# Constructive Practice of Generating ADaM Datasets

Amos Shu, AstraZeneca, Gaithersburg, MD
Charles Ling, AstraZeneca, Gaithersburg, MD

## ABSTRACT

Many programmers may not know there are actually three ADaM standard data structures because the latest ADaM Implementation Guide (ADaMIG) (v 1.1, 2016) still only describes two of them - the subject-level analysis dataset (ADSL) and the Basic Data Structure (BDS). A new separate document called "ADaM Data Structure for Occurrence Data" (released on 16 Fed 2016) describes the third standard data structure - the Occurrence Data Structure (OCCDS). Questions still remain as to what are the derived and collected variables in each dataset and how many ADaM datasets should be created for a study. It is virtually impossible to have a universal standard format for all ADaM datasets because of the different analysis needs of different therapeutic areas. There is also much more room for personal discretion in creating ADaM datasets than their corresponding SDTM datasets. ADaM datasets contain both source and derived data, so some variables may contain the same values, e.g. AVALC, VSORRES, and VSSTRESC in an ADVS dataset. This tends to confuse programmers, statisticians, and clinicians because they do not like or are not aware of traceability consideration in creation of ADaM datasets. It is also time consuming to go over such long documents as ADaM and ADaMIG. This paper provides a basic guide to generating ADaM datasets.

## INTRODUCTION

Analysis Data Model (ADaM) datasets are much more flexible than SDTM (Study Data Tabulation Model) datasets, but they are also more complex. Based on our observations over the years in the industry, and regardless of how long you have worked in the clinical biostatistical programming field, you'll probably still need to read and understand some CDISC and FDA standard documents [1, 2, 3, 4] before generating ADaM datasets. Some reasons include refreshing your memory on some seldom used details (usually the most common reason), or to get updates from FDA or CDISC requirements and standards, or to keep up with new publications or papers related to ADaM from statistical programming colleagues. ADaM datasets must also be analysis-ready and contain traceability between ADaM and SDTM. Reading those documents for most all of these reasons can be extremely tedious and time consuming. Hence, the programming process of generating ADaM datasets is not an easy task. Programmers are often frustrated from this process because these documents do not always have a clear or easily assessable answer for their specific study related questions. As many have pointed out, ADaM is not perfect and it does not always provide the necessary data that easily supports all necessary analyses for review. All ADaM datasets, however, must pass validation by using Pinnacle 21 (formerly as OpenCDISC) Validator before NDA submission.

Several factors will impact the quality of ADaM datasets – personal preference, fully understanding of the needs of the specific study requirements, the optimum number of analysis datasets to be generated, and the degree of self-sufficiency to allow analysis and review with minimum programming or data processing, i.e. one proc away or analysis-ready. This paper outlines the specific aspects to generate ADaM datasets to help programmers to achieve their ultimate tasks.

## READ THE BASIC DOCUMENTS

The following publications are basic standard documents, must be included in the reading list:

- Study Data Tabulation Model (SDTM) (v 1.4, 2014) – CDISC
- Study Data Tabulation Model Implementation Guide (SDTMIG) (v 3.2, 2013) – CDISC
- Analysis Data Model (ADaM) (v 2.1, 2009) - CDISC
- Analysis Data Model Implementation Guide (ADaMIG) (v 1.1, 2016) – CDISC
- ADaM Data Structure for Occurrence Data – CDISC
- ADaM Data Structure for Adverse Event Analysis – CDISC
- The ADaM Basic Data Structure for Time-to-Event Event Analyses – CDISC
- Analysis Data Model (ADaM) Examples in Commonly Used Statistical Analysis Methods - CDISC
- eSubmissions Standard Study Data – FDA
- Study Data Technical Conformance Guide – FDA

- CDER Common Data Standards Issues Document – FDA

There are also many ADaM standards specifically related to different therapeutic areas which are published by CDISC.

To have a solid understanding on a specific therapeutic area, you'll need to read other related publications in order to complete the study specific programming work. Reading those documents is extremely tedious and time consuming, but there is no shortcut to understanding the ADaM datasets; it must be done.

## THREE ADAM DATA STRUCTURES

There are three ADaM standard data structures: 1. The subject-level analysis dataset (ADSL), 2. The Basic Data Structure (BDS), and 3. The Occurrence Data Structure (OCCDS). The ADaMIG document only describes the first two ADaM standard data structures and the third one is in a separate document – "ADaM Data Structure for Occurrence Data". This could easily mislead readers. Our hope is that for the next ADAMIG version, all three of these documents will reside in one place.

