final document

Note: Please see Appendix 8.8 for Representations and Warranties; Limitations of Liability, and Disclaimers.

Contents

1. INTRODUCTION / PURPOSE

The objective of the CDISC Analysis Dataset Modeling Team (ADaM) is to provide metadata models and examples of analysis datasets used to generate the statistical results for a regulatory submission. The ADaM models and examples will build on Study Data Tabulation Model (SDTM) metadata models, adding attributes and examples specific to statistical analysis. Throughout its work, the ADaM team acknowledges that clinical trials are unique and that the design of analysis datasets is driven by the scientific and medical objectives of the study.

The Analysis Data Model describes the general structure, metadata, and content typically found in Analysis Datasets and accompanying documentation. The three types of metadata associated with analysis datasets (analysis dataset metadata, analysis variable metadata, and analysis results metadata) are described and examples provided.

The statistical analysis data models (in both this document and in satellite documents) presented by the ADaM team represent consensus across a large number of statistical professionals experienced in regulatory submissions. It is recognized that these models represent only one approach and other data set structures may be more appropriate for a specific indication, study design or analysis. In the end, the structure and content of the analysis datasets should be designed to provide clear, unambiguous communication of the science and statistics of the trial. We cannot emphasize enough that early and effective cross-communication between regulatory reviewers and sponsors is requisite for mutual success.

The descriptions in this document build on the nomenclature of the SDTM V3.1.1 standard, adding attributes, variables, and data structures required as appropriate for statistical analyses.

Analysis Datasets will be discussed in the context of their relationship to other Clinical Data Interchange Standards Consortium (CDISC) and Food and Drug Administration (FDA) initiatives. A set of statistical analysis variables that are useful in most Analysis Data models will be presented.

The concept of analysis results metadata will be introduced. This metadata describes the attributes of each important analysis and refers to the Analysis Datasets used by each analysis. Finally, a model for a subject-level analysis dataset will be presented. This dataset is the minimum requirement for analysis datasets, assuming any are produced.

This document provides the core of the ADaM concepts and standards. A library of on-line satellite documents [3] will be created as additional statistical analysis models are developed or other analysis topics and concepts, such as the creation of define.XML for analysis datasets, are addressed. The general concepts, descriptions and definitions in this document will be used to build these documents.

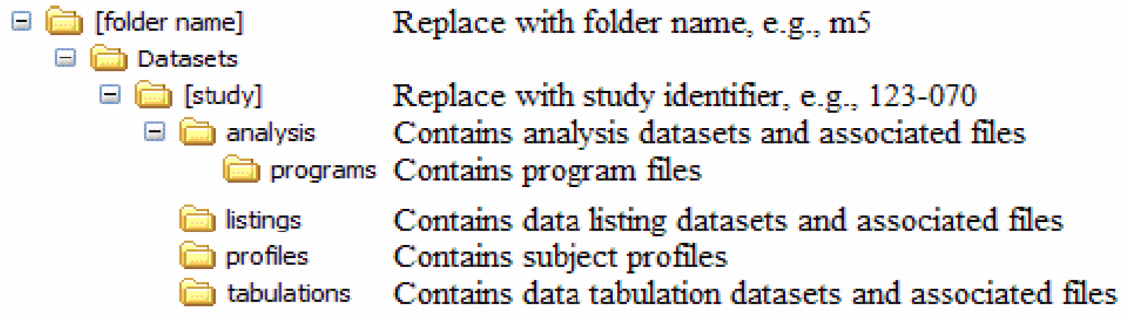
In an effort to provide illustration of ADaM analysis dataset concepts, examples will be provided that make reference to specific programming languages. Throughout ADaM documents, references to specific vendor products are examples only and should not be interpreted as an endorsement.

2. BACKGROUND / MOTIVATION

The marketing approval process for regulated human and animal health products often includes the submission of data from clinical trials. In the United States, data are a required element of a submission to the FDA as expressed in the Code of Federal Regulations (CFR). The FDA established the regulatory basis for wholly electronic submission of data in 1997 with the publication of regulations on the use of electronic records in place of paper records (21 CFR Part 11). In 1999, the FDA standardized the file format (SAS Version 5 Transport Files) for electronically submitting non-clinical and clinical data collected in clinical trials with the first of a series of guidance documents that describe the submission of clinical data and data definition (i.e., metadata) files for those clinical data in PDF format (Define.PDF). As of 2005, metadata can be submitted via the XML metadata (Define.XML) in place of the Define.PDF, as described in the FDA document regarding study data specifications. [4] More information about Define.XML can be found on the CDISC website. [2]

In parallel with the development of new clinical data submission guidance, the FDA has adopted the International Conference on Harmonization (ICH) standards for regulatory submissions and has issued a draft guidance on the electronic Common Technical Document (eCTD) as its framework for electronic communications regarding pharmaceutical product applications.

According to public presentations made by FDA representatives and the latest FDA guidance documents on the eCTD, submitted data can be classified into four types: 1) Data tabulations, 2) Data listings, 3) Analysis datasets, and 4) Subject profiles. These data are referred to in 21 CFR 11 as Case Report Tabulations (CRTs) and in ICH E3 as Individual Patient Data Listings (E3 16.4). The specification for organizing datasets and their associated files in folders is summarized in the following figure. [4]



[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

Historically, listings and subject profiles have been submitted as documents, not datasets. Data tabulations and analysis datasets are typically submitted as datasets and are defined as:

- **Study Data Tabulations (SDTM)** – datasets containing data collected during the study and organized by clinical domain. These datasets are described by CDISC Study Data Tabulation Model Implementation Guide (Version 3.1.1). [1]
- **Analysis Datasets –** datasets used for statistical analysis and reporting by the sponsor. These datasets are submitted in addition to the study data tabulation (SDTM) datasets and are described within this document.

For the purposes of simplifying this document, analysis datasets will be discussed within the context of electronic submissions to the FDA. However, analysis data models are applicable to a wide range of drug development activities in addition to regulatory submissions. They provide a standard for transferring datasets between sponsors and CROs, development partners and independent data monitoring committees. As adoption of the models becomes more universal, they will also facilitate in–licensing, out–licensing and mergers by providing a common model for analysis data and documentation across sponsors. The same principles and standards will apply, regardless of the purpose of the analysis datasets.

3. OVERVIEW OF ANALYSIS DATA MODELS

3.1 Key Principles

The overall principle in designing Analysis Datasets and related metadata is that there must be clear and unambiguous communication of the content, source and quality of the datasets supporting the statistical analyses performed in a clinical study.

Sponsors should strive to submit analysis datasets that can be analyzed with little or no programming or complex data manipulations. Such datasets are said to be “Analysis-ready” or “One Statistical Procedure Away” from the statistical results. This approach eliminates or greatly reduces the amount of programming required by the statistical reviewers. Appendix 8.4 gives an example in SAS, but the concepts apply to all statistical packages.

Analysis Datasets should be useable by currently available tools, but should provide machine-readable metadata to facilitate future standard analysis tool development. Metadata and other documentation should provide clear, concise communication of the analytic results of a clinical trial from the sponsor to the regulatory reviewers, including statistical methods, transformations, assumptions, derivations and imputations performed. The metadata, programs and other documentation serve to codify the analyses described in the Statistical Analysis Plan (SAP) as well as other analyses performed, and are discussed in detail in Sections 5 and 6.

Key Principles for Analysis Datasets

Analysis datasets should:

- facilitate clear and unambiguous communication
- be useable by currently available tools
- be linked to machine-readable metadata
- be analysis-ready

3.2 Analysis Data Flow Diagram in Research Process

The general flow diagram of data from its source through the analysis results is shown in Figure 1 below.

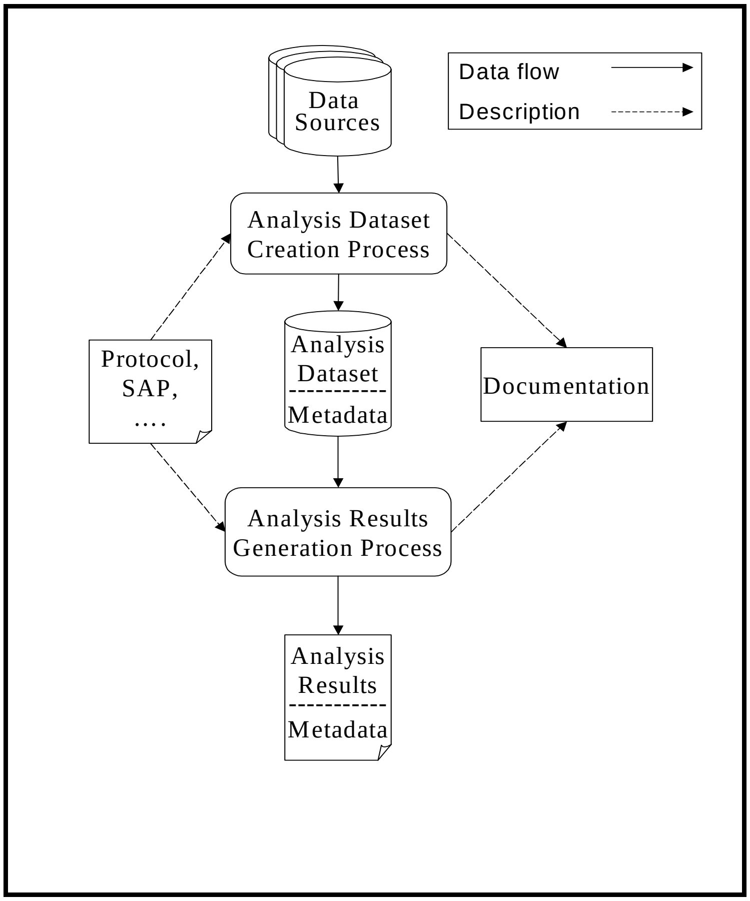


Figure 1: Analysis Data Flow

A variety of sources are possible for analysis datasets. One source could be the SDTM datasets submitted as part of a regulatory submission. In all cases, the data sources should be clearly described in the metadata and the analysis dataset creation documentation (see Section 5.4).

