PharmaSUG 2018 - Paper SS21

eSubmission - Are you really Compliant?

Majdoub Haloui, Merck & Co., Inc., Upper Gwynedd, PA, USA

Suhas R. Sanjee, Merck & Co., Inc., Upper Gwynedd, PA, USA

ABSTRACT

Pinnacle 21 Enterprise is a valuable tool for SDTM, ADaM and define.xml validation. However, the tool is still evolving and has some limitations, so additional quality checks need to be created to supplement the tool's assessments to ensure accuracy, validity and completeness of SDTM/ADaM datasets as well as the define.xml. Adding manual quality checks can be cumbersome and time consuming.

In this paper, we will discuss the extra quality checks that are implemented beyond what is available in the Pinnacle 21 Enterprise tool. The paper focuses on how these checks are created and an efficient approach for converting define.xml into SAS datasets so they can be validated against the actual SDTM/ADaM datasets.

INTRODUCTION

Pinnacle 21 Enterprise is the leading industry web-based application used by sponsors and CROs to validate SDTM / ADaM datasets and define.xml against CDISC standards. The FDA and PMDA are also using Pinnacle 21 Enterprise to review submission data from sponsors. Pinnacle 21 Enterprise has many useful features, including validating SDTM, ADaM and define.xml as well as generating define.xml version 2.0.

This paper introduces new SAS-based quality checks that supplement what is already available in the Pinnacle 21 Enterprise tool.

QUALITY CHECK DESCRIPTIONS

Having SDTM/ADaM dataset attributes consistent with the define.xml can contribute to an efficient and seamless regulatory review process, potentially expediting drug approval.

The following table provides a list of checks developed by Merck & Co., Inc. to address checks that are not performed by Pinnacle 21 Enterprise but are needed to meet the requirements defined by the current regulatory agency standards and expectations. Consistency checks between the define.xml documents and the submission data are among the supplemental checks needed to produce quality deliverables.

SAS Macro	Checks Implemented
check0csr0xpt	All submission XPT datasets must match the final sas7bdat datasets that were used to develop CSR TFLs.

check0define/
check0def0data0consistency
	• The define.xml must contain controlled terminology for all variables that use controlled terminology, and must be consistent with the SDTM-IG/ADaM IG version being used for the study.
	• Check consistency between data and define.xml (define).
	  ○ All datasets in submission must be defined in define.xml and vice versa
	  ○ Dataset label must match between define and data
	  ○ All variables in XPT files must be defined in define.xml and vice versa
	  ○ Variable labels must match between define and data
	  ○ Variable order must match between define and data
	  ○ Variable type must match between define and data
	  ○ All controlled terminology values in data must be present in the define

SAS Macro	Checks Implemented
	• The ISO 8601 format must not be provided as a Codelist or an External codelist • Datasets must be listed in alphabetical order by name attribute within each class in the define.xml file. **SDTM** o Trial Design Datasets o Special-purpose Domains o Interventions Domains o Events Domains o Findings Domains o Findings About o Relationship Datasets **ADaM** o Subject Level Analysis Dataset o Basic Data Structure o ADaM Other
check0file0names	File names must only contain letters, numbers or hyphens and must be in lower case. No underscores or blanks are allowed
check0keys0dups	All Keys listed in the define ,when used, must represent uniqueness in the data
check0meta check0meta0adam0sdtm	• Dataset labels and variable names /labels must not contain: punctuation, dashes, spaces, or other non-alphanumeric symbols (like line breaks and carriage returns) • Across all ADaM datasets, core variables (i.e., AGEGRP, TRT01P, population flags, etc.) must have consistent metadata (i.e., label, type, etc) • Dataset labels must be unique across SDTM & ADaM datasets. (ADaM dataset labels are required to be different from SDTM labels) • Variables inherited from SDTM must retain the same attributes except length • For ADaM datasets, only SAS built-in Date or Datetime formats may be used. Check if the format is consistent across the same submission: date7, datetime, etc
check0orphans	Check for RELREC or SUPPQUAL domains that have references to non-existing records
check0ts	Study Start Date in the TS domain, the TSVAL must be populated where TSPARMCD = 'SSTDTC'
check0visit	• The Trial Visits (TV) and Subject Visits (SV) domains should be consistent with other domains • VISIT, VISITNUM (if it exists) from all of the domains should be consistent with TV/SV • Usage of VISITNUM values must be consistent across domains. All values of Visit Number (VISITNUM) must be the same for a given value of Visit Name (VISIT)
check0xpt	• Variables must not contain customized format in XPT for SDTM • STUDYID in all XPT files must be consistent and correct for the given study
file0exist	• DM, define.xml, style sheet ,Trial Summary (TS) dataset and csdrg must be present in the tabulations\sdtm folder • ADSL, define.xml, style sheet and adrg must be present in the analysis\adam folder • Correct spelling of the file name (define.xml) is included in the submission
check0arm	• All analysis variables and parameters referenced in the ARM must exist in the dataset • All programs, macros, datasets and variables referenced in the define/ARM must exist in the submission folder