### 1. GENERATE THE SUBJECT-LEVEL ANALYSIS DATASET (ADSL)

The structure of subject-level analysis dataset (ADSL) contains one record per subject regardless of the type of clinical trial design. This is the golden rule that every programmer must follow. There are cases programmers break this rule by creating more than one record per subject. This usually leads to horrible results.

Each study must have only one ADSL as a starting point in terms of the ADaM build. ADSL is a central location for important variables that describe a subject's experience in the trial, which can then be used in all types of analysis. For example, the variable SAFFL (Safety Population Flag) would be used for a variety of efficacy analyses and safety analyses.

In ADaMIG v1.0, the standard variables that are required to be in every ADSL fall into five categories. The ADaMIG v1.1 not only includes two more categories – Dose Variables and Subject-Level Trial Experience Variables, but also brings more group variables such as REGIONy, AGEGRy, and TSEQPGy, etc., which are certainly helpful for subgroup analysis.

Here is the comparison of the two ADaM versions (v1.0 & v1.1) in term of ADSL. Some new group variables are listed here as well.

| V1.0 (2009-12-17) | V1.1 (2016-02-12) |
|---|---|
| ADSL Variable Category<br>1. Study Identifiers<br>2. Subject Demographics<br>3. Population Indicators<br>4. Treatment Variables<br>5. Trial dates | ADSL Variable Category<br>1. Study Identifiers<br>• REGIONy, REGIONyN<br>2. Subject Demographics<br>• AGEGRy, AGEGRyN, AAGE<br>3. Population Indicators<br>4. Treatment Variables<br>• ACTARM<br>• TSEQPGy, TSEQPGyN, TSEQAGy, TSEQAGyN<br>5. Dose Variables<br>• DOSExxP, DOSExxA, DOSExxU<br>6. Treatment Timing Variables (including Subject-Level Period, Subperiod, and Phase Timing Variables)<br>• PxxSw, etc.<br>• APHASEw<br>• PHwSDT, etc.<br>7. Subject-Level Trial Experience Variables<br>• EOSSTT, EOSDT, DCSREAS, DCSREASP<br>• EOTSTT, DCTREAS, DCTREASP<br>• EOTxxSTT, EOPxxSTT, etc.<br>• RFICDT, ENRLDT, RFICyDT, etc.<br>• LSTALVDT<br>• TRCMP, etc.<br>• TRxxDURD, etc.<br>• DTHDT, etc. |

In real practice, some programmers like to include some vital signs variables with their values at baseline, such as height, weight, BMI, and BSA in ADSL. We personally do not recommend this way because those variables are only used for a few of tables and figures. We prefer to have all vital signs variables in one place – ADVS.

The new version includes some trial experience variables such as DCSREAS ("Reason for Discontinuation from Study") and DCTREAS ("Reason for Discontinuation of Treatment"). Keeping these variables in ADSL, would save time and reduce the number of ADaM datasets needed for a simple phase 1 study. For those complicated studies, however, it may be better to have those variables added into a new ADaM dataset such as ADDS to avoid too many variables in ADSL.

Some programmers like to have the numeric version of flag variables (e.g. "1"/"0" for "Yes"/"No" values). These numeric variables actually do not add any value to the dataset or programming process.

Besides those required variables, how many conditionally required variables and permissible variables that should be added into the ADSL dataset, depend on the actual needs (e.g. how many unique tables/figures, what are their structure, etc.) of the specific study? We prefer to keep these variables that will apply for all types of tables and figures in the ADSL.

ADSL will be used to merge with some SDTM datasets to generate other ADaM datasets. A word of caution is that ADSL is not a dataset that you want to over populate with variables. An ADSL dataset with many variables (e.g. 50+) does not necessarily bring convenience to your programming work and data review. It is important to keep each ADaM dataset as lean as possible and have only the necessary variables that describe attributes/experience of a subject.

Please do not forget that the ultimate goal for the creation of ADaM datasets is to facilitate the regulatory review of new drug applications. All programmers need to put themselves into the regulatory statisticians' shoes. Doing so will help get your NDA/BLA reviewed by the FDA quickly.