To facilitate clear communication, we distinguish between the processes of Analysis Dataset Creation and Analysis Results Generation. These two processes have distinct purposes and should each be clearly described and documented.

- **Analysis Dataset Creation –** The processing and programming steps used to create the Analysis Datasets using the data sources as input. The analysis dataset and variable metadata are defined in this step. Additional documentation may include programs or code fragments and links to the Protocol or Statistical Analysis Plan.
- **Analysis Results Generation –** The programming steps used to generate an analysis result, using submitted data as input. The analysis results metadata are defined in this step. Additional documentation may include analysis results programs or code fragments and links to the Statistical Analysis Plan or statistical appendix of the final report. The output is the results presentation and display objects (e.g., tables, data for graphics, test statistics, p-values, etc.).

These processes, datasets, results, metadata and documentation are discussed in detail in the following sections of this document.

3.3 Metadata Components

The analysis datasets and related metadata will facilitate the review of the clinical trial data and the analyses performed. There are three types of metadata described in this document. These include:

- Analysis dataset metadata provides certain key pieces of information describing each analysis dataset, including documentation and/or analysis dataset creation programs. (Refer to Section 5.1)
- Analysis variable metadata describes the variables within the analysis datasets, including links to relevant documentation providing additional details about the source and creation of the analysis variables, e.g. detailed descriptions of algorithms involved and/or references to analysis dataset creation programs. (Refer to Section 5.4)
- Analysis results metadata provides certain key pieces of information describing each important display, including which analysis dataset was used and links to relevant documentation providing details about the analyses performed, e.g. a specific section of the statistical analysis plan and/or analysis generation programs. (Refer to Section 6)

Analysis results metadata provides a link between an analysis result and the analysis dataset used to calculate the result. The other two types of metadata relate directly to the analysis dataset, with the analysis dataset metadata describing the analysis dataset as a whole and the analysis variable metadata describing the variables within the dataset.

4. ANALYSIS DATASETS

4.1 Practical Considerations

An analysis dataset will gather from various sources (e.g., data tabulation datasets) all of the variables required for performing the statistical analysis it is designed to support. For example, data may be required from the disposition, demographics, subject characteristics, vital signs, questionnaires, and exposure domains. By gathering the data into an analysis dataset, including any derived variables, further complicated data manipulation will not be required prior to the analysis. An example of a composite endpoint requiring complex algorithms and source variables from multiple datasets is shown in Appendix 8.7.

In creating analysis datasets, one goal should be to have the optimum number of analysis datasets needed to accomplish the various analyses, with the minimum requirement being a subject-level analysis dataset. Analysis datasets should be designed to allow analysis and review with little or no programming or data processing. Redundancy between analysis datasets will often be necessary so that the datasets are analysis-ready (e.g., age in the adverse event analysis data set as well as an efficacy dataset). In addition, redundancy between analysis datasets and SDTM domain datasets is acceptable. If a variable exists in an SDTM domain that can be used for an analysis without any change, then this variable should be included in the analysis dataset “as is”, with all SDTM attributes retained. An analysis dataset can be designed so that it can be used for multiple analyses. To aid in the review and use of analysis variables, there may be variables included that are not actually used in any of the submitted analyses, but are still of interest to the sponsor or reviewer (e.g., an identification flag for subjects who had an event of clinical interest). Analysis datasets will be provided to support the key analyses in a report or submission. The determination of which analyses are key analyses will be agreed between the sponsor and the recipient of the data.

Analysis datasets will be named using the convention "ADxxxxxx." The subject-level analysis dataset will be named "ADSL" as described in Section 7. For all other analysis datasets the xxxxxx portion of the name will be sponsor-defined, using a common naming convention across a given submission or multiple submissions for a product. Naming conventions for variables created (not to be confused with any standard variables required by SDTM) within the analysis should use the SDTM naming fragments where feasible. Otherwise the analysis variable names will be sponsor-defined, and should also follow a common naming convention across a given submission or multiple submissions for a product. This should allow for optimum clarity for any reviewer.

Analysis datasets must

- include a subject-level analysis dataset named "ADSL" (refer to Section 7).
- consist of the optimum number of analysis datasets needed to allow analysis and review with little or no programming or data processing.
- maintain SDTM variable attributes if the identical variable also exists in an SDTM dataset.
- be named using the convention "ADxxxxxx."
- follow naming conventions for datasets and variables that are sponsor-defined and applied consistently across a given submission or multiple submissions for a product, yet use published SDTM naming fragments for variables where feasible.

Although this document discusses some of the statistical and programming issues that arise in the creation of an analysis dataset, it is by no means a complete list. Trial design, statistical methods, sponsor SOPs and "real world" issues that arise during the conduct of the trial may complicate the definitions and derivations shown here. Additionally, the variables presented are by no means exhaustive and real world analysis files will likely contain additional variables used for analyses.

The following comments identify some statistical and programming issues to be considered in creating analysis datasets, but should not be interpreted as the only issues for a specific trial. To facilitate review and comprehension of the analysis datasets and analysis results, these issues may be important to represent in either Analysis Dataset or Analysis Results documentation or metadata.

- How are missing values handled in the analysis dataset? Should they be re-coded in some way in the analysis dataset, i.e. replaced with a special value? Missing values should be clearly encoded such that they can be identified and handled correctly in analysis computations.
- If a missing value is replaced by an imputed value (such as the last observation or the mean of existing values), what indication of that will be included in the analysis dataset? This imputation should be clearly documented and represented in the analysis dataset.
- Is there a difference in the source data between data that are actually missing and other coded values (e.g., does a missing value mean the data were not collected, not entered, or

not a “Yes”)? Is there a way to incorporate that difference in the coded value in the analysis dataset (e.g., Y=yes, N=no, X=data not available)?

- The visit window is often computed using the decision rules from the SAP. On rare occasions (hopefully), this may also require human intervention for cases not anticipated in the SAP. It is possible that the visit window will need to be computed in an interim dataset before endpoints can be computed. In most cases, this interim dataset would not be submitted. All decisions and processing steps of the visit windowing process should be fully documented.
- If the analysis results in p-values or other comparative statistics, data should be included in the analysis dataset that will allow the statistic to be produced with minimal additional computation. The documentation accompanying the analysis dataset should specify clearly how the statistic was produced, including any multiple comparison procedures that might have been used. For example, if the analysis is a Cochran-Mantel-Haenszel comparison between treatment groups of the proportion of subjects who responded to treatment, controlling for age group, the age group of the subject as well as whether or not the subject responded to treatment will be included in the analysis dataset.
- If multiple records are eligible for analysis, the record actually analyzed should be clearly identified. For example, if the maximum on-treatment value is to be summarized, that record should be flagged. Or if the value closest to the protocol-defined scheduled visit is to be analyzed, that record should be flagged.
- Variables that are changed or derived (e.g., logarithmic transformation, percent change from baseline) from the original data should be clearly identified. Depending on the structure of the data and the conventions being followed, a flag variable or a special naming convention may be used to indicate changed or derived variables. The algorithm used for the change or derivation, including the names of the variables containing the source or original data, and the reason for the change or derivation should be documented within the metadata.
- When a statistical analysis is based on a derived variable that is obtained from multiple records, such as a derived value that is calculated as the average across several records, or when a statistical analysis uses just a subset of records, such as using just those visits that adhere to a visit windowing rule, the decision must be made whether to retain all of the original records in the analysis datatset. As a general rule, if any complex derivations or decisions were made in the course of creating a derived record or selecting a subset of records, it is prudent to include all of the original data so that the reviewer can substantiate these decisions and/or explore the impact of different decision rules. If instead the derivations were simple and straightforward with no exceptions to an analysis rule and the metadata documentation is clear written, then this may be a situation where including redundant information is not necessary

4.2 Analysis Dataset Variables

Analysis dataset variables should conform to the CDISC Submission Metadata Model and be consistent with the SDTM V3.x standard, where practicable. (Note that "SDTM V3.x" refers to SDTM Version 3.1 and all subsequent versions.) If SDTM variables are included in Analysis Datasets without any changes, the SDTM variable attributes should be retained. Refer to Section 7.4 for an example.

Table 1 presents statistical analysis variables that should be considered for inclusion in most Analysis Datasets.

Table 1: Analysis Dataset Variables

Variable Class	**Examples[1]**	**Naming Rules**	**Content Rules**	**Comments**
Identifiers	STUDYID, USUBJID, SUBJID, SITEID INVID	Be consistent with SDTM V3.x variable names	Variable values should be identical to the values in SDTM domains.	STUDYID and USUBJID are **required;** additional variables should be included as needed. Variables may be given additional roles specific to analyses. If an analysis requires pooling of investigator sites, a derived pooled investigator identifier variable should be added.

[1] **ADaM variables are shown in BOLD,** SDTM V3.1 variables are in non-bold

Table 1: Analysis Dataset Variables

Variable Class	Examples[1]	Naming Rules	Content Rules	Comments
Analysis Population Indicators	ITT SAFETY PPROT FULLSET **ITTV** **SAFV** **ITTM** **SAFM** **FULLV** **FULLM**	These variable names should be used to identify these specific populations. These variable names are referenced in SDTM V3.x and may be represented in a supplemental qualifier domain. Append a V (for visit) or M (for measurement) to SDTM population variables as needed	Variable values should be identical to the values in SDTM domains, if present	The SDTM V3.x standard allows for population indicators or flags in the Supplemental Qualifiers (SUPPQUAL) dataset and standard name codes are suggested. Nearly all statistical analyses will **require** at least one population indicator and it is **required** to include all indicators for all populations for which a given analysis is performed. If the appropriate flags already exist in the SDTM domains, they should be used in the analysis datasets. Depending on the analyses performed, multiple population flags may be needed, for example, Efficacy versus Safety. Longitudinal study designs may require population flags at the Visit or Outcome level. Analysis datasets should include analysis population indicator variables at whatever level (e.g. subject, visit, or measurement) is necessary to clearly communicate the analysis and study design.

