TDF_ADaM: ADaMIG v1.1 Test Datasets

Test Data Factory Project Team

November 2018

1 Introduction and Background

1.1 General Approach

The PhUSE organization initiated the Test Data Factory (TDF) project to create CDISC-compliant test datasets and make these test datasets publicly available. The goal of the project is not to create complete submission packages, but it will focus on datasets that comply with CDISC standards (focusing on SDTM, ADaM, and SEND) and that can be used to support testing of CDISC-based processes and programs. In general, the TDF project team will use the following iterative approach:

- For the datasets under development the team will run the validation tool of Pinnacle 21 Community v2.2 using the appropriate settings for the standard version.
- Each individual dataset is then updated based on the reported findings so that no more unexplained findings are reported by the validation tool.
- The team will then decide which errors and warning should be addressed and which should be left untouched and should be explained in the documentation for the dataset package.
    - Note that there could be different possible reasons for leaving errors and warning in the dataset: For example, a certain type of issue could be be seen in real world datasets as well or the issue is caused by some inefficiency of the validation tool.
- As a regression test, Pinnacle 21 Community v2.2 is used on updated datasets at the same time to find additional issues resulting from inconsistencies between individual datasets. The datasets are then updated as needed.
- For updating the datasets, the team will use SAS and R programs as appropriate.
- Finally, the test datasets will be published as a package. The objective is not to create a complete submission package but rather a set of files that can be used for testing and that includes sufficient information for the user to decide how to use the datasets. Typically, a package will include the following files:
    - The datasets as .xpt files
    - A define.xml and corresponding .xsl stylesheet that transforms the define.xml file into an html page for easy reading and navigating.
    - If applicable supporting documents as required and deemed necessary by the TDF team.
    - A Word document (like this file) that describes the datasets, the process, and explains remaining issues reported by the validation tool.
    - A final report from the validation tool

The team used Pinnacle 21 as a conformance checking tool because it is a commonly used tool to evaluate conformance and is openly available to everyone. The team is very well aware of shortcomings and knows that this tool is not perfect and comprehensive but decided that for lack of better alternatives, this would be the right choice.

1.2 CDISCPILOT02 Study and ADaM Datasets

The CDISC organization published the CDISCPILOT02 study as an example and test case in earlier development phases of the CDISC standards. The original CDISCPILOT02 data may be downloaded from https://www.cdisc.org/sdtmadam-pilot-project, if desired. For the initial phase of the TDF project,

the team decided to use the existing SDTM and ADaM datasets of the CDISCPILOT02 study from 2013 as a starting point for its work to gain experience with the process and to be able to deliver some results more quickly. This document describes the SDTM datasets from the CDISCPILOT02 that were updated to conform to ADaM 1.0 as the Configuration and 2016-06-24 as the CDISC Controlled Terminology. Note the remarks in section 3 about conformance issues between ADaMIG 1.0 and ADaMIG 1.1.

In order to resolve some of the issues found in the original datasets, the team needed to make some decisions on how the datasets should be updated. These changes should not be construed as definitive approaches for resolving conformance issues in general. Users of the TDF_ADaM datasets should be aware that depending on the situation, several options can and need to be considered.

Note that the rather large ADLBC dataset was split (see section 2.1) to accommodate limitations of the repository that was used during the work. For consistency, the team decided to split the ADLBH as well using the same rules. This was not done because of any guidance or other requirement to split these domains. Three datasets were renamed to follow CDISC guidance (see section 2.2).

In section 2.4, some files are listed that were not updated but copied from the original CDISCPILOT02 submission package. Users should be aware that the content of these files (typically .pdf files) is likely not consistent with the data in the TDF_ADaM datasets because they were created with the original CDISCPILOT02 datasets. In addition, users need to understand that the location of these files in a subfolder names “cdisc_docs” does not conform with submission requirements.

Section 3 of this document provides explanations for the Pinnacle 21 findings that are still seen in the report. Some of these findings cannot be fully addressed because information on how the data was originally collected was not included in the pilot materials.