## 2. GENERATE THE BASIC DATA STRUCTURE (BDS) DATASETS

This type of ADaM dataset must have the variables PARAM (Parameter) and AVAL (Analysis Value) and/or AVALC (Analysis Value (C)). The difference between ADSL and BDS is that a subject may have one or more records in the BDS dataset. Their predecessors are the SDTM findings class datasets. Here is the list of some of them:

- ADDA (Drug Accountability Analysis Dataset)
- ADEG (ECG Test Results Analysis Dataset)
- ADIE (Inc./Exc. Criterion Not Met Analysis Dataset)
- ADLB (Laboratory Test Results Analysis Dataset)
- ADPE (Physical Examination Analysis Dataset)
- ADQS (Questionnaires Analysis Dataset)
- ADSC (Subject Characteristics Analysis Dataset)
- ADVS (Vital Signs Analysis Dataset)
- ADPC (PK Concentrations Analysis Dataset)
- ADPP (PK Parameters Analysis Dataset)
- ADMB (Microbiology Specimen Analysis Dataset)
- ADMS (Microbiology Susceptibility Test Analysis Dataset)
- ADFA (Findings About Analysis Dataset)

All SDTM findings class datasets have a similar structure (see blow SDTM Finding Class datasets). In addition to the three common variables – STUDYID, DOMAIN, and USUBJID, they all have eight similar variables – XXSEQ, XXTESTCD, XXTEST, XXCAT, XXSCAT, XXORRES, XXSTRESC, and XXDTC. The 'XX' in the variable names stands for the domain name (e.g. VS) of a SDTM findings class dataset.

| Findings Class | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **IE** | **PE** | **EG** | **LB** | **VS** | **QS** | **MB** | **MS** | **PC** | **PP** | **SC** | **DA** | **FA** |
| STUDYID | STUDYID | STUDYID | STUDYID | STUDYID | STUDYID | STUDYID | STUDYID | STUDYID | STUDYID | STUDYID | STUDYID | STUDYID |
| DOMAIN | DOMAIN | DOMAIN | DOMAIN | DOMAIN | DOMAIN | DOMAIN | DOMAIN | DOMAIN | DOMAIN | DOMAIN | DOMAIN | DOMAIN |
| USUBJID | USUBJID | USUBJID | USUBJID | USUBJID | USUBJID | USUBJID | USUBJID | USUBJID | USUBJID | USUBJID | USUBJID | USUBJID |
| IESEQ | PESEQ | EGSEQ | LBSEQ | VSSEQ | QSSEQ | MBSEQ | MSSEQ | PCSEQ | PPSEQ | SCSEQ | DASEQ | FASEQ |
| IETESTCD | PETESTCD | EGTESTCD | LBTESTCD | VSTESTCD | QSTESTCD | MBTESTCD | MSTESTCD | PCTESTCD | PPTESTCD | SCTESTCD | DATESTCD | FATESTCD |
| IETEST | PETEST | EGTEST | LBTEST | VSTEST | QSTEST | MBTEST | MSTEST | PCTEST | PPTEST | SCTEST | DATEST | FATEST |
| IECAT | PECAT | EGCAT | LBCAT | VSCAT | QSCAT | MBCAT | MSCAT | PCCAT | PPCAT | SCCAT | DACAT | FACAT |
| IESCAT | PESCAT | EGSCAT | LBSCAT | VSSCAT | QSSCAT | MBSCAT | MSSCAT | PCSCAT | PPSCAT | SCSCAT | DASCAT | FASCAT |
| IEORRES | PEORRES | EGORRES | LBORRES | VSORRES | QSORRES | MBORRES | MSORRES | PCORRES | PPORRES | SCORRES | DAORRES | FAORRES |
| IESTRESC | PESTRESC | EGSTRESC | LBSTRESC | VSSTRESC | QSSTRESC | MBSTRESC | MSSTRESC | PCSTRESC | PPSTRESC | SCSTRESC | DASTRESC | FASTRESC |
| | PEDTC | EGDTC | LBDTC | VSDTC | QSDTC | MBDTC | MSDTC | PCDTC | PPDTC | SCDTC | DADTC | FADTC |

XXTESTCD and XXTEST are used to generate variables PARAMCD and PARAM, respectively. XXORRES and XXSTRESC are used to generate variable AVAL or AVALC. ADT will usually be derived from XXDTC. Most SDTM findings class domains have variables – XXSTRESN, XXBLFL, VISITNUM, and VISIT. If XXSTRESN is available, Variable AVAL can be generated from XXSTRESN. Variables ABLFL, AVISITN, and AVISIT are normally derived from XXBLFL, VISITNUM, and VISIT correspondently. Then AVAL, ABLFL, AVISITN, and AVISIT can be used to generate values for variables - BASE and CHG, etc.