[1] **ADaM variables are shown in BOLD,** SDTM V3.1 variables are in non-bold

Table 1: Analysis Dataset Variables

Variable Class	Examples[1]	Naming Rules	Content Rules	Comments
Analysis Date Variables	See the ADaM Implementation Guide [3]	--DT for date variables --DTM for date/time variables	Analysis date variables are numeric, such as SAS Dates. Due to possible imputation, these values may differ from companion date-time variables in SDTM domains	SDTM V3.x has eliminated SAS numeric date/time variables and uses ISO 8601 character date strings whose names end in --DTC. In order to use dates for computations or graphical presentation, numeric versions of Date and Date-Time variables (such as SAS Dates) are **required** to be included on analysis datasets. The choice of presenting Date or Date-Time variables should be made depending on the context of the analysis. Partial dates are **required** to be identified by imputation indicator variables (--DTF) to indicate level of imputation. See table in Section 4.2.1 for further documentation.

[1] **ADaM variables are shown in BOLD,** SDTM V3.1 variables are in non-bold

Table 1: Analysis Dataset Variables

Variable Class	**Examples**[1]	**Naming Rules**	**Content Rules**	**Comments**
Indicator Variables	See ADaM Implementation Guide [3]	--DTF for date imputation flags -----FN for numeric indicator flag variables and ------FL for character indicator flag variables. (note: because of the 8 character limit for XPORT compliant files, the FL or FN prefix may need to be truncated; the variable name should be as close as possible to the parent variable name)	The value of indicator variables can be either informative, such as date imputation indicators (e.g. D, M, Y to indicate which part of a date value was imputed), or flag variables, where a value of 1 is used to flag the value or record as being important in some manner to an analysis. It is not required to use 0 for the remaining values or records.	Indicator variables are **required** when it is necessary to inform the reviewer that there is something unique about a variable or individual data record. For example, an indicator variable may be used to identify dates or calculated values that were imputed or that a particular value was used (or not used) in a per-protocol assessment. They may also be used on a record basis to indicate that the flagged record contains the values used for a particular time point, such as endpoint assessment. Thus, indicator variables may be created for individual variables or for individual records. Any indicator variable used to identify imputations should be accompanied by variable-level metadata that defines the imputation rules used. Similarly, any indicator variables that are created to identify values or records used in a specific way for an analysis should have well documented metadata to explain their use. See Section 4.2.1 below for an example of date imputation flags. Indicator flag variables with two levels are **required** to be numeric and of the form (1 and 0) or (1 and missing), when the variables are used to perform simple arithmetic computations.

[1] **ADaM variables are shown in BOLD,** SDTM V3.1 variables are in non-bold

Table 1: Analysis Dataset Variables

Variable Class	**Examples**[1]	**Naming Rules**	**Content Rules**	**Comments**
Analysis Study Day Variables	ANLDY	ANLDY (in days) ANLDYT (in seconds)	ANLDY (in days) = --DT – Reference Date +1. ANLDYT (in seconds) = --DTM – Reference Date/time + 1 sec.	This is a **required** variable IF the computation of duration from a pre-treatment day to a post-treatment day is needed since using the SDTM V3.x study day variable (--DY) will lead to mathematical inconsistencies due to the omission of a value of 0. Reference date may be numeric equivalent of RFSTDTC from DM Domain. See Section 4.2.2 below for a comparison of SDTM V3.x study day variable and ADaM Analysis study day variable.
Visit Timing Variables	VISIT, VISITNUM, VISITDY, **AVISITDY**	Be consistent with SDTM V3.x variable names AVISITDY is a derived variable representing the analysis day (ANLDY) of a planned visit and is used where visits occur at negative visit days	Variable values should be identical to the values in SDTM datasets.	Timing variables should be consistent with SDTM V3.x wherever applicable. Common timing variables include VISIT, VISITNUM, and VISITDY. At least one variable is **required**. Caution should be taken with VISITDY (planned visit day) if it follows the SDTM V3.x convention of omitting Day 0 and visits occur at negative visit days. In these cases, use the analysis planned visit day

[1] **ADaM variables are shown in BOLD,** SDTM V3.1 variables are in non-bold

Table 1: Analysis Dataset Variables

Variable Class	Examples[1]	Naming Rules	Content Rules	Comments
Numeric Code Variables	**AESEVN** **SEXN** **RACEN**	Append an 'N' to the SDTM V3.x variable name IF there is a one-to-one correspondence between the character SDTM V3.x variable and the numeric variable. Some truncation of the name may be needed to meeting the SAS V5 transport 8-character restriction. If there is not a one to one correspondence between the character SDTM V3.x variable and the analysis variable (such as the combining SDTM values into one analysis group), then derive a new variable name.	Use numeric values that are most appropriate for the desired sort order of the values	Existing FDA Guidance documents have discouraged the use of format libraries for user-defined formats and suggest that character variables with meaningful values be used instead. The character SDTM V3.x variables should be utilized where appropriate. There are, however, cases where a numeric version of a categorical variable is **required** for statistical purposes. An ordered analysis or report table may require a sort order other than the alphabetic values of the variable. In other cases, a statistical model may require numeric 0/1 variables as indicators. In order to accommodate these statistical requirements, additional numeric variables may be needed. These variables should only be added if it is anticipated that they are needed for specific analysis requiring ordered or numeric values. Numeric code variables should be named by using a common variable naming convention across a given submission. See Section 4.2.3 below for an example.

[1] **ADaM variables are shown in BOLD,** SDTM V3.1 variables are in non-bold

Table 1: Analysis Dataset Variables

Variable Class	Examples[1]	Naming Rules	Content Rules	Comments
Analysis Treatment Variables	**TRTP** **TRTPN** **TRTA** **TRTAN** **ARM** **ARMN** **ARMA** **ARMAN**	TRTP (character) and TRTPN (numeric) for planned treatment TRTA (character) and TRTAN (numeric) for actual treatment Corresponding rules for ARM but planned treatment variables do not need a 'P' in the variable name since by definition these variables represent the planned treatment. For numeric variables, the naming convention adds "N" to the end of the character variable name. In SAS V5 transport, some truncation of the name may be needed to meet the 8-character restriction.		Treatment variables are **required** to be present in all analysis datasets. Analyses by treatment group require a single variable indicating the subject treatment. In some simple trial designs, this may correspond to the SDTM ARM variable and the SDTM variable should be used without change to attributes. For reviewer consistency, both ARM and TRT variables should be present in at least the ADSL analysis data set, even if they are identical. All other analysis data sets should include the TRT variables. In other designs such as cross-over trials, the treatment planned for a particular segment may require ARM, ELEMENT and VISIT or EPOCH. If an analysis is performed on the actual treatment received, instead of the planned treatment, the variables corresponding to the actual treatment are **required** to be present in addition to the planned treatment variables. For clarity, these variable values are character and may not sort in ascending order of dosage. For any analyses that require ordered categories, a numeric code variable is required, yet numeric versions should only be used when needed for display or ordering purposes.. See Section 4.2.4 below for an example of Analysis Treatment Variables.

[1] **ADaM variables are shown in BOLD,** SDTM V3.1 variables are in non-bold

4.2.1 Analysis Date Imputation Flag Example

	ADaM Date Variables		**SDTM V3.1.1 Date-Time Variables**
Missing Elements	**--DT Date **Imputed**	**--DTF**	**Corresponding --DTC String**
None	YYYY-MM-DD	blank	YYYY-MM-DD
Day	YYYY-MM-**	D	YYYY-MM
Month (and Day)	YYYY-**-(**)	M	YYYY
Year (and M, D)	****-(**)-(**)	Y	

Missing years are typically only imputed when the year can be inferred from the context (e.g. the year of collection).

4.2.2 Comparison of Study Day Variable versus Analysis Study Day Variable

The following illustrates the difference between the SDTM V3.1.1 study day variable and the analysis study day variable.

Date Jan	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Visit	SCR				TRT				FU
SDTM –DY	-4	-3	-2	-1	1	2	3	4	5
ANLDY	-3	-2	-1	0	1	2	3	4	5

4.2.3 Example of Numeric Code Variables

Example of Analysis Variable Metadata for Numeric Code Variables

Variable Name	**Variable Label**	**Type**	**Controlled Terms or Format**[1]	**Source**
AESEV	Adverse Event Severity	Char	Mild, Moderate, Severe, Life-threatening	AE.AESEV
AESEVN	Adverse Event Severity Numeric Code	Num	1=Mild 2=Moderate 3=Severe 4=Life-threatening	Coded from AE.AESEV

[1] Throughout the document, the information in the “Controlled Terms or Format” column in the analysis variable metadata examples should be considered example only and not meant to represent terminology to be adopted.

In the example above, AE.AESEV refers to the AESEV variable within the SDTM adverse event domain.

4.2.4 Analysis Treatment Variables

Example of Analysis Variable Metadata for Analysis Treatment Variables

Variable Name	Variable Label	Type	Controlled Terms or Format	Source
TRTP	Planned Treatment Group	Char		Derived from DM.ARM, SE.ELEMENT
TRTPN	Planned Treatment Group Numeric Code	Num	0=Placebo 1=5 mg XX 2=10 mg XX	Coded from TRTP
TRTA	Actual Treatment Group	Char		Derived from the exposure dataset
TRTAN	Actual Treatment Group Numeric Code	Num	0=Placebo 1=5 mg XX 2=10 mg XX	Coded from TRTA

In the example above, the variable name is a two part name; incorporating the domain and the variable within the domain (e.g. DM.ARM refers to the ARM variable within the SDTM DM domain).

5. ANALYSIS DATASET DOCUMENTATION

Analysis dataset documentation provides the link between the general description of the analysis (as found in the Protocol Data Analysis Section, SAP or the reported analysis methods) and the source data. The source(s) of the Analysis Dataset should be clearly documented, allowing the reviewer to trace back data items to their source. The analysis dataset metadata and analysis variable metadata form an important part of this documentation. Depending on the complexity of the algorithms involved, the trial design, and the content and structure of the analysis dataset, written documentation and analysis file generation programs may also be submitted as part of the analysis dataset documentation.

5.1 Analysis Dataset Metadata

The Analysis Dataset Metadata conforms to the CDISC Submission Metadata Model. The datasets should have descriptive names, should indicate “analysis” or “statistics” in the dataset label, and should have a PURPOSE of Analysis in the dataset metadata. The dataset names should not duplicate any of the SDTM domain names within a submission. Refer to Section 7.3 for an illustration of analysis dataset metadata.

5.2 Analysis Dataset Creation Documentation

Written documentation may include descriptions of the source datasets, processing steps, and scientific decisions pertaining to creation of the dataset. This documentation should clearly distinguish those derivations and decision rules that were specified a priori from those changes and decisions that were data-driven. Key issues for consideration in analysis dataset creation documentation include (but are not limited to):

- Derived variables
- Added observations (for time-point analysis or imputed data capture)
- Visit windows
- Omitted observations
- Multiple observations
- Imputed data
- Missing data
- Dropouts
- Data item-specific derivations, i.e. changes to a data value for a specific observation.

5.3 Analysis Dataset Creation Programs

Statistical software programs may also be included as part of the analysis dataset documentation. These programs may be classified into three levels of increasing functionality and complexity:

- As pseudo-code embedded in written documentation of the creation of the dataset
- As code fragments that a reviewer could include in a program
- As stand alone, fully-functioning programs that replicate the creation of the dataset in another programming environment.

It should be noted that FDA requirements on submission of programs and how they will be used in the review of a submission are currently under development. In the interim, the alternatives listed above might be appropriate documentation of analysis datasets transferred between sponsors and other parties, independent of FDA guidance.

5.4 Analysis Variable Metadata

- The analysis variable metadata describes each variable in the analysis dataset. The Source column provides details about where the variable came from in the source data or how the variable was derived. This column should be used to identify the immediate predecessor data file and can contain hyperlinked text which will refer to the reviewer to additional information. This column differs from the ORIGIN attribute since Origin identifies the location of the first occurrence of the variable.

Throughout this document, the information in the “Controlled Terms or Format” column in the analysis variable metadata examples should be considered example only and is not meant to represent terminology to be adopted.

5.5 Analysis Variable Value-Level Metadata

For variables containing more than one type of measure, metadata is needed at the value level in addition to the variable level. Value-level metadata will facilitate the viewing of “vertical” data sets by providing labels and formats for each value. It will allow the transformation of data back and forth between “vertical” and “horizontal” dataset structures while preserving all metadata at the value-level. Value-level metadata is described as part of the proposed DEFINE.XML standard [2].

6. ANALYSIS RESULTS METADATA

Analysis results metadata describes the major attributes of each important analysis result in a report. (Analysis results metadata may not be necessary for every analysis included in a report or submission, but only for the key analyses. The determination of which analyses are key analyses will be agreed between the sponsor and the recipient of the data.) Analysis results may include statistical statements in the report such as treatment effect and p-values, tables or figures. Analysis results metadata will provide critical information concerning an analysis in a standard format in a predictable location. This will allow reviewers to link from a statistical result to metadata describing the analysis, the reason for performing the analysis, and the datasets and programs used to generate the analysis. Note that analysis results metadata is not part of an analysis dataset, but that one of the attributes of analysis metadata describes the analysis datasets used in the analysis. The following attributes can be used to describe each key analysis.

- **ANALYSIS NAME –** A unique identifier for the specific analysis. The column may include a table number or other sponsor-specific reference, such as the title of the display.
- **DESCRIPTION –** A text description of the contents of the display. This will normally contain more information than the title of the display.
- **REASON –** The high-level reason for performing this analysis. It will indicate when the analysis was planned and the purpose of the analysis within the body of evidence. Using consistent terminology in this field will allow ease in searching and identifying analyses. Refer to Appendix 8.2.
- **DATASET –** the name of the dataset(s) used in the analysis. In most cases, this will be a single dataset. If multiple datasets are used, they should all be listed here. The column may also include specific selection criteria so that the reviewer can easily identify the appropriate records from the analysis dataset (e.g., “where ITT=Y”).
- **DOCUMENTATION –** contains the information about how the analysis was performed. This information could be a text description, or a link to another document such as the protocol or statistical analysis plan, or a link to an analysis generation program (i.e., a statistical software program used to generate the analysis result). The analysis method could be documented in the protocol or the statistical analysis plan, or somewhere on the display itself. What the documentation column contains will depend on the level of detail required to describe the analysis itself, whether or not the sponsor will be providing a corresponding analysis generation program, and sponsor-specific requirements and standards.

Additional information that the sponsor may consider important for inclusion in the analysis results metadata include the type of analysis (e.g., patient-level summary, event-level summary, line listing) and a list of the variables in the analysis dataset that are used in the analysis.

Refer to Section 7.5 for an illustration of analysis results metadata.

7. SUBJECT-LEVEL ANALYSIS DATA MODEL

This model will present some of the issues that should be considered when creating a subject-level analysis dataset. It should be stressed that the subject-level analysis dataset described here can be used for multiple types of analyses, including descriptive, categorical, and modeling, depending on what variables are included in it. Refer to other ADaM models for information pertaining to a specific type of statistical analysis.

A subject-level analysis dataset and its related dataset documentation are the minimum requirement if any analysis datasets are submitted. The dataset will have one record per subject and will be named "ADSL."

A subject-level analysis dataset will contain all of the variables that are important for describing the target population to whom the study results are generalizable. These variables will include data that either describe the subjects or events in a clinical trial prior to treatment, or that group the subjects or events in some way for analysis purposes. This will include critical demographic and baseline characteristics of the subjects, as well as other factors arising during the study that could affect response. There may be variables included that are not actually used in any of the submitted analyses, but are still of interest to the sponsor or reviewer. It should be noted that the data assembled into a subject-level analysis dataset can be used as the source for data used in other analysis datasets for grouping subjects or events.

ICH Guidance recommends that "in addition to tables and graphs giving group data for baseline variables, relevant individual patient demographic and baseline data, … for all individual patients randomized (broken down by treatment and by centre for multi-center studies) should be presented in by-patient tabular listings." (Ref: ICH E3 Guidance for Industry: Structure and Content of Clinical Study Reports, Section 11.2.) Often a reviewer and sponsor will agree that submission of subject-level data will meet this requirement. If that is the case, variables included in a subject-level analysis dataset will need to include those needed to meet this regulatory requirement.

The critical variables included in a subject-level analysis dataset will depend on the specific nature of the disease and on the protocol, but will usually include (refer to the ICH E3 Guidance for Industry: Structure and Content of Clinical Study Reports for a more detailed listing):

- Demographic variables (age, sex, race, other relevant factors)
- Disease factors (including baseline values for critical clinical measurements carried out during the study or identified as important indicators of prognosis or response to therapy)
- Treatment code/group
- Other factors that might affect response to therapy
- Other possibly relevant variables (e.g., smoking, alcohol intake, menstrual status for women)

Therefore, the critical variables will include factors that are either descriptive, known to affect the subject's response to drug (in terms of either efficacy or safety), used as strata for randomization, or identify the subject or event as belonging to specific subgroups (e.g., population flags). For example, subjects may be randomized after being stratified by age group because it is believed that younger subjects respond differently to the study drug. In this situation, a subject's age category would be considered a critical variable for a study and included in the descriptive analysis dataset.

The following example illustrates a subject-level analysis dataset that will be the basis of the descriptive analyses that provide the foundation for any set of statistical analyses. By adding other variables to the dataset, it can also be used for purposes other than descriptive analyses, such as a primary efficacy analysis that uses a single assessment for each subject.. However, the sponsor should consider the advantage of having analysis datasets that specifically address primary efficacy or other key analysis. Separate analysis datasets may be advantageous since they include only the variables that are needed for a specific set of analyses. Including many variables in one analysis dataset for the sole reason that the dataset structure is similar may impede clear and concise communication with the reviewer.

The components of the example, including the layout of the mock tables, variable names, and dataset names, should be considered as examples only, not as standards. Within this document, numbers are used to identify displays and mock tables; this does not imply that numbers are the “correct” way of identifying and linking information. Other methods of identification, such as a short descriptive name, are acceptable.

7.1 Example: Disposition and Baseline Summaries in a Hypothetical Clinical Trial

This example illustrates the subject-level analysis dataset (and associated metadata) used to describe the analysis population (i.e., the subject disposition and the subject demographic and baseline characteristics) in a hypothetical clinical trial. Subjects are randomized to one of two treatments: Drug A or placebo. Subject demographic data, baseline characteristics, and medical history are collected at randomization.

The mock tables to be used for the displays to be included in the statistical analysis package are illustrated in Appendix 8.5.1; the example will illustrate an analysis dataset and analysis data metadata that could be used to generate the displays.

7.2 Example Analysis Dataset for Subject-Level Model

The mock tables in the statistical analysis plan can be generated from several different analysis dataset structures. The structure illustrated below is the one that the ADaM team determined would be most analysis-ready for the example. The use of a particular structure in the example is not meant to imply that it is the recommended format. The structure to be used for a given analysis dataset will be determined by the sponsor, and should produce the most analysis-ready dataset based on the analysis performed by the sponsor. In this example, the demographics and subject characteristics analysis dataset is used for both the disposition summary and the subject demographics and baseline characteristics summary.

In this example it is assumed that the analysis datasets use data tabulation datasets as source data. The hypothetical variables of interest are from the DM, DS, SE, SUPPQUAL, and VS SDTM V3.x domains. (Though it is possible to have different SUPPQUAL datasets for various domains, this example assumes one SUPPQUAL dataset.) The variable name is a two part name; incorporating the domain and the variable within the domain (e.g. DS.DSDECOD represents the DSDECOD variable within the DS domain). For the purposes of this example, it is assumed that the sponsor defined standard codes and variables listed in the following table are used for the specified data variables of interest. The reasons for discontinuation and the population flags used in an actual analysis dataset will vary based on sponsor-specific and study-specific requirements and standards.

Data of interest	SDTM Variable	Valid Values
Subject's discontinuation status	DS.DSDECOD	COMPLETED ADVERSE EVENT PROTOCOL VIOLATION LOST TO FOLLOW-UP OTHER
Subject's population flags	SUPPQUAL.QNAM	SAFETY ITT PPROT

7.3 Example Analysis Dataset Metadata for Subject-Level Model

The analysis dataset documentation for the example can be found in Appendix 8.5.2.

