WITH STANDARDS – UNLOCK THE POWER OF DATA

cdisc

2022
US
INTERCHANGE
26-27 OCTOBER | AUSTIN

SDTM Metadata Cross Checks to Improve Quality

Presented by Nicole Jones, Senior Statistical Programmer
BARDS, Merck & Co., Inc., Rahway, NJ, USA

Meet the Speaker

Nicole Jones

Title: Senior Statistical Programmer

Organization: Merck & Co., Inc., Rahway, NJ, USA

I am a Senior Scientist Statistical Programmer at Merck. I have a Masters degree in Public Health with a concentration in Epidemiology and a Post-Baccalaureate Certificate in Applied Biostatistics from Rutgers University School of Public Health. I currently support R projects including package qualification, package development and R Shiny development. I have been at Merck for nearly two years and have supported SDTM programming outside of my current role.

Disclaimer and Disclosures

- *The views and opinions expressed in this presentation are those of the author(s) and do not necessarily reflect the official policy or position of CDISC.*
- *The author(s) have no real or apparent conflicts of interest to report.*

Agenda

1. Background & Problem Statements
2. Metadata Quality Checks
3. Design Considerations & Proposed Solution
4. Development of Application
5. Output
6. Discussion

Background and Problem Statements

Background

- SDTM And ADaM Metadata are vital components of eSubmissions
- Metadata components are generated as part of different workflows
- SDTM metadata is the focus of this presentation
  - SDTM define.xml
  - Annotated Case Report Form (aCRF)
  - Value Level Metadata (VLM) in SDTM datasets

Problem Statements

- Despite automation efforts, discrepancies between the metadata and SDTM data can still exist
- No commercial tool exists in the current state to perform the unique list of checks performed by our proposed tool
- No readily available tool /software could reliably parse the aCRF in PDF format

Metadata Quality Checks

Metadata Quality Checks

Seven checks to identify discrepancies between Define.xml, aCRFs and SDTM datasets were defined

Number/Group	Message	Description
1. VLM Presence **Define.xml and SDTM Data**	QVAL found in Define but no matching QNAM in dataset	When a QVAL VLM record exist in the define, there should be a record in the data with the referenced QNAM
2. VLM Presence **Define.xml and SDTM Data**	QNAM found in dataset but no matching QVAL in define	When the data have a value for QNAM, there should be a corresponding QVAL VLM record in the define
3. VLM Presence **Define.xml, SDTM Data and aCRF**	QNAM annotated on CRF and missing from either define or dataset. Note: If missing from both, this is okay because it is a common scenario that variables are annotated but later dropped because no data were collected.	When there is a QVAL VLM annotation on the aCRF, there should be a record in the data with the referenced QNAM and there should be a QVAL VLM record in the define with corresponding QNAM

Metadata Quality Checks

Number/Group	Message	Description
4. VLM Origin SDTM Data and aCRF	QORIG = CRF but its not annotated	When the data have QORIG = CRF, there should be a corresponding QVAL VLM annotation on the aCRF
5. VLM Origin Define.xml and SDTM Data	QORIG does not match the Origin in the define	The QVAL VLM record in the define should have a define origin that matches the QORIG value in the data
6. VLM Label Define.xml and SDTM Data	QLABEL does not match Description in the Define	The QVAL VLM record in the define should have a description that matches the value of QLABEL in the data **Value Level Metadata - SUPPCM [QVAL]** Variable: QVAL Where: QNAM EQ ATC4CODE (WHODRUG ATC Level 4 Code) Type: text Length / Display Format: 5
7. Variable Origin Define.xml and aCRF	Annotated on aCRF but define origin != CRF	When a variable is annotated on the aCRF and referenced in the define, the define origin should be CRF

Design Considerations & Proposed Solution

Design Considerations

- Inputs:
  - aCRF
  - SