

How to write ADaM specifications like a ninja.

Caroline Francis, Independent SAS & Standards Consultant, Torrevieja, Spain

ABSTRACT

To produce analysis datasets from CDISC Study Data Tabulation Model (SDTM) datasets, best practice is to follow CDISC Analysis Data Model (ADaM) standards. Indeed, the FDA may request that you submit some ADaM datasets to support a submission. Writing these specifications can be a daunting task: what inputs do you need? What documents should you reference? How many datasets do I need? How should the data be structured? Does the input data need to be available? This paper attempts to answer these questions in a step by step, easy to follow fashion.

INTRODUCTION

In order to produce tables, figures and listings for a Clinical Study Report (CSR) it is advisable to produce analysis datasets. The pharmaceutical industry standard for this is the CDISC Analysis Dataset Model (ADaM). These are dataset structured in a standard way, which are analysis ready (an output should be able to be created straight from the dataset, with no additional manipulation). ADaM specifications detail how to create an ADaM dataset: How it is structured, what variables you need and how they are created. This paper assumes that you are well versed in the CDISC Standard Data Tabulation Model (SDTM) – the input dataset structure – and the ADaM standard. You should also have a basic grasp of data collection for clinical trials and the process through which data goes through to create the final outputs for the CSR.

ESSENTIAL READING

There are documents that contain all the information about the clinical study: how it is conducted, what assessments are involved, what data is collected and how it should be analyzed. In this section we will go through these documents one by one, discussing what information you should be looking for and how to apply it to writing ADaM specifications.

PROTOCOL

This document is the backbone of the study. It details how the study will be run, study timings, study assessments, how data will be collected and what the endpoints are. This document must absolutely be read before reading any other documents connected with the study. So what points should you be looking for in this document? Here are initial questions to be asked when reading the protocol. These may change when considering the Statistical Analysis Plan (SAP), output shells, CRF, SDTM specification, but considering these points is a good first step for crafting ADaM specifications:

- Blinding – will any data (apart from any randomization schema) be blinded until Database Lock (DBL)?
- How many subjects are there (what volume of data are you expecting=?)
- What is the duration of the study?
- Visit timings
 - Are these the same for all trial arms?
 - Are visits split into timepoints?
 - Are visits held across more than 1 day?
 - Are they grouped into study periods and are these likely to be used in the analysis?
- Primary, secondary and exploratory objectives:
 - What are they?
 - How are they to be measured?
 - Think about the datasets you will need: are there any corporate templates? Will you need to request a new standard? What structure will it be? What are the data sources? Are multiple sources required for one efficacy / safety / exploratory dataset?
 - Are any statistical tests specified in the protocol?
 - Are any tests specifically mentioned? If so, are they done at particular timepoints and is this important?
- Inclusion / exclusion criteria:
 - Will these need to be programmed?
 - Will these need to be considered when defining protocol deviations?
- Concomitant medications:

PhUSE EU Connect 2018

- Are any prohibited?
 - Are rescue medications defined?
 - Are there any medications that are not investigational medication products (IMP) that need to be included in the analysis?
- Is the study blinded? What steps are being taken to maintain the blind?
- Subject disposition:
 - Is this done at study level?
 - Treatment discontinuation?
 - How is this handled? Will a subject remain in the study if they have discontinued IMP?
 - What visit schedule is followed for subjects who discontinue early?
- Subject demographics and characteristics – into ADSL.
- What is the medical history / condition of these subjects?
 - Will they have an underlying disease?
 - Are they to have surgery? Additional diagnostic tests that affects their progress through the study?
- Subgroup analysis
 - These will probably also be in ADSL.
 - Check the SDTM origin of these data.
- Covariates
 - Consider if these will be from ADSL or is it more efficient if they are created in the efficacy dataset?
 - What is company policy for covariates?
- Exploratory analysis
 - How is this collected?
 - Are any specialist instruments used?
 - What are the company standards?
 - Is this similar enough to the primary efficacy that it will be included in the same ADaM or will an additional dataset be required?
- Is any data collected that may be analyzed by a different team, for example PK data? If so, does an ADPK dataset specification still required?

Often, it will take more than 1 reading of the protocol to get all the information you will need to answer these questions. Even if the study is being analyzed under compressed timelines, it is worth taking the time to analyze the protocol thoroughly as it saves re-work later on (for example, after an initial review of the datasets and specifications by the Statistician and others in the study team).

CRF

This is the Case Report Form (CRF) that Investigators use to input the data into the database. How this is set up will mirror the data collection modules in the database. It will also include code lists for questions with lists of answers that Investigators can choose from. Annotating the CRF with the SDTM domain is helpful if this has not been done when the SDTM mappings have been drafted. This will give you the format of the data as well as the anticipated values of the variables.

SDTM SPECIFICATION

This document contains the mappings from the raw data to the SDTM datasets. If the database is set up to be CDASH compliant then this mapping will be straightforward. In here will be the number and type of domains to be mapped, the variables that will be included, the mappings, code lists, any derivations (for example of Baseline) and the final structure of the dataset. This will be your reference document when writing the ADaM specifications as it will give you the variable names and values which are required for writing derivations. This is especially true for the more complicated SDTM domains such as Disposition (DS) and Laboratory test results (LB).

SAP

The Statistical Analysis Plan (SAP) expands on the analysis from the protocol. It gives much more detail to enable programmers to create the ADaM datasets and the final outputs. When reading the SAP, bear in mind the following points / questions. If the SAP is not clear then queries must be raised with the Statistician. The sooner this can be done, unnecessary rework and lost time can be avoided. The details that should be noted are:

- Analysis sets
 - Subject level flags in ADSL.
 - Any observation level flags will need to be added to the relevant datasets.
 - Pay attention to naming conventions. Values of the flags are Y or N and must be populated for all subjects / observations in a dataset.
 - Ensure all populations are defined, as described in the SAP.
- Disposition
 - Also in ADSL.

PhUSE EU Connect 2018

- ADaM Implementation Guide version 1.1 has naming conventions for these variables.
 - Ensure disposition for the subject in the study can be clearly distinguished from that of study treatment disposition.
 - Decide whether Screen Failure subjects will be included in ADaM and if so, which datasets they will be included in.
- Protocol Deviations
 - Define the dataset that will contain the deviations. This can be variables in ADSL or a separate dataset, depending on Sponsor choice.
 - Work with the Statistician and study team to define the deviations.
 - Use the CRF and SDTM specifications to explicitly define the programmable deviations.
 - Some deviations won't be able to be defined due to the way data is captured in the CRF: Decide how to incorporate deviations logged at study sites.
- Baseline
 - How is this defined for Safety?
 - Efficacy?
 - Is it the same for all measurements, or will this need to be defined differently across datasets?
 - Can the definition / flags from SDTM be used or does the SAP require a more complex derivation, for example of multiple observations?
 - Are multiple Baselines required, for example for multiple analysis periods?
- Missing data
 - How is this imputed for efficacy?
 - Is multiple imputation required? How will the dataset be structured for this?
 - How is this imputed for Safety?
 - Should dates be imputed? Times? Is this done for all data or only selected domains, such as Adverse Events or Concomitant Medications?
- Analysis time points
 - Should visit windowing be applied? To which datasets?
 - How are the analysis visits structured? Are time points to be considered?
 - Are APERIOD variables required? These should be consistent with the analysis periods in ADSL.
 - Are APHASE required? These are a less granular grouping of APERIOD. If a more granular grouping of APERIOD is required, then sub-period variables can be used.
- Efficacy models
 - Primary endpoints; derivations, imputations, timepoints, models, covariates, subgroups must all be considered and included in the efficacy dataset(s).
 - Sensitivity analysis of these primary endpoints – will these be included in the same ADaM dataset?
 - Definition of a responder vs a non-responder. Will this need to be a flag in ADSL as well as the efficacy ADaM?
 - Secondary endpoints and exploratory endpoints; are these similar to the primary endpoint? Included in the same ADaM or are additional datasets required?
 - How is the data captured for these endpoints? Will the ADaM dataset have a similar structure to the SDTM? Are multiple input SDTM required?
- Safety summaries
 - How complex are these summaries?
 - Are all safety data analyzed in the same way?
 - Are any imputations required?
 - What dataset structure is appropriate? ADaM Basic Data Structure (BDS), Occurrence Dataset Structure (OCCDS) or other?
 - Is coding or standardizing of tests required?
- Covariates
 - What SDTM datasets are they included in?
 - Should they be added to ADSL?
 - Are they to be included in all ADaM?
- Subgroups
 - How are these grouped?
 - Usually stored in ADSL, should they be copied to all other ADaM?

Note that if data is only to be listed with minimal manipulation of SDTM data (such as re-merging SUPPQUAL data or adding an analysis flag) then this can be done straight from SDTM without creating an ADaM dataset.

PhUSE EU Connect 2018

When deciding what datasets to create, it may be useful to find out what datasets are required for a submission. Agencies may only require a subset of datasets. However, for an analysis dataset submission to be considered ADaM compliant, it must contain an ADSL dataset.

TFL SHELLS

Usually the mock-ups of the outputs are created to give the programmer a template for the tables, figures and listings. These usually follow standard template, however, for some efficacy analyses new templates may be required.

These usually contain programming notes that may contain additional information not included in the SAP. This information needs to be included in the ADaM dataset specifications. This may include sort orders, parameters that must be included, selected coded variables. If this cannot be included in the ADaM specification itself, then it must be included in the Analysis Data Reviewer's Guide.