In this example, "ADSL" is the name of the analysis dataset, as required in Section 4.1.

Analysis Dataset Metadata						
Dataset	**Description of Dataset**	**Location**	**Structure**	**Purpose**	**Key Variables**	**Documentation**
ADSL	Demographic and subject baseline characteristics analysis dataset, includes population flags	*pathname*/adsl.xpt	One record per subject	Analysis	USUBJID	SAP, DS_ADSL.SAS

7.4 Example Analysis Variable Metadata for Subject-Level Model

The example of analysis variable metadata below assumes that specific flags have already been set in the data tabulation datasets (e.g., population flags) and that the creation of those flags will have been documented in the appropriate metadata. Many variables (e.g. additional identifiers) that would be in a submitted analysis dataset are not included in this example, to simplify the document. The reasons for discontinuation used in this example are hypothetical. Those used in an actual analysis dataset will vary based on the protocol-specific reasons for discontinuation. In the example, the baseline height and weight could also be brought in from the SDTM datasets where the baseline flag is set to "Y" (e.g., VSBLFL), if the derivation of the baseline flag is the same as that needed for the analysis datasets.

Analysis Variable Metadata for Analysis Dataset ADSL – 1 record per subject

Variable Name	Variable Label	Type	Controlled Terms or Format	Source	ADaM Notes
STUDYID	Study Identifier	Char		DM.STUDYID	
USUBJID	Unique Subject Identifier	Char		DM.USUBJID	
…	Additional identifiers				
SAFETY	Safety population flag	Char	Y or N	SUPPQUAL.QVAL (where QNAM=SAFETY)	All flag variables used to identify subject populations for which analysees were conducted should be included in ADSL
ITT	Intent to treat population flag	Char	Y or N	SUPPQUAL.QVAL (where QNAM=ITT)	
PPROT	Per Protocol set flag	Char	Y or N	SUPPQUAL.QVAL (where QNAM=PPROT)	
COMPLT	Completers population flag	Char	Y or N	SUPPQUAL.QVAL (where QNAM=COMPLT)	Additional completer flags should be added as needed, such as flags to indicate completion of primary efficacy visit or completion of a pre-specified number of weeks of exposure to study drug
TRTSTDT	Start Date of Treatment	Date		Derived from EX.EXSTDTC	
…	Additional analysis variables such as milestone dates				Milestone dates such as date of discontinuation, date of last dose, date of randomization should be added as necessary, especially if the dates are used to derived variables in other analysis files or used in statistical models.

Analysis Variable Metadata for Analysis Dataset ADSL – 1 record per subject

Variable Name	Variable Label	Type	Controlled Terms or Format	Source	ADaM Notes
DSREAS	Reason for discontinuation	Char	ADVERSE EVENT, PROTOCOL VIOLATION, LOST TO FOLLOW-UP, OTHER	DS.DSDECOD where DS.DSCAT=DISPOSITION EVENT and DS.DSDECOD not equal COMPLETED, OTHER if DSDECOD is other non-missing value, missing if DSDECOD=COMPLETED	
AGE	Age (yr) at reference start date	Num	3.0	DM.AGE	
AGEGRP	Age Group	Char	<21 21-35 36-50 >50	Derived from DM.AGE	
AGEGRPN	Age Group, Numeric	Num	1= <21 2= 21-35 3= 36-50 4= >50	Coded from AGEGRP	
SEX	Sex	Char	M, F, U	DM.SEX	
RACE	Race	Char		DM.RACE, using controlled terminology as specified in SDTM	
RACEN	Race, Numeric	Num	1=WHITE 2=BLACK 3=ASIAN 4=OTHER	Coded from DM.RACE	
ARM	Description of Planned Arm	Char	DRUG A PLACEBO	DM.ARM	Both ARM and TRT variables should be included even if they are identical. All other analysis data sets should contain the TRT variables. Sponsors should choose consistent values for TRTPN across the submission. The
TRTP	Planned Treatment Group	Char	DRUG A PLACEBO	Derived from DM.ARM	

Analysis Variable Metadata for Analysis Dataset ADSL – 1 record per subject					
Variable Name	**Variable Label**	**Type**	**Controlled Terms or Format**	**Source**	**ADaM Notes**
TRTPN	Planned Treatment Group, Numeric	Num	0=PLACEBO 1=DRUG A	Coded from TRTP	value of 0 is often used to indicate the reference treatment group. Often this is the Placebo arm yet the choice of numeric values for treatment are sponsor specific
HEIGHTBL	Baseline Height (cm)	Num	3.0	VS.VSSTRESN (where VSTESTCD=HEIGHT and VISIT=RAND), imputed from Screening visit if missing at Randomization	
WEIGHTBL	Baseline Weight (kg)	Num	5.1	VS.VSSTRESN (where VSTESTCD=WEIGHT and VISIT=RAND), imputed from Screening visit if missing at Randomization	
BMIBL	Baseline BMI (kg/m^2)	Num	5.2	Derived: WEIGHTBL / ((HEIGHTBL/100)**2)	

7.5 Example Analysis Results Metadata for Subject-Level Model

The following illustrates analysis results metadata for the example. The analysis results documentation for the example can be found in Appendix 8.5.

Analysis Results Metadata				
Analysis name	**Description**	**Reason**	**Dataset**	**Documentation**
Table 1.1 – Subject Disposition Summary	Summary of number of subjects in each analysis population, reasons for discontinuation	Pre-specified in Protocol	link to analysis variable metadata for adsl.xpt	SAP Sections X.W and X.X.
Table 1.2 – Demographic and Subject Characteristics, ITT Population	Summary statistics for selected demographic and subject baseline characteristics variables.	Pre-specified in Protocol	link to analysis variable metadata for adsl.xpt select records where ITT=Y	SAP Section X.Y

8. APPENDICES

8.1 Links referenced in document

[1] CDISC Submission Data Standards (SDS) Team documents: Study Data Tabulation Model (SDTM) and the SDTM Implementation Guide (SDTM-IG). Retrieved from <http://www.cdisc.org/models/sdtm/v1.1/index.html> January 2006.

[2] Define.XML: Submitting Machine-readable Metadata. CDISC webpage retrieved from <http://www.cdisc.org/models/define/define.html> January 2006.

[3] Guide to documents relating to the CDISC Analysis Data Model (ADaM), found on the CDISC website under the ADaM Standards <http://www.cdisc.org/models/adam/V1.0/index.html>.

[4] Study Data Specifications, Version 1.2. FDA Guidance Document (2006). Retrieved from <http://www.fda.gov/cder/regulatory/ersr/Studydata-v1.2.pdf> August 2006.

8.2 Suggested Terminology to be used in “Reason” within Analysis Results Metadata

Analyses can be grouped into categories, based on the reason for the analysis. Consistent terminology is helpful to reviewers and others using the reason category for searching or sorting. The following list contains suggestions for terms to be used as the reason for the analysis within the analysis results metadata. Additional categories may be added to the following list as the ADaM team continues to develop models.

Suggested Terminology (select one or more as the reason for the analysis):

- Pre-specified in XXX = Documented plans for the analysis are located in XXX. Valid terminology for XXX includes:
    - Protocol
    - SAP (statistical analysis plan)
    - Other a priori plans (genetics analysis plan, PK/PD analysis plans, pharmacoeconomics analysis plan).
- Data driven = Analysis performed in response to issue seen in the released data, not pre-specified, i.e. ad hoc analysis.
- Requested by YYY = Analysis performed in response to request from group external to the sponsor (e.g., FDA, EMEA, DSMB).

8.3 Definitions

Analysis Dataset Creation Program – Statistical software program used to create the analysis dataset.

Analysis Dataset Documentation - Written documentation may include descriptions of the source datasets, processing steps, and scientific decisions pertaining to creation of the dataset. Analysis dataset creation programs may also be included.

Analysis Dataset Metadata – provides information describing each analysis dataset

Analysis Datasets – datasets used for statistical analysis and reporting by the sponsor; submitted in addition to the data tabulation datasets.

Analysis Generation Programs – Statistical software programs used to generate an analysis, provide an "audit trail" (e.g., step-by-step process of how a result was obtained) for important results.

Analysis Variable Metadata – describes the variables within the analysis dataset

Analysis Variable Value-Level Metadata – describes the various possibilities included in variables in the analysis dataset that contain more than one type of measure

Analysis Results Documentation – Written documentation will include descriptions of planned and ad hoc analyses. The documentation may consist of the protocol, the statistical analysis plan, the statistical methods section of the study report, and analysis generation programs.

Analysis Results Metadata – describes the major attributes of each important analysis result in a report

CDISC – Clinical Data Interchange Standards Consortium

Data Tabulation Datasets - Datasets in which each record is a single observation for a subject. (Study Data Specifications [4])

Submission Data Domain Standards – Released by the CDISC Submission Data Standards (SDS) Team, Version 3.1.1 consists of two documents: SDTM and SDTM-IG. [1]

SDTM - Study Data Tabulation Model – Document which represents the underlying conceptual model behind the SDS standards. It defines a standard structure for study data tabulations that are to be submitted as part of a product application to a regulatory authority.

SDTM-IG - SDTM Implementation Guide - Document which includes the detailed domain descriptions, assumptions, and examples for human clinical trials.

8.4 Analysis-Ready Datasets

Consider the demographic table shown below. It has been determined that Age and Race are relevant to the study outcomes and the comparability of the treatment groups for these characteristics is computed and displayed (ICH E3, Section 11.2). Analysis-ready does not mean that this formatted table can be generated in a single statistical procedure or PROC. Rather it means that each statistic in the table can be replicated by running a standard statistical procedure (SAS PROC, S-PLUS function…) using the appropriate analysis dataset as input. This means that reviewers can replicate and explore these results with little or no data manipulation, allowing reviewers to concentrate on the results, not on programming.