Note that this document is not intended to represent or even resemble a Study Data Reviewer’s Guide (SDRG). Instead, its purpose is to serve as documentation for the updated test data, and to explain some of the decisions that were made during the update process. Similarly, the TDF_SDTM test datasets are intended to be used for the development and testing of standard reporting and analysis scripts and are not meant to represent a complete regulatory submission package.

2 Specific Comments

2.1 Split datasets

The ADLBC and ADLBH datasets were split based on the values of the indicated variable. Note that this splitting was done to reduce the size of the resulting datasets and to demonstrate split datasets and not because of any guidance or other requirement to split these domains.

Split Dataset	Value of split Variable	Comment about Split Variable
adlbc.xpt	Preceding "_" in PARAMCD	The dataset contains two sets of observations, one is marked as "change from previous visit, relative to normal range" and is marked in PARAMCD using a preceding "_". Using the PARAMCD value the dataset was split into adlbc.xpt and adlbcpv.xpt.
adlbh.xpt	Preceding "_" in PARAMCD	The dataset contains two sets of observations, one is marked as "change from previous visit, relative to normal range" and is marked in PARAMCD using a preceding "_". Using the PARAMCD value the dataset was split into adlbh.xpt and adlbhpv.xpt.

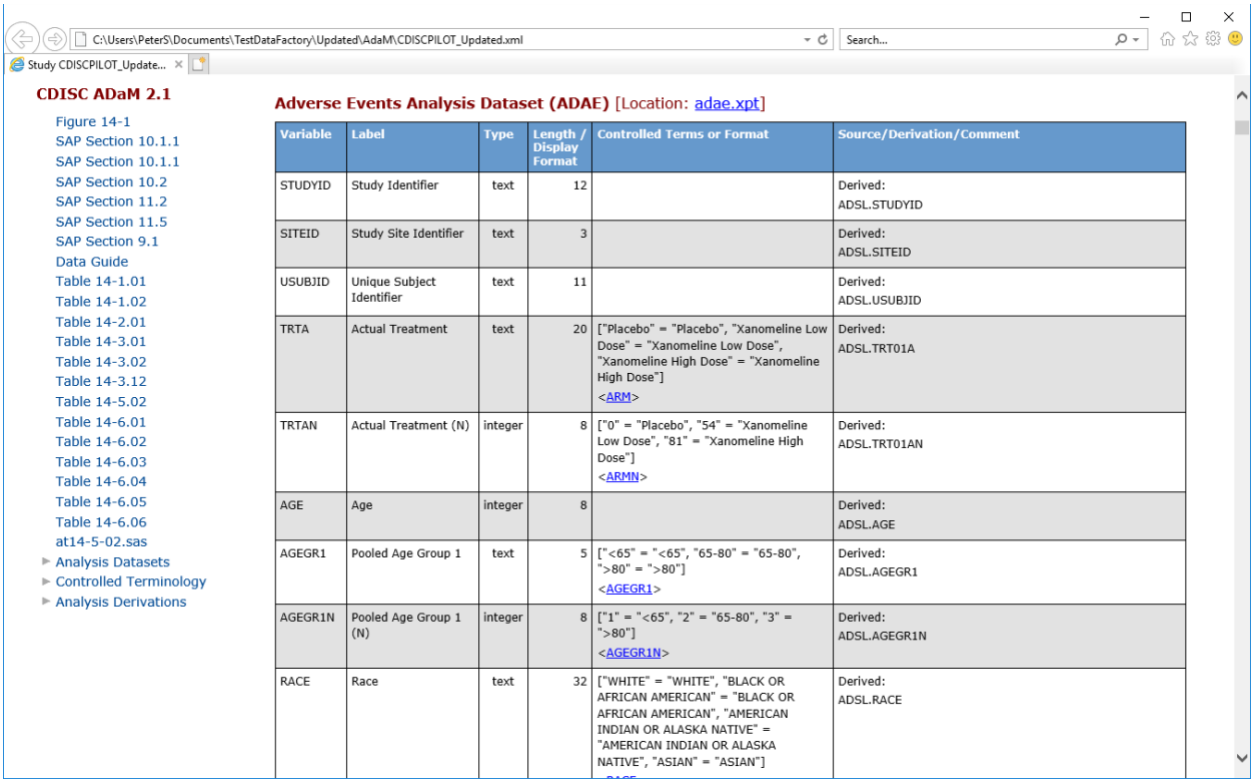
2.2 Renamed Datasets

The following table shows a list of datasets that were renamed to confirm to CDISC guidance.