I strongly recommend annotating these shells with the dataset, any observation selection method, which variables are to be summarized / categorized / selected for the model. This not only ensures that ADaM has been created for all outputs and nothing has been missed but also will help the programming team to select the correct variables for the analysis, as created by the specification writer. This ensures that the protocol and SAP are interpreted correctly and consistently, giving high quality outputs to be used in the clinical study reports.

PUTTING THE ADAM SPECIFICATION TOGETHER

So now that you have analyzed all the input documentation, you are ready to craft your ADaM specifications like a ninja. It is helpful if the SDTM data is available, but a true sensei will be able to write the specifications without doing any programming or checking any data.

Most Sponsors will have a template for the ADaM specifications, usually structured to provide the metadata for the define.xml file. Follow this specification and make sure there is enough detail that there is no ambiguity. Use dataset and variable names, specific manipulations and actual variable values. Formats are not allowed to be used in SDTM nor ADaM data.

DATASETS

The define.xml requires metadata about the dataset itself. This includes ADaM dataset type (ADSL, BDS, OCCDS; other), name, description, primary key.

One of the ways to maintain traceability between SDTM and ADaM is to name the ADaM dataset after the SDTM input domain, such as ADAE for the analysis adverse event domain, especially if the ADaM is structured similarly to the SDTM, so OCCDS for ADAE which gives an occurrence structure to the SDTM AE events dataset.

VARIABLES

For each dataset you will need create the variables required for the analysis, as identified in the previous section.

The ADaM data structure will determine what variables are to be included in the dataset. For example, for BDS datasets a PARAMCD (parameter code), PARAM (parameter name) and AVAL/C (analysis value) are required. How these are named and created will be determined by the metadata gather from the input documents in the previous section. Follow Sponsor guidelines and standards as well as the CDISC ADaM standards to ensure compliance and traceability.

This paper does not go into the detail for the variables that are required for each dataset. Ensure you include identifier variables, timepoint (analysis visit) variables, and topic variables (coded terms or parameter names and analysis values) so that the dataset is structured correctly and contains all the detail identified in analyzing the input documents so that these ADaM datasets are analysis ready. Add Baseline, analysis periods, and analysis enabling variables as required / permitted for the dataset structure.

CODE LISTS

Often overlooked in specification creation, it is useful to create code lists for categorical or grouping variables such as subgroups, analysis categories, analysis visits or time points and analysis periods. This guides the programmer on how to populate variables and ensures consistency with the final outputs.

Remember to use the CDISC SDTM and ADaM standards as much as possible. Where code lists can be extended, ensure these are included in the metadata so that this is included in the define.xml and is therefore compliant with the CDISC standards.

Code lists are sometimes omitted due to time constraints, but experience has shown that if these are created upstream in the ADaM specification process it eliminates rework and the additional time this takes further on down the output creation process.

COMPUTATIONAL ALGORITHMS

For derivations or imputations that are required across multiple datasets, such as visit windowing or analysis period creation, or an imputation technique to be applied in more than one dataset, it is useful to create a computational algorithm. This can then be used as the basis for a macro that can be called in multiple dataset creation programs. Again, this should be structured following Sponsor guidelines to ensure the metadata is created correctly.

PhUSE EU Connect 2018

SPECIFICATION UPDATES

ADaM dataset specifications are constantly evolving. An ADaM specification ninja knows this and plans accordingly. As conversations with the Statistician and study team evolve, so does the analysis and the specification. Updating the specification should be controlled and changes logged so they are easily identified. This needs to be done in a clear and consistent way throughout the study. Some methods of doing this are:

- Putting updates in a different colour to highlight any differences.
- Dating and logging the updates in a dedicated part of the specification.
- Detailing them in a separate document or email to the programming team.
- Logging questions and decisions in a dedicated document that can be uploaded into the Trial Master File (TMF).

Whilst creating the ADaM datasets, data issues may be encountered. These should be logged, and a communication strategy agreed with the Data Management team. If certain errors are identified consistently then these should be discussed with Data Management and additional checks or study site retraining may be required.

CONCLUSION

To become an ADaM specification ninja, thorough understanding of the input documents: protocol, SAP, CRF, SDTM specifications, are required. Detailed preparation early in the process enables the ADaM specifications to be produced without necessarily seeing the input data. Close collaboration with the Statistician where uncertainty around the SAP or shells is identified enables the specification to be clarified earlier and the outputs to be created as required. This reduces rework closer to database lock, taking programming off the critical path as database lock approaches.

REFERENCES

CDISC ADaM model version 2.1.
CDISC ADaM Implementation Guide version 1.1.
CDISC ADaM OCCDS model version 1.0.
CDISC Define.xml model version 2.0.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Caroline Francis
Independent SAS and Standards Consultant
Torrevieja, Spain
Work Phone: +34 966 84 34 50
Email: sunshineprog@yahoo.co.uk

Brand and product names are trademarks of their respective companies.