Table DEM1 – Demographics by Treatment Assignment for all randomized patients

		Placebo	Drug A	P-value*
NUMBER OF SUBJECTS RANDOMIZED		nn	nn	
Number of subjects eligible per protocol		nn (xx%)	nn (xx%)	
Age (yrs) Mean(SD)		xx (xx.x)	xx (xx.x)	0.xxx
Sex N(%)	Female	nn (xx%)	nn (xx%)	
	Male	nn (xx%)	nn (xx%)	
Race N(%)	White	nn (xx%)	nn (xx%)	0.xxx
	Black	nn (xx%)	nn (xx%)	
	……	nn (xx%)	nn (xx%)	
Baseline Weight (kg) Mean(SD)		xxx (xx.x)	xxx (xx.x)	
Baseline Height (cm) Mean(SD)		xxx (xx.x)	xxx (xx.x)	

*Continuous variables will be analyzed using t-test. Categorical variables will be compared using chi-square.

NOTE: This is an illustrative example of analysis-ready datasets. It is not a recommendation to do hypothesis tests for baseline characteristics.

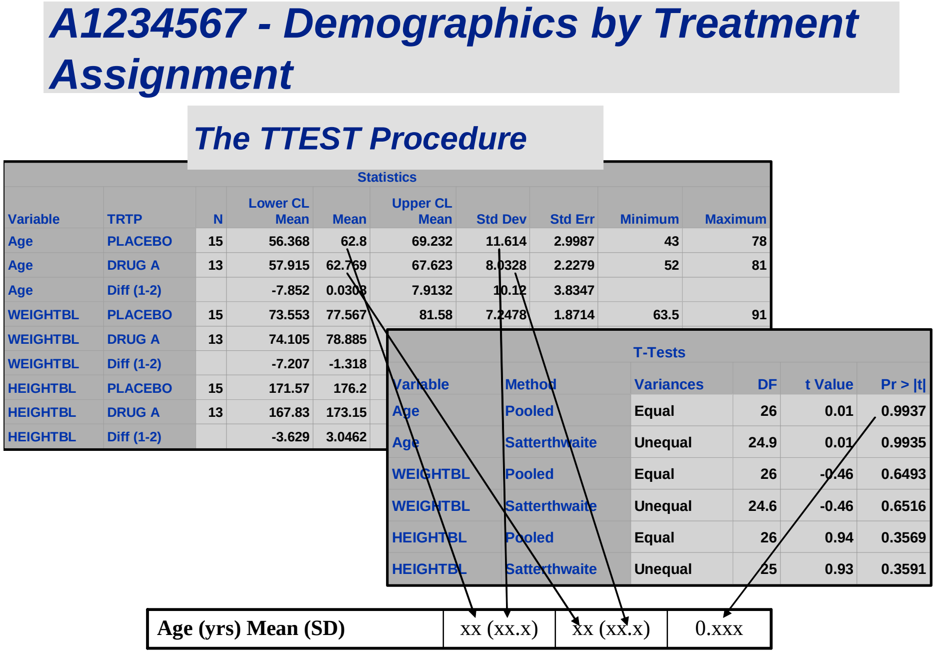
For example, the following SAS code will replicate results of Table DEM1 using an analysis dataset containing the appropriate variables.

```
PROC tabulate data=r.ADSL f=4.0;
class pprot trtp;
table all pprot, trtp*(n pctn<all pprot>);
run;

PROC freq data=r.ADSL;
table (race)*trtp/chisq nopercent norow;
run;
```

```
PROC ttest data=r.ADSL ci=none;
class trtp;
var age weightbl heightbl;
run;
```

The following annotated SAS procedure output results relate the SAS output with the corresponding elements of Table DEM1.



A1234567 - Demographics by Treatment Assignment

The TTEST Procedure

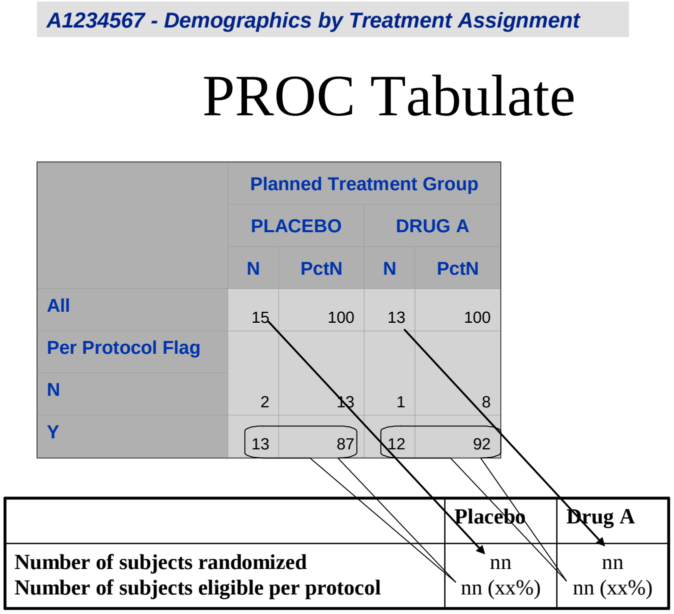
Statistics

Variable	TRTP	N	Lower CL Mean	Mean	Upper CL Mean	Std Dev	Std Err	Minimum	Maximum
Age	PLACEBO	15	56.368	62.8	69.232	11.614	2.9987	43	78
Age	DRUG A	13	57.915	62.769	67.623	8.0328	2.2279	52	81
Age	Diff (1-2)		-7.852	0.0308	7.9132	10.12	3.8347		
WEIGHTBL	PLACEBO	15	73.553	77.567	81.58	7.2478	1.8714	63.5	91
WEIGHTBL	DRUG A	13	74.105	78.885					
WEIGHTBL	Diff (1-2)		-7.207	-1.318					
HEIGHTBL	PLACEBO	15	171.57	176.2					
HEIGHTBL	DRUG A	13	167.83	173.15					
HEIGHTBL	Diff (1-2)		-3.629	3.0462					

T-Tests

Variable	Method	Variances	DF	t Value	Pr > |t|
Age	Pooled	Equal	26	0.01	0.9937
Age	Satterthwaite	Unequal	24.9	0.01	0.9935
WEIGHTBL	Pooled	Equal	26	-0.46	0.6493
WEIGHTBL	Satterthwaite	Unequal	24.6	-0.46	0.6516
HEIGHTBL	Pooled	Equal	26	0.94	0.3569
HEIGHTBL	Satterthwaite	Unequal	25	0.93	0.3591

Age (yrs) Mean (SD)	xx (xx.x)	xx (xx.x)	0.xxx

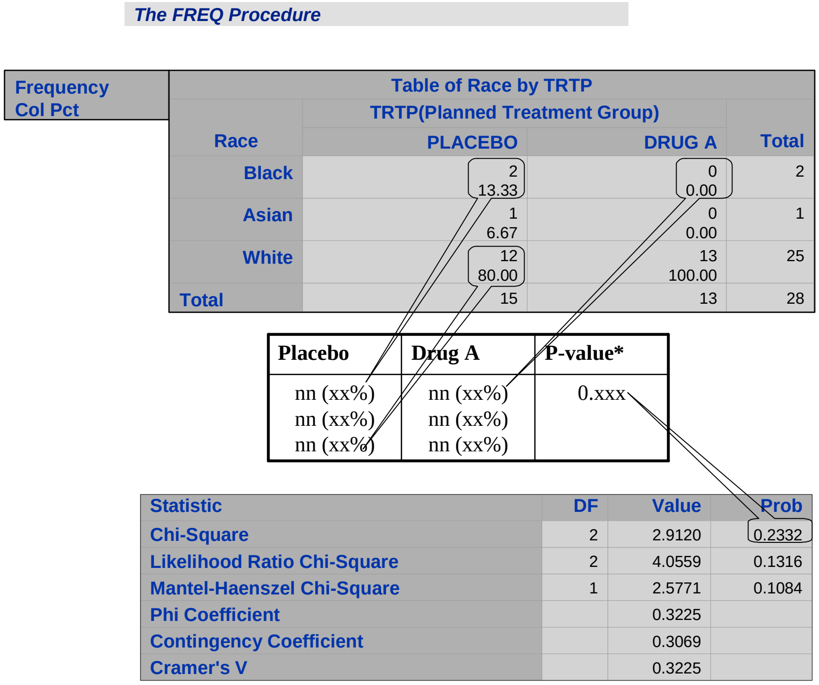


A1234567 - Demographics by Treatment Assignment

PROC Tabulate

	Planned Treatment Group			
	PLACEBO		DRUG A	
	N	PctN	N	PctN
All	15	100	13	100
Per Protocol Flag				
N	2	13	1	8
Y	13	87	12	92

	Placebo	Drug A
Number of subjects randomized	nn	nn
Number of subjects eligible per protocol	nn (xx%)	nn (xx%)



The FREQ Procedure

Frequency Col Pct	Table of Race by TRTP		
	TRTP(Planned Treatment Group)		
Race	PLACEBO	DRUG A	Total
Black	2 13.33	0 0.00	2
Asian	1 6.67	0 0.00	1
White	12 80.00	13 100.00	25
Total	15	13	28

Placebo	Drug A	P-value*
nn (xx%) nn (xx%) nn (xx%)	nn (xx%) nn (xx%) nn (xx%)	0.xxx

Statistic	DF	Value	Prob
Chi-Square	2	2.9120	0.2332
Likelihood Ratio Chi-Square	2	4.0559	0.1316
Mantel-Haenszel Chi-Square	1	2.5771	0.1084
Phi Coefficient		0.3225	
Contingency Coefficient		0.3069	
Cramer's V		0.3225	

It is often the case that analysis-ready datasets can also be used for subset analyses without additional programming. For example, the following SAS code can be used to generate a table similar to Table DEM1 for only those subjects meeting the “per protocol” criteria.

```
PROC freq data=r.ADSL(where=(pprot eq ’Y’));
table (race)*trtp/chisq nopercent norow;
run;

PROC ttest data=r.ADSL(where=(pprot eq ’Y’)) ci=none;
class trtp;
var age weightbl heightbl;
run;
```

8.5 Further Information for Example of Subject-Level Model

8.5.1 Analysis Results Documentation for Subject-Level Model Example

For this example, the protocol and statistical analysis plan provide sufficient description of the analyses.

Protocol:

The following are excerpts from the hypothetical study protocol:

> **Primary endpoint:**
>
> - The percentage of ITT population subjects pain-free 2 hours after treatment in the Drug A treatment group versus the Placebo treatment group.