Original Dataset Name	New Dataset Name
adqscibc.xpt	adcibc.xpt
adqsadas.xpt	adadas.xpt
adqsnpix.xpt	adnpix.xpt

2.3 Missing Information in define.xml

An updated define.xml document is included with the updated ADaM datasets. This define.xml document has been created using the 'Generate Define' feature of the Pinnacle 21 Community edition from an Excel specification document. Note that there is incomplete information in the define.xml file. More specifically, Source/Derivation/Comment information is not included for all datasets and variables. The TDF team believes that the information included in the define.xml is sufficient to show how this information would be provided.



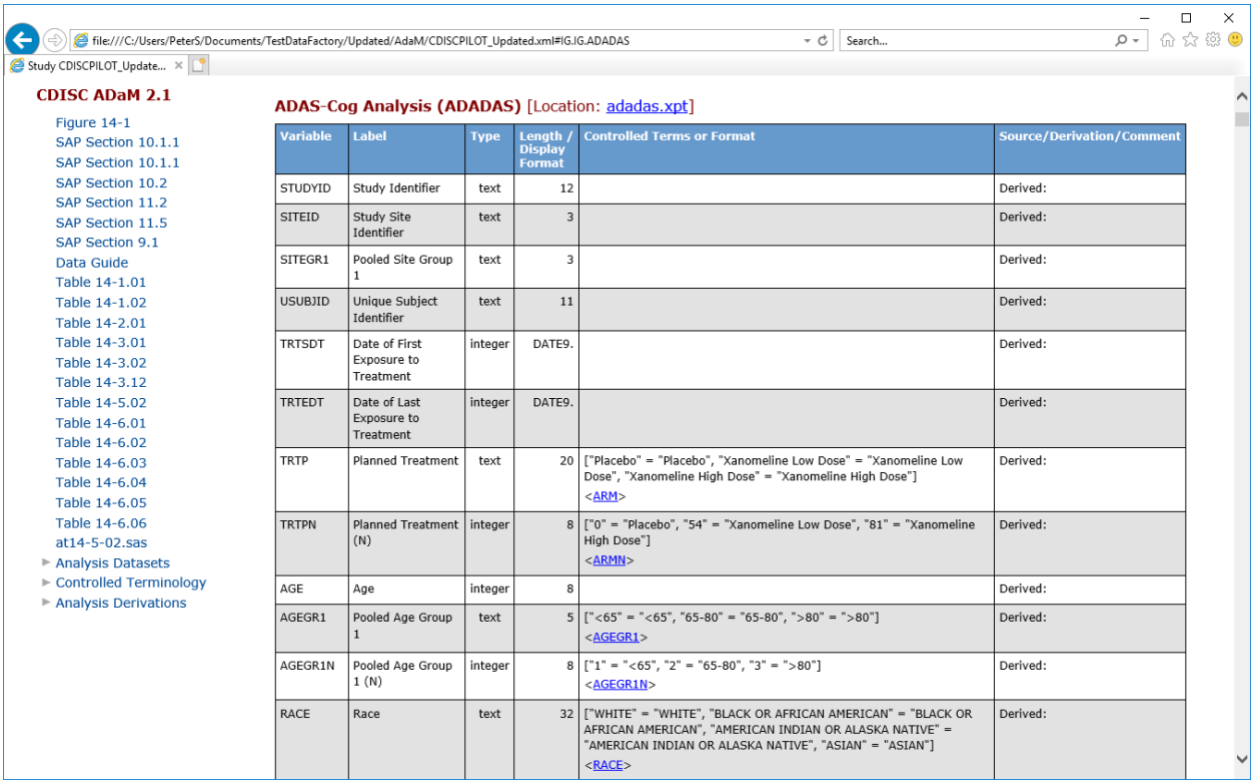
CDISC ADaM 2.1

- Figure 14-1
- SAP Section 10.1.1
- SAP Section 10.1.1
- SAP Section 10.2
- SAP Section 11.2
- SAP Section 11.5
- SAP Section 9.1
- Data Guide
- Table 14-1.01
- Table 14-1.02
- Table 14-2.01
- Table 14-3.01
- Table 14-3.02
- Table 14-3.12
- Table 14-5.02
- Table 14-6.01
- Table 14-6.02
- Table 14-6.03
- Table 14-6.04
- Table 14-6.05
- Table 14-6.06
- at14-5-02.sas
- Analysis Datasets
- Controlled Terminology
- Analysis Derivations

Adverse Events Analysis Dataset (ADAE) [Location: adae.xpt]

Variable	Label	Type	Length / Display Format	Controlled Terms or Format	Source/Derivation/Comment
STUDYID	Study Identifier	text	12		Derived: ADSL.STUDYID
SITEID	Study Site Identifier	text	3		Derived: ADSL.SITEID