MACRO DESIGN AND ARCHITECTURE

The independent, interchangeable SAS macros described in the previous section perform the supplemental compliance/quality checks. This modular approach also provides users with flexibility to execute only applicable checks. This also provides improved quality, efficiency and ease of maintenance. Each macro/module outputs its own dataset containing the issues that it identifies. The suite of modules also creates or updates a "Summary" dataset to provide an overview of the issues. Figure 1 shows the block diagram of the macro design and architecture.

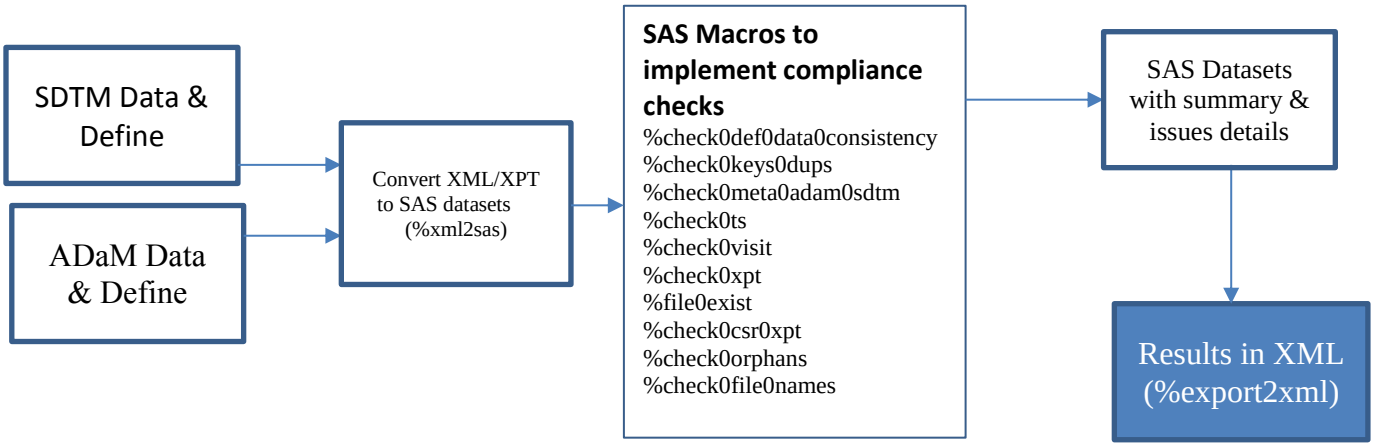


Figure 1. Macro design and architecture

Conversion from XML to SAS dataset(s) %xml2sas

This macro converts the ADaM & SDTM define.xml to SAS datasets. It takes in the XML file, XML map as input and generates a SAS dataset that is used for compliance checks. Here is the sample code that extracts required information from the XML file (based on the XML map) and generates SAS datasets:

```
filename   define "c:\protxxx\metadata\define.xml";
filename   SXLEMAP "c:\protxxx\metadata\define.map";
*Assign library reference to XML file using XMLV2 engine;
libname    define xmlv2 xmlmap=SXLEMAP access=READONLY;

*Create output datasets;
proc copy in=define out=outlib;
run;
```

The map files were generated using SAS XML mapper. Figure 2 shows an XML map with "Type" and SDTM Core information.

