TDF_SDTM: SDTMIG v3.2 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 SDTM 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 SDTM 3.2 as the Configuration and 2016-06-24 as the CDISC Controlled Terminology.

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_SDTM datasets should be aware that depending on the situation, several options can and need to be considered.

Note that the rather large LB and QS domain datasets were split (see section 2.1) to accommodate limitations of the repository that was used during the work. This also demonstrates splitting of domains. This was not done because of any guidance or other requirement to split these domains.

In section 2.3, one file is listed that was not updated but copied from the original CDISCPILOT02 submission package. Users should be aware that the content of this .pdf file is likely not consistent with the data in the TDF_SDTM datasets because it was created with the original CDISCPILOT02 datasets. In addition, users need to understand that the location of the file 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 LB, SUPPLB, and QS domain datasets were split based on the values of the indicated variable as recommended in the SDTMIG. 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.

Domain	Split Domain	Value of split Variable	Comment about Split Variable
QS	qsco.xpt	QSCAT: ADAS-COG	Alzheimer's Disease Assessment Scale-Cognitive CDISC Version Questionnaire
QS	qsda.xpt	QSCAT: DAD	Disability Assessment for Dementia Questionnaire
QS	qsgi.xpt	QSCAT: ADCS-CGIC	Alzheimer's Disease Cooperative Study-Clinical Global Impression of Change Questionnaire
QS	qshi.xpt	QSCAT: MHIS-NACC	Modified Hachinski Ischemic Scale-NACC Version Questionnaire
QS	qsmm.xpt	QSCAT: MMSE	Mini-Mental Status Exam
QS	qsni.xpt	QSCAT: NPI	Neuropsychiatric Inventory Questionnaire
LB	lbch.xpt	LBCAT: CHEMISTRY	Chemistry
LB	lbhe.xpt	LBCAT: HEMATOLOGY	Hematology
LB	lbur.xpt	LBCAT: URINALYSIS	Urinalysis
SUPPLB	supplbch.xpt	IDVARVAL	Split based on IDVARVAL matching LBSEQ in lbch.xpt
SUPPLB	supplbhe.xpt	IDVARVAL	Split based on IDVARVAL matching LBSEQ in lbhe.xpt
SUPPLB	supplbur.xpt	IDVARVAL	Split based on IDVARVAL matching LBSEQ in lbur.xpt

2.2 AE Domain Modification

AE information was collected by visit, with a new record entered each time. In the CDISCPILOT02 dataset, these duplicate records caused 230 validation warnings “Duplicate records in AE domain” for the AE variables “AEDECOD, AETERM, AESEV, AESTDTC, USUBJID”. These records have been collapsed.

2.3 CDISCPILOT02 files, that are not updated

The project team decided that one file that is included in the CDISCPILOT02 should be included in the TDF_SDTM package but does not need to be updated. This file is added to satisfy a link in the define.xml but the content of the file might not be consistent with the updated datasets. The file named "annotated_crf.pdf" is placed in a subfolder named "cdisc_docs".

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 before the update of the define.xml file and is included so that users of the updated CDISC Pilot dataset can understand the extend 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 created by running the tool against the datasets.

Domain	Variables	Values	Message	Severity	Explanation
AE	VARIABLE	AEDY	Model permissible variable added into standard domain	Warning	AEDY was included because AEDTC is included.
AE	VARIABLE	EPOCH	Model permissible variable added into standard domain	Warning	Per FDA Study Data Technical Conformance Guide, the variable EPOCH should be included for clinical subject-level observations
AE	VARIABLE	AEDTC	Model permissible variable added into standard domain	Warning	AEDTC is included because AE information was collected by visit.
AE	AEENDTC	null	Missing End Time-Point value	Warning	AEENDTC is missing for some AEs and no value was collected to state that AEs were ongoing or not
CM	VARIABLE	EPOCH	Model permissible variable added into standard domain	Warning	Per FDA Study Data Technical Conformance Guide, the variable EPOCH should be included for clinical subject-level observations
CM	VARIABLE	CMDTC	Model permissible variable added into standard domain	Warning	CMDTC is included because CM information was collected by visit.
CM	VARIABLE	VISITDY	Model permissible variable added into standard domain	Warning	VISITDY is included because CM data was collected by visit.
CM	VARIABLE	VISIT	Model permissible variable added into standard domain	Warning	VISIT is included because CM data was collected by visit.
CM	VARIABLE	CMDY	Model permissible variable added into standard domain	Warning	CMDY was included because CMDTC is included.
CM	VARIABLE	VISITNUM	Model permissible variable added into standard domain	Warning	VISITNUM is included because CM data was collected by visit.