USUBJID	Unique Subject Identifier	text	11		Derived: ADSL.USUBJID
TRTA	Actual Treatment	text	20	["Placebo" = "Placebo", "Xanomeline Low Dose" = "Xanomeline Low Dose", "Xanomeline High Dose" = "Xanomeline High Dose"] <ARM>	Derived: ADSL.TRT01A
TRTAN	Actual Treatment (N)	integer	8	["0" = "Placebo", "54" = "Xanomeline Low Dose", "81" = "Xanomeline High Dose"] <ARMN>	Derived: ADSL.TRT01AN
AGE	Age	integer	8		Derived: ADSL.AGE
AGEGR1	Pooled Age Group 1	text	5	["<65" = "<65", "65-80" = "65-80", ">80" = ">80"] <AGEGR1>	Derived: ADSL.AGEGR1
AGEGR1N	Pooled Age Group 1 (N)	integer	8	["1" = "<65", "2" = "65-80", "3" = ">80"] <AGEGR1N>	Derived: ADSL.AGEGR1N
RACE	Race	text	32	["WHITE" = "WHITE", "BLACK OR AFRICAN AMERICAN" = "BLACK OR AFRICAN AMERICAN", "AMERICAN INDIAN OR ALASKA NATIVE" = "AMERICAN INDIAN OR ALASKA NATIVE", "ASIAN" = "ASIAN"]	Derived: ADSL.RACE

Screenshot of ADAE.xpt details with Source/Derivation/Comment information included



CDISC ADaM 2.1

- Figure 14-1
- SAP Section 10.1.1
- SAP Section 10.1.1
- SAP Section 10.2
- SAP Section 11.2
- SAP Section 11.5
- SAP Section 9.1
- Data Guide
- Table 14-1.01
- Table 14-1.02
- Table 14-2.01
- Table 14-3.01
- Table 14-3.02
- Table 14-3.12
- Table 14-5.02
- Table 14-6.01
- Table 14-6.02
- Table 14-6.03
- Table 14-6.04
- Table 14-6.05
- Table 14-6.06
- at14-5-02.sas
- Analysis Datasets
- Controlled Terminology
- Analysis Derivations

ADAS-Cog Analysis (ADADAS) [Location: adadas.xpt]