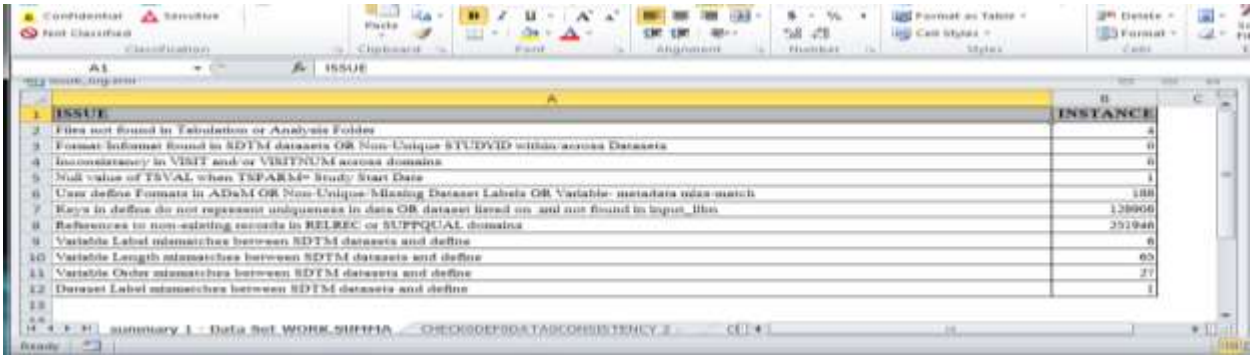
```
<?xml version="1.0" encoding="UTF-8"?>
<SXLEMAP name="AUTO_GEN" version="2.1">
    <NAMESPACES count="3">
        <NS id="1" prefix="xlink">http://www.w3.org/1999/xlink</NS>
        <NS id="2" prefix="">http://www.cdisc.org/ns/odm/v1.3</NS>
        <NS id="3" prefix="def">http://www.cdisc.org/ns/def/v2.0</NS>
    </NAMESPACES>
    <TABLE description="ItemRef" name="ItemRef">
        <TABLE-PATH syntax="XPathENR">/{2}ODM/{2}Study/{2}MetaDataVersion/{2}ItemGroupDef/{2}ItemRef</TABLE-PATH>

        <COLUMN name="OID">
            <PATH syntax="XPathENR">/{2}ODM/{2}Study/{2}MetaDataVersion/{2}ItemGroupDef/{2}ItemRef/@ItemOID</PATH>
            <TYPE>character</TYPE>
            <DATATYPE>string</DATATYPE>
            <LENGTH>35</LENGTH>
        </COLUMN>
        <COLUMN name="ItemRef_OrderNumber">
            <PATH syntax="XPathENR">/{2}ODM/{2}Study/{2}MetaDataVersion/{2}ItemGroupDef/{2}ItemRef/@OrderNumber</PATH>
            <TYPE>numeric</TYPE>
            <DATATYPE>integer</DATATYPE>
        </COLUMN>
        <COLUMN name="ItemRef_Mandatory">
            <PATH syntax="XPathENR">/{2}ODM/{2}Study/{2}MetaDataVersion/{2}ItemGroupDef/{2}ItemRef/@Mandatory</PATH>
            <TYPE>character</TYPE>
            <DATATYPE>string</DATATYPE>
            <LENGTH>3</LENGTH>
        </COLUMN>
    </TABLE>
</SXLEMAP>
```

Figure 2. Snapshot of XML map

Output XML File

The output XML file "Summary" tab provides issue descriptions and frequencies (Figure 3). An individual tab for each compliance check module provides a detailed list of issues that it detects (Figures 4a, 4b & 4c).



	A	B
1	ISSUE	INSTANCE
2	Files not found in Tabulation or Analysis Folder	4
3	Format Informat found in SDTM datasets OR Non-Unique STUDYID within across Datasets	0
4	Inconsistency in VISIT and or VISITNUM across domains	0
5	Null value of TSVAL when TSPARM= Study Start Date	1
6	User define Formats in ADaM OR Non-Unique Missing Dataset Labels OR Variable- metadata miss-match	188
7	Keys in define do not represent uniqueness in data OR dataset listed on and not found in input_libn	128906
8	References to non-existing records in RELREC or SUPPQUAL domains	251946
9	Variable Label mismatches between SDTM datasets and define	0
10	Variable Length mismatches between SDTM datasets and define	63
11	Variable Order mismatches between SDTM datasets and define	27
12	Dataset Label mismatches between SDTM datasets and define	1
13		