Statistical Analysis Plan:

The following are excerpts from the hypothetical statistical analysis plan:

> **Section X.W. Analysis Populations**
>
> Subjects are included in the Safety population if they have been randomized and have taken at least one dose of study medication.
>
> Subjects are included in the Intent-to-treat (ITT) population if they have taken at least one dose of study medication and have attended at least one post baseline efficacy assessment.
>
> Subjects are included in the Per Protocol population if they are included in the ITT population and have no major protocol deviations. The following criteria will be used to exclude subjects from the PPROT population:
>
> - Use of any additional medication before the 2 hour assessment,
> - Use of prohibited drugs during the pre-treatment period.
>
> **Section X.X. Subject Disposition**
>
> The number of subjects in each analysis population will be provided. Subjects who are in the safety population but do not complete the study will be summarized by reason for discontinuation. (Refer to Mock Table 1.1.) All subjects randomized will be accounted for. ITT population subjects who are not in the Per Protocol population will be summarized by reason for exclusion.
>
> **Section X.Y. Demographic and Baseline Characteristics**
>
> Demographic and subject characteristic variables will be summarized within each treatment group for the safety and intent-to-treat populations. (Refer to Mock Table 1.2.) Categorical variables (sex, race, migraine history, age category) will be summarized by frequency counts. Continuous variables (age, weight, height, and BMI) will be summarized by descriptive statistics (n, mean, median, minimum, maximum and standard deviation). BMI will be derived from weight and height using the algorithm BMI=[weight in kg] / ([height in cm]/100)^2.

The following mock tables are excerpts from the statistical analysis plan. These are the displays to be included in the statistical analysis package.

Mock Table 1.1 Subject Disposition Table

```
                              Table 1.1
                   Summary of Subject Disposition
                     for all subjects randomized

                                    Placebo      Drug A       Total
Number of Subjects Randomized       nn           nn           nn
Safety Population [1]               nn           nn           nn
ITT Population [2]                  nn (xxx%)    nn (xxx%)    nn (xxx%)
Per Protocol Population [3]         nn (xxx%)    nn (xxx%)    nn (xxx%)

Completed Study [4]                 nn (xxx%)    nn (xxx%)    nn (xxx%)
Discontinued Study Prematurely      nn (xxx%)    nn (xxx%)    nn (xxx%)

Reasons for Discontinuation:
   Adverse event                    nn (xxx%)    nn (xxx%)    nn (xxx%)
   Lost to follow-up                nn (xxx%)    nn (xxx%)    nn (xxx%)
   Protocol violation               nn (xxx%)    nn (xxx%)    nn (xxx%)
   Other                            nn (xxx%)    nn (xxx%)    nn (xxx%)
-------------------------------------------------------------------------------
Note: Percentages are based on Safety population.
[1] Subjects are included in the Safety population if they have been
randomized and have taken at least one dose of study medication.
[2] Subjects are included in the ITT population if they have taken at least
one dose of study medication and have at least one post baseline efficacy
assessment.
[3] Subjects are included in the Per Protocol population if they are
included in the ITT population and have no major protocol deviations.
[4] Subjects completed the study if they were in the ITT population and
provided all follow-up information.
```

Mock Table 1.2 Demographic and Subject Characteristics

use for ITT and Safety population summaries (appropriate term inserted in table title)

Table 1.2
Summary of Demographics and Subject Characteristics of the
XXX Population

		Placebo	Drug A	Total
		(N=xxx)	(N=xxx)	(N=xxx)
Age (yr)	n	nn	nn	nn
	Mean	xx.x	xx.x	xx.x
	SD	xx.xx	xx.xx	xx.xx
	Median	xx.x	xx.x	xx.x
	Min.	xx	xx	xx
	Max.	xx	xx	xx
Age Group	n	nn	nn	nn
	<21	xx (xx%)	xx (xx%)	xx (xx%)
	21-35	xx (xx%)	xx (xx%)	xx (xx%)
	36-50	xx (xx%)	xx (xx%)	xx (xx%)
	>50	xx (xx%)	xx (xx%)	xx (xx%)
Race	n	nn	nn	nn
	White	xx (xx%)	xx (xx%)	xx (xx%)
	Black	xx (xx%)	xx (xx%)	xx (xx%)
	Asian	xx (xx%)	xx (xx%)	xx (xx%)
	Other	xx (xx%)	xx (xx%)	xx (xx%)
Sex	n	nn	nn	nn
	Female	xx (xx%)	xx (xx%)	xx (xx%)
	Male	xx (xx%)	xx (xx%)	xx (xx%)
Baseline Height (cm)	n	nn	nn	nn
	Mean	xx.x	xx.x	xx.x
	SD	xx.xx	xx.xx	xx.xx
	Median	xx.x	xx.x	xx.x
	Min.	xx	xx	xx
	Max.	xx	xx	xx
Baseline Weight (kg)	n	nn	nn	nn
	Mean	xx.x	xx.x	xx.x
	SD	xx.xx	xx.xx	xx.xx
	Median	xx.x	xx.x	xx.x
	Min.	xx	xx	xx
	Max.	xx	xx	xx
Baseline BMI (kg/m^2)	n	nn	nn	nn
	Mean	xx.x	xx.x	xx.x
	SD	xx.xx	xx.xx	xx.xx
	Median	xx.x	xx.x	xx.x
	Min.	xx	xx	xx
	Max.	xx	xx	xx

8.5.2 Analysis Dataset Documentation for Subject-Level Model Example

In this example, the written documentation that accompanies the analysis dataset metadata describes the source datasets, as well as provides programming code to illustrate how the baseline weight and height were derived.   It should also be noted that in this example the height and weight used to calculate BMI could be from two different visits. This may not be appropriate for all studies.

SAS program to generate subject description analysis dataset ADSL:

If the baseline flag (VSBLFL) in the data tabulation vital signs dataset appropriately identifies the value to be used (i.e., screening if randomization is missing), then the following code will not be necessary. However, if the flag does not meet the analysis requirements, it is possible that the sponsor would choose to provide SAS code to illustrate the derivation (including imputation of missing values) of the baseline height and weight. The following is an example of SAS code used to create a working dataset containing one record per subject.

DS_ADSL.SAS

```
*select screening and randomization height and weight data from
*  data tabulation datasets;
proc sort data=VS(where=(
  VISIT in (’SCR’,’RAND’) AND
  VSTESTCD in (’HEIGHT’,‘WEIGHT’)))
  out=vs_step1(keep=USUBJID VSTESTCD VSSTRESN VISIT);
  by USUBJID VSTESTCD;
  run;
*reformat the data to be one height record and one weight per subject;

  by usubjid vstestcd;
  var vsstresn;
  id visit;
  run;
*determine the randomization value (take from screening if
randomization value is missing);
data vs_step3;
  set vs_step2;
  if rand=. and scr ne . then rand=scr;
  run;
*re-format the data to be one record per subject, containing the
variables HEIGHTBL and WEIGHTBL;
proc transpose data=vs_step3 out=vital(rename=(HEIGHT=HEIGHTBL
WEIGHT=WEIGHTBL));
  by usubjid;
  var rand;
  id vstestcd;
  run;
```

8.5.3 Sample Dataset for Subject-Level Model Example – ADSL

Because of the length of the records, the dataset is portrayed in two segments in order to fit on the page.

SAMPLE DATASET FOR ADSL									
Obs	**Studyid**	**USUBJID**	**SAFETY**	**ITT**	**PPROT**	**COMPLT**	**DSREAS**	**AGE**	**AGEGRP**
1	XX0001	0001-1	Y	Y	Y	Y		30	21-35
2	XX0001	0001-2	Y	Y	N	N	ADVERSE EVENT	38	36-50

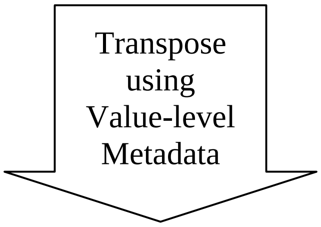
SAMPLE DATASET FOR ADSL (continued)									
Obs	**AGEGRPN**	**SEX**	**RACE**	**RACEN**	**TRTP**	**TRTPN**	**HEIGHTBL**	**WEIGHTBL**	**BMIBL**
1	2	F	WHITE	1	DRUG A	1	170	63.5	21.97
2	3	M	ASIAN	3	PLACEBO	0	183	86.2	25.74

8.6 Analysis Variable Value-Level Metadata

Value-level metadata is particularly valuable when transposing the “vertical” structure of the SDTM V3.x datasets to a more “horizontal” structure for analysis or display. For example, in SDTM V3.1.1 Vital Signs, the values of Height, Weight, and Frame Size are all stored in the same variable. A more “horizontal” analysis dataset may require separate variables for Height, Weight and Frame Size. A derived variable BMI, computed from the transposed height and weight will also be added to the analysis dataset. The vertical SDTM V3.1.1 Vital Signs dataset and the transposed, horizontal Vital Signs Analysis Dataset are shown below.

SDTM V3.1.1 Vital Signs Findings - 1 Rec/subject/visit/finding

USUBJID	VISIT	VSTESTCD	VSTEST	VSSTRESN	VSSTRESC
00001	RAND	HEIGHT	Height	183	183
00001	RAND	WEIGHT	Weight	88.2	88.2
00001	RAND	FRMSIZE	Frame Size		Large



Transpose using Value-level Metadata

ADVS -Vital Signs Analysis - 1 Rec/subject/visit

USUBJID	VISIT	HEIGHT	WEIGHT	BMI	FRMSIZE
00001	RAND	183	88.2	26.34	Large

There is no place to store individual attributes for values of VSTESTCD in the standard metadata model. An analysis dataset with a “more horizontal” structure of 1 record per subject per visit with variables HEIGHT, WEIGHT, **FRMSIZE** could not get variable information from the SDTM V3.1.1 metadata.