Variable	Label	Type	Length / Display Format	Controlled Terms or Format	Source/Derivation/Comment
STUDYID	Study Identifier	text	12		Derived:
SITEID	Study Site Identifier	text	3		Derived:
SITEGR1	Pooled Site Group 1	text	3		Derived:
USUBJID	Unique Subject Identifier	text	11		Derived:
TRTSDT	Date of First Exposure to Treatment	integer	DATE9.		Derived:
TRTEDT	Date of Last Exposure to Treatment	integer	DATE9.		Derived:
TRTP	Planned Treatment	text	20	["Placebo" = "Placebo", "Xanomeline Low Dose" = "Xanomeline Low Dose", "Xanomeline High Dose" = "Xanomeline High Dose"] <ARM>	Derived:
TRTPN	Planned Treatment (N)	integer	8	["0" = "Placebo", "54" = "Xanomeline Low Dose", "81" = "Xanomeline High Dose"] <ARMN>	Derived:
AGE	Age	integer	8		Derived:
AGEGR1	Pooled Age Group 1	text	5	["<65" = "<65", "65-80" = "65-80", ">80" = ">80"] <AGEGR1>	Derived:
AGEGR1N	Pooled Age Group 1 (N)	integer	8	["1" = "<65", "2" = "65-80", "3" = ">80"] <AGEGR1N>	Derived:
RACE	Race	text	32	["WHITE" = "WHITE", "BLACK OR AFRICAN AMERICAN" = "BLACK OR AFRICAN AMERICAN", "AMERICAN INDIAN OR ALASKA NATIVE" = "AMERICAN INDIAN OR ALASKA NATIVE", "ASIAN" = "ASIAN"] <RACE>	Derived:

Screenshot of ADADAS.xpt details where Source/Derivation/Comment information is missing

2.4 CDISCPILOT02 files, that are not updated

The project team decided that certain files that are included in the CDISCPILOT02 should be included in the TDF_ADaM package but do not need to be updated. These files are added to satisfy links in the define.xml but the content of these files might not be consistent with the updated datasets. These files are placed in a subfolder named “cdisc_docs” and include a dataguide document, the cdiscpilot02 study report, and a SAS program. Note that links in these files might be broken because the team decided not to include all documents that are included in the CDISCPILOT02 submission package.

3 Data Conformance Summary – Explanation on Remaining Pinnacle 21 Findings

The following table explains the remaining warnings from a Pinnacle 21 Community Edition validation. This list represents the status after the update of the define.xml file and is included so that users of the updated CDISC Pilot dataset can understand the extent of changes that were applied to the datasets. A final validation report from Pinnacle 21 Community Edition is included with the datasets or can easily be created by running the tool against the datasets.

Dataset	Variables	Values	Message	Severity	Explanation
ADADAS	AWU, AWLO, AWHI	DAYS, null, 1 or DAYS, 141, null	AWU is populated but both AWLO and AWHI are not populated	Error	A one-sided window is incorrectly marked by the current Pinnacle 21 version as an error.
ADLBC	VARIABLE, LABEL	ANL01FL, Analysis Record Flag 1	Variable label mismatch between dataset and ADaM standard	Error	Pinnacle 21 Community Edition validation uses ADaMIGv1.0. The value for the variable label conforms to ADaMIGv1.1 as required.
ADLBCPV	VARIABLE, LABEL	ANL01FL, Analysis Record Flag 1	Variable label mismatch between dataset and ADaM standard	Error	Pinnacle 21 Community Edition validation uses ADaMIGv1.0. The value for the variable label conforms to ADaMIGv1.1 as required.
ADLBH	VARIABLE, LABEL	ANL01FL, Analysis Record Flag 1	Variable label mismatch between dataset and ADaM standard	Error	Pinnacle 21 Community Edition validation uses ADaMIGv1.0. The value for the variable label conforms to ADaMIGv1.1 as required.
ADLBHPV	VARIABLE, LABEL	ANL01FL, Analysis Record Flag 1	Variable label mismatch between dataset and ADaM standard	Error	Pinnacle 21 Community Edition validation uses ADaMIGv1.0. The value for the variable label conforms to ADaMIGv1.1 as required.
ADNPIX	AWU, AWLO, AWHI	DAYS, null, 1 or DAYS, 176, null	AWU is populated but both AWLO and AWHI are not populated	Error	A one-sided window is incorrectly marked by the current Pinnacle 21 version as an error.
ADVS	VARIABLE, LABEL	ANL01FL, Analysis Flag 01	Variable label mismatch between dataset and ADaM standard	Error	Pinnacle 21 Community Edition validation uses ADaMIGv1.0. The value for the variable label conforms to ADaMIGv1.1 as required.