The majority of other variables in this type of ADaM dataset are directly copied from ADSL and its corresponding SDTM dataset. Some other ADaM variables mentioned in ADaMIG may be also needed depending upon the analysis needs of a specific study. For example, vital signs shift tables require variables – ANRIND (Analysis Reference range Indicator) and BNRIND (baseline Reference range Indicator). So, it is quite straightforward to generate the BDS type of ADaM datasets.

| **SDTM Variables** | **ADaM Variables** |
|---|---|
| XXTESTCD | PARAMCD |
| XXTEST | PARAM |
| XXORRES or XXSTRESN | AVAL or AVALC |
| XXSTRESC | AVALC |
| XXDTC | ADT |
| XXBLFL | ABLFL |
| VISITNUM | AVISITN |
| VISIT | AVISIT |

There are no major differences between ADaMIG v1.0 and ADaMIG v1.1. Most of the differences are caused by the newly added ADSL variables in v1.1. The category “Lab Related Analysis variables” in v1.0 has been changed to “Toxicity and Range Variables” in v1.1. Here is the comparison of the BDS parts in the two ADaM versions. Again, some new group variables are listed here as well.

| V1.0 (2009-12-17) | V1.1 (2016-02-12) |
|---|---|
| BDS<br>1. Study Identifiers<br>2. Treatment Variables<br>3. Timing Variables<br>4. Analysis Parameter Variables<br>5. Analysis Descriptor Variables<br>6. Time-to-Event Variables<br>7. Lab Related Analysis variables<br>8. Indicator Variables<br>9. Datapoint Traceability Variables<br>10. Analysis-Enabling Variables | BDS<br>1. Study Identifiers<br>• ASEQ<br>2. Record-Level Treatment and Dose Variables<br>• DOSEP, DOSCUMP, DOSEA, DOSCUMA, DOSEU<br>3. Timing Variables (including Subject-Level Period, Subperiod, and Phase Timing Variables)<br>• APHASEN<br>• ASPER, ASPERC<br>4. Analysis Parameter Variables<br>• CRITyFL, etc.<br>• MCRITy, etc.<br>5. Analysis Descriptor Variables<br>6. Time-to-Event Variables<br>• STARTDTM, etc.<br>• CNSDTDSC<br>7. Toxicity and Range Variables<br>• ANRLOC, ANRHIC, AyLOC, etc.<br>8. Indicator Variables<br>9. Datapoint Traceability Variables<br>10. Analysis-Enabling Variables |

Some programmers, statisticians, and clinicians do not like the BDS, which vertically presents the data. Instead, they like to have all tests in one row within the same visit for a subject in datasets such as ADLB and ADVS. We totally understand the convenience for them but the essential reason to follow the ADaM structure is that you want to have your NDA/BLA reviewed by the FDA as quickly as possible. The quicker the review and approval, the earlier your approved drug will reach the market obtaining a winning competitive advantage, and can begin to generate revenue. If the sales from the drug are 365 million dollars a year that means a one million dollars lose for every one day delay pending approval. Can you imagine how much loses there will be in just one day for a blockbuster drug? Anyway, there is no perfect world.

## 3. GENERATE THE OCCURRENCE DATA STRUCTURE (OCCDS) DATASETS

In contrast to the BDS datasets, the OCCDS type of ADaM Datasets do not have variables PARAM (Parameter) and AVAL (Analysis Value) and/or AVALC (Analysis Value (C)) because they are from the SDTM interventions class and events class datasets. When using these SDTM datasets to generate ADaM datasets, they do not fit in well with the BDS structure and are more appropriately analyzed using their SDTM structure with added analysis variables [2]. Here are the reasons from the introduction section of the ADaM document titled “Analysis Data Model (ADaM) Data Structure for Adverse Event Analysis” (version 1.0, 2012):

- There is no need for AVAL or AVALC. Occurrences are counted in analysis, and there are typically one or more records for each occurrence.
- A dictionary is used for coding the occurrence, and it includes a well-structured hierarchy of categories and terminology. Mapping this hierarchy to BDS variables PARAM and generic *CAT variables would lose the structure and meaning of the dictionary.
- Dictionary content is typically not modified for analysis purposes. In other words, there is no need for analysis versions of the dictionary hierarchy.