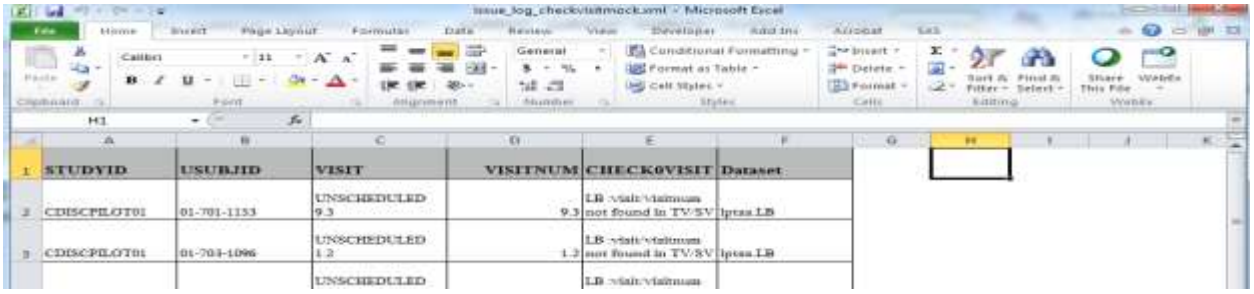
Figure 3. Snapshot of "Summary" tab in Output XML file

Dataset	CHECKSMETADATADAMSSDTM	ADaM Dataset	Variable	LENGT	ALENGT	ADSL LABEL	ADaM LABEL
ADSL	Miss match to variable length observed	ADCM	TRT01A	40	100		
ADSL	Miss match to variable length observed	ADCM	TRT01P	40	100		
ADSL	Miss match to variable length observed	ADSFF2	A1CORI	20	30		
ADSL	Miss match to variable length observed	ADSFF2	ACCPFE04	20	30		
ADSL	Miss match to variable length observed	ADSFF2	ACCPFE06	99	80		
ADSL	Miss match to variable label observed	ADSFF2	A1CORI			Baseline A1C group (<8 , >= 8)	Baseline A1C group (<8.0 , >= 8.0)
ADSL	Miss match to variable label observed	ADSFF2	A1CORLN			Baseline A1C (N) group (<8 , >= 8)	Baseline A1C (N) group (<8.0 , >= 8.0)
ADSL	Miss match to variable label observed	ADSFF2	ACCPFE06			Subj With Drug Compliance <75%	Subj With Drug Compliance <75%
ADSL	Miss match to variable label observed	ADSFF2	ACCPFE06			Subj With Incorrect Study Medication	Subj with Incorrect Study Medication

Figure 4a. Snapshot showing details about the issues in ADaM attributes

Dataset	Issue
TS	TS: TSPARM = SSTDTC (Study Start Date) does not exist

Figure 4b. Snapshot showing details about the issue with TS



STUDYID	USUBJID	VISIT	VISITNUM	CHECK0VISIT	Dataset
CDISCPILOT01	01-701-1153	UNSCHEDULED 9.3	9.3	LB visit/visitnum not found in TV/SV	lptss LB
CDISCPILOT01	01-703-1086	UNSCHEDULED 1.2	1.2	LB visit/visitnum not found in TV/SV	lptss LB
		UNSCHEDULED		LB visit/visitnum	

Figure 4c. Snapshot showing details about the issue where VISIT/VISITNUM variables are not consistent with SV/TV

CONCLUSION

High quality study data and accompanying metadata are extremely important for the regulatory review process. They add significant value to the submission by allowing reviewers to efficiently interpret and understand submitted data. The result is a time savings for the regulatory reviewer, which can potentially expedite the drug approval. The modular SAS-based quality checks presented in this paper can help expedite the quality review time.

REFERENCES

1. FDA STUDY DATA TECHNICAL CONFORMANCE GUIDE October 2017 https://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM384744.pdf
2. Sergiy Sirichenko, Max Kanevsky What is high quality study metadata? PharmaSUG 2016
3. Lex Jansen Understanding the define.xml and converting it to a relational database PharmaSUG 2009

CONTACT INFORMATION

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

Majdoub Haloui
Merck & Co., Inc.
351 North Sumneytown Pike
North Wales PA 19454-2505
Phone: 267-305-2629
E-mail: majdoub.haloui@merck.com

Suhas R. Sanjee
Merck & Co., Inc.
351 North Sumneytown Pike
North Wales PA 19454-2505
Phone: 267-305-0878
E-mail: suhas_sanjee@merck.com