Domain	Variables	Values	Message	Severity	Explanation
DM	ARMCD, ACTARMCD, USUBJID	Xan_Lo, Xan_Lo, 01-703-1119	No baseline result in LB for subject	Warning	According to the study rules for setting the baseline flag, this subject did not have a specific baseline visit.
DM	ARMCD, ACTARMCD	Xan_Hi, Xan_Lo	ACTARMCD does not equal ARMCD	Warning	Some subjects were not treated according to protocol. This deviation was recorded as required and correctly marked as warning.
DS	VARIABLE	VISIT	Model permissible variable added into standard domain	Warning	VISIT was added to DS because protocol specified events were collected at particular visits.
DS	VARIABLE	VISITNUM	Model permissible variable added into standard domain	Warning	VISITNUM was added to DS because protocol specified events were collected at particular visits.
DS	VARIABLE	DSDY	Model permissible variable added into standard domain	Warning	DSDY was added because DSDTC is captured.
EX	VARIABLE	VISIT	Model permissible variable added into standard domain	Warning	VISIT was added to EX was collected by visits.
EX	VARIABLE	VISITNUM	Model permissible variable added into standard domain	Warning	VISITNUM was added to EX was collected by visits.
EX	VARIABLE	VISITDY	Model permissible variable added into standard domain	Warning	VISITDY was added to EX was collected by visits.
LB	VARIABLE	EPOCH	Model permissible variable added into standard domain	Warning	Per FDA Study Data Technical Conformance Guide, the variable EPOCH should be included for clinical subject-level observations
LB (LBHE)	LBORRES, LBTEST, LBORRESU, LBTESTCD		Missing value for LBORRESU, when LBORRES is provided	Warning	There are lab tests that do not have units. This is acceptable.
LB (LBHE)	LBTEST, LBSTRESN, LBTESTCD, LBSTRESC, LBSTRESU		Missing value for LBSTRESU, when LBSTRESC is provided	Warning	There are lab tests that do not have units. This is acceptable.

Domain	Variables	Values	Message	Severity	Explanation
LB (LBUR)	LBORRES, LBTEST, LBORRESU, LBTESTCD		Missing value for LBORRESU, when LBORRES is provided	Warning	There are lab tests that do not have units. This is acceptable.
MH	VARIABLE	VISITDY	Model permissible variable added into standard domain	Warning	VISITDY is included because MH data was collected by visit.
MH	VARIABLE	VISITNUM	Model permissible variable added into standard domain	Warning	VISITNUM is included because MH data was collected by visit.
MH	VARIABLE	VISIT	Model permissible variable added into standard domain	Warning	VISIT is included because MH data was collected by visit.
MH	VARIABLE	MHSEV	Model permissible variable added into standard domain	Warning	MH data to be analyzed in consideration with AE data; similar data allows cross domain comparisons. AESEV is in AE
MH	VARIABLE	MHHLGT	Model permissible variable added into standard domain	Warning	MH data to be analyzed in consideration with AE data; similar data allows cross domain comparisons. AEHLGT is in AE
MH	VARIABLE	MHLLT	Model permissible variable added into standard domain	Warning	MH data to be analyzed in consideration with AE data; similar data allows cross domain comparisons. AEHLLT is in AE
MH	VARIABLE	MHHLT	Model permissible variable added into standard domain	Warning	MH data to be analyzed in consideration with AE data; similar data allows cross domain comparisons. AEHHLT is in AE
QS	VARIABLE	EPOCH	Model permissible variable added into standard domain	Warning	Per FDA Study Data Technical Conformance Guide, the variable EPOCH should be included for clinical subject-level observations
QS (QSCO)	QSSTRESC, QSBLFL	null, Y	Missing QSSTRESC value for Baseline record	Warning	A baseline value was not recorded as it might happen in trials.
SE	VARIABLE	SESTDY	Model permissible variable added into standard domain	Warning	Per Pinnacle 21 validation SESTDY variable is required when SESTDTC variable is present.
SE	VARIABLE	SEENDY	Model permissible variable added into standard domain	Warning	Per Pinnacle 21 validation SEENDY variable is required when SEENDTC variable is present.

Comments about CDISC PILOT Updated for SDTMIG v3.2

Domain	Variables	Values	Message	Severity	Explanation
SV	VARIABLE	EPOCH	Model permissible variable added into standard domain	Warning	Per FDA Study Data Technical Conformance Guide, the variable EPOCH should be included for clinical subject-level observations
TS	TSPARM	Age Group	TSPARM value not found in 'Trial Summary Parameter Test Name' extensible codelist	Warning	A value was added to the extensible code list.
TS	TSPARMCD	AGESPAN	TSPARMCD value not found in 'Trial Summary Parameter Test Code' extensible codelist	Warning	A value was added to the extensible code list.
VS	VARIABLE	EPOCH	Model permissible variable added into standard domain	Warning	Per FDA Study Data Technical Conformance Guide, the variable EPOCH should be included for clinical subject-level observations