The most common OCCDS datasets are ADAE, ADCM, ADMH, ADEX, ADCE, and ADDS. Most of the variables in this type of datasets come directly from the ADSL and its corresponding SDTM datasets such as AE, CM, MH, etc. Only a few variables that need to be created, e. g., the following numeric variables AESEVN, ARELN, AEACNN, and AEOUTN, timing variables such as ASTDT, AENDT, ASTDY, AENDY, and ADURN, and a flag variable TRTEMFL, need to be created from SDTM.AE and ADSL. These variable are extremely helpful in increasing programming efficiencies.

| Category | Variable Name | Variable Label | Source/Derivation |
|---|---|---|---|
| Analysis Descriptive Variables | AESEVN | Severity/Intensity (N) | Derived from AE.AESEV |
| | AERELN | Causality (N) | Derived from AE.AEREL |
| | AEACNN | Action Taken with Study Treatment (N) | Derived from AE.AEACN |
| | AEOUTN | Outcome of Adverse Event (N) | Derived from AE.AEOUT |
| | AETOXGRN | Analysis Toxicity Grade (N) | Derived from AE.AETOXGR |
| Analysis Timing Variables | ASTDT | Analysis Start Date | Derived from AE.ASTDT |
| | AENDT | Analysis End Date | Derived from AE.AENDT |
| | ASTDY | Analysis Start Relative Day | AE.AESTDY or use the following formula to derive ASTDT - ADSL.TRTSDT + 1 if ASTDT >= TRTSDT, else ASTDT - ADSL.TRTSDT if ASTDT < TRTSDT |
| | AENDY | Analysis End Relative Day | AE.AEENDY or use the following formula to derive AENDT - ADSL.TRTSDT + 1 if AENDT >= TRTSDT, else AENDT - ADSL.TRTSDT if AENDT < TRTSDT |
| | ADURN | AE Duration (N) | Derived from ASTSDT and AENDT |
| Indicator Variables | TRTEMFL | Treatment-Emergent Analysis Flag | Derived |

## HOW MANY ADAM DATASETS ARE NEEDED

ADaM is optimized to support data derivation and analysis. So, besides the unique ADSL, how many other ADaM datasets with the BDS and OCCDS structures are needed? The answer depends on how many tables and figures are needed and what are their structures (shells).

Not every SDTM dataset needs a corresponding ADaM dataset. For programming efficiency, we need to optimize the number of ADaM datasets. In general, the number of ADaM datasets should be much smaller than the number of SDTM datasets. Some programmers like to create ADaM datasets from the SDTM special-purpose class datasets such as ADDM. We do not recommend this practice because most of the SDTM special-purpose class datasets contain a few variables. Although SDTM.DM has 20 variables, most of these variables also exist in ADSL. It is not necessary to have a separated ADDM even in this case.

## HOW MANY VARIABLES ARE NEEDED IN AN ADAM DATASET

How many variables should an ADaM dataset contain? In other words, how many variables are enough to support the efficient generation, replication, and review of analysis results? If the primary programmer, validation programmer, and study statisticians all put themselves in the FDA statisticians' shoes, they will have the best answer. Some variables, such as those directly taken from SDTM datasets, provide traceability between the analysis data and its source data (ultimately SDTM). Some derived variables facilitate clear and unambiguous communication of the scientific and statistical aspects of the trial. Variables that will be used for multiple tables/figures will stay within their respective ADaM datasets. Variables that are used to generate only one table or figure, may be not necessarily need to be created in the ADaM dataset depending on the programming efficiency. It may be more convenient to have this variable derived within the table or figure program.

We also need to consider the size issue of a dataset. The more variables a dataset contains, the bigger its' size. Datasets with very large size, may not be accepted by the FDA IT storage system.

## CONCLUSION

BDS and OCCDS datasets discussed above are interchangeable in order, depending on your preference and actual programming needs. The documents - Analysis Data Model (ADaM) (v 2.1, 2009), Analysis Data Model Implementation Guide (ADaMIG) (v 1.1, 2016), and ADaM Data Structure for Occurrence Data, are good reference

documents as an overview of the process, but the key to generating ADaM datasets is the actual needs of the specific study, traceability, and statistical programming and analysis efficiency.

## REFERENCES

[1]. http://www.cdisc.org/standards
[2]. http://www.cdisc.org/stuff/contentmgr/files/0/5aee16f59e8d6bd2083dbb5c1639f224/misc/adam_ae_final_v1.pdf
[3]. http://www.fda.gov/Drugs/default.htm
[4]. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM292334.pdf

## ACKNOWLEDGEMENTS

Special thanks to Tracy Turschman.

## CONTACT INFORMATION

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

Amos Shu
AstraZeneca Pharmaceuticals, LP
200 Orchard Ridge Dr.
Gaithersburg, MD 20878
Email: amos.shu@astrazeneca.com

## TRADEMARK INFORMATION