SDTM V3.1.1 Vital Signs Findings - 1 Rec/subject/visit/finding

Variable Name	Variable Label	Type	Controlled Terms or Format	Origin	Role
USUBJID	Subject ID	Char		Sponsor Defined	Identifier
VISIT	Visit Name	Char		Vital CRF	Timing

SDTM V3.1.1 Vital Signs Findings - 1 Rec/subject/visit/finding					
Variable Name	**Variable Label**	**Type**	**Controlled Terms or Format**	**Origin**	**Role**
VSTESTCD	Vital SignsTest ShortName	Char		CRF/Derived	Topic
VSSTRESN	Num. Result in Standard Units	Num		Derived	Qualifier
VSSTRESC	Char Result in Standard Units	Char		Derived	Qualifier

Value-level metadata for each test (VSTESTCD) is needed for formats, labels and origins at the value-level. Note that the value-level metadata also identifies the character (VSSTRESC) and numeric (VSSTRESN) variables containing the results of the measures named in VSTESTCD.

Vital Signs Value-level Metadata for VSTESTCD					
Value (VSTESTCD)		**Type**	**Controlled Terms or Format**	**Origin**	**Role**
HEIGHT	Height	Num	3.0	VSSTRESN	
WEIGHT	Weight	Num	5.1	VSSTRESN	
FRMSIZE	Frame Size	Char	Small, Medium, Large	VSSTRESC	

In the ADVS metadata shown below, some metadata elements from the value-level metadata are “passed through” from the source datasets while others are added by the statistical programmers during the creation of the dataset. In this example, the SDTM variable, VSSTRESU, that indicates the unit for each VSTESTCD, is not passed through to the analysis file. In many situations, this information can more succinctly be placed in the source/computational method/etc associated with each of the transposed VSTESTCD variables.

ADVS -Vital Signs Analysis - 1 Rec/subject/visi					
Variable	**Label**	**Type**	**Controlled Terms or Format**		**Role**
USUBJID	Subject ID	Char		DM.USUBJID	Identifier
SEX	Sex	Char	M, F, U	DM.SEX	Qualifier, Selection
VISIT	Visit Name	Char		VS.VISIT	Timing
HEIGHT	Height	Num	3.0	VS.VSSTRESN (where VS.VSTESTCD=’HEIGHT’)	Analysis
WEIGHT	Weight	Num	5.1	VS.VSSTRESN (where VS.VSTESTCD=’WEIGHT’)	Analysis

ADVS -Vital Signs Analysis - 1 Rec/subject/visit					
Variable	**Label**	**Type**	**Controlled Terms or Format**		**Role**
BMI	Body Mass Index in kg/m^2	Num	5.2	WEIGHT/ (.01*HEIGHT)**2	Analysis
FRMSIZE	Frame Size	Char	Small Medium Large	VS.VSSTRESC (where VS.VSTESTCD='FRMSIZE')	Analysis

8.7 Composite Endpoint Example

This is an example of an analysis endpoint requiring complex algorithms and source variables from multiple datasets.

Because of complexity of some derived variables, it may not be possible to describe a variable as simply as done in earlier examples in the Source column of the metadata.

For complex derived variables the Source field should provide a link to external documentation that explains the various sources of data and the algorithms involved in creating the variable.

The following example, drawn from the International Headache Society Guidelines, describes a composite endpoint that requires data from an efficacy dataset (Headache severity at different time points), Adverse Experiences, and Concomitant Medications datasets. It illustrates how an apparently simple binary outcome variable (outcome of the treatment of a single headache episode) has complex underpinnings and draws from data elements from different source datasets.

The variable metadata for this binary variable, named HASTPNFR on a headache analysis dataset ADHA, might look as follows:

Exa Analysis Variable Metadata for AD TPNFR					
Variable	**Variable Label**	**Type**	**Controlled** **Format**	**Source**	**Role**
HASTPNFR	Sustained total pain and symptom free	Num	0 = No 1= Yes	International Headache Society Guidelines	Analysis

The Source column might additionally contain a list of source variables, such as those listed below. However, this example illustrates that this list could be quite lengthy. Note that each of the variables named in this example (e.g., HASEV0, HAPHOT) will also have variable metadata – for brevity not shown here – which will point back to variables on source datasets.

The Source could hyperlink to the supporting documentation, *International Headache Society Guidelines,* as follows.

Endpoint: Sustained total pain and symptom free (Yes/No)

Defined as:

- Headache severity of either Moderate or Severe at Baseline AND
- Headache severity of No Pain by 2 hours AND

- No headache recurrence within 48 hours AND
- No rescue medications for analgesia or antiemetic prior to 2 hours or for 48 hours post baseline AND
- No associated symptoms (nausea, vomiting, photophobia, phonophobia) from 2 through 48 hours.

Definitions:

Headache severity

Headache severity is subjectively rated by patients at pre-specified time points (baseline, 0.5, 1, 1.5, 2, 3, and 4 hours post-initial dose) on a scale from Grade 0 to 3:

Grade 0 - No pain;

Grade 1 - Mild pain;

Grade 2 - Moderate pain;

Grade 3 - Severe pain.

Variables (all Y/N): HASEV0, HASEV30, HASEV60, HASEV90, HASEV120, HASEV180, HASEV240 from Headache Efficacy Dataset; HASEV0 will be used in the endpoint derivation

Associated Symptoms

The patient will record whether the following associated symptoms were present or absent at regular time points:

Photophobia, Phonophobia, Nausea, Vomiting

Patients should be instructed to list any of the above symptoms as an “Adverse Symptom” on the diary card if it: (1) shows an unusual increase in intensity after they have taken their test medication or, (2) otherwise shows an important change in character after they have taken their test medication, as compared with their usual migraine symptoms. All such symptoms will be recorded by the investigator as adverse experiences.

Variables (all Y/N): HAPHOT, HAPHON, HANAU, HAVOM derived from AE Datataset

Headache Recurrence

Headache recurrence is defined as the return of headache to Grade 2 or 3 (moderate or severe) within 48 hours of the initial dose in patients who report pain relief at 2 hours after the initial treatment.

Variables (Y/N): HAREC48H from Headache Efficacy Dataset

Rescue Medications

The patient will record any additional analgesics/anti-emetics taken after any test dose, documenting date, clock time (AM/PM), name of drug (e.g., codeine), the number of tablets/capsules, and the dose per tablet/capsule.

Variables (all Y/N): HARESC derived from Concomitant Medications and Exposure Datasets

8.8 Representations and Warranties; Limitations of Liability, and Disclaimers

CDISC Patent Disclaimers. It is possible that implementation of and compliance with this standard may require use of subject matter covered by patent rights. By publication of this standard, no position is taken with respect to the existence or validity of any claim or of any patent rights in connection therewith. CDISC, including the CDISC Board of Directors, shall not be responsible for identifying patent claims for which a license may be required in order to implement this standard or for conducting inquiries into the legal validity or scope of those patents or patent claims that are brought to its attention.

Representations and Warranties. Each Participant in the development of this standard shall be deemed to represent, warrant, and covenant, at the time of a Contribution by such Participant (or by its Representative), that to the best of its knowledge and ability: (a) it holds or has the right to grant all relevant licenses to any of its Contributions in all jurisdictions or territories in which it holds relevant intellectual property rights; (b) there are no limits to the Participant's ability to make the grants, acknowledgments, and agreements herein; and (c) the Contribution does not subject any Contribution, Draft Standard, Final Standard, or implementations thereof, in whole or in part, to licensing obligations with additional restrictions or requirements inconsistent with those set forth in this Policy, or that would require any such Contribution, Final Standard, or implementation, in whole or in part, to be either: (i) disclosed or distributed in source code form; (ii) licensed for the purpose of making derivative works (other than as set forth in Section 4.2 of the CDISC Intellectual Property Policy ("the Policy")); or (iii) distributed at no charge, except as set forth in Sections 3, 5.1, and 4.2 of the Policy. If a Participant has knowledge that a Contribution made by any Participant or any other party may subject any Contribution, Draft Standard, Final Standard, or implementation, in whole or in part, to one or more of the licensing obligations listed in Section 9.3, such Participant shall give prompt notice of the same to the CDISC President who shall promptly notify all Participants.

No Other Warranties/Disclaimers. ALL PARTICIPANTS ACKNOWLEDGE THAT, EXCEPT AS PROVIDED UNDER SECTION 9.3 OF THE CDISC INTELLECTUAL PROPERTY POLICY, ALL DRAFT STANDARDS AND FINAL STANDARDS, AND ALL CONTRIBUTIONS TO FINAL STANDARDS AND DRAFT STANDARDS, ARE PROVIDED "AS IS" WITH NO WARRANTIES WHATSOEVER, WHETHER EXPRESS, IMPLIED, STATUTORY, OR OTHERWISE, AND THE PARTICIPANTS, REPRESENTATIVES, THE CDISC PRESIDENT, THE CDISC BOARD OF DIRECTORS, AND CDISC EXPRESSLY DISCLAIM ANY WARRANTY OF MERCHANTABILITY, NONINFRINGEMENT, FITNESS FOR ANY PARTICULAR OR INTENDED PURPOSE, OR ANY OTHER WARRANTY OTHERWISE ARISING OUT OF ANY PROPOSAL, FINAL STANDARDS OR DRAFT STANDARDS, OR CONTRIBUTION.

Limitation of Liability. IN NO EVENT WILL CDISC OR ANY OF ITS CONSTITUENT PARTS (INCLUDING, BUT NOT LIMITED TO, THE CDISC BOARD OF DIRECTORS, THE CDISC PRESIDENT, CDISC STAFF, AND CDISC MEMBERS) BE LIABLE TO ANY OTHER PERSON OR ENTITY FOR ANY LOSS OF PROFITS, LOSS OF USE, DIRECT, INDIRECT, INCIDENTAL, CONSEQUENTIAL, OR SPECIAL DAMAGES, WHETHER UNDER CONTRACT, TORT, WARRANTY, OR OTHERWISE, ARISING IN ANY WAY OUT OF THIS POLICY OR ANY RELATED AGREEMENT, WHETHER OR NOT SUCH PARTY HAD ADVANCE NOTICE OF THE POSSIBILITY OF SUCH DAMAGES.

Note: The CDISC Intellectual Property Policy can be found at
http://www.cdisc.org/about/bylaws_pdfs/CDISCIPPolicy-FINAL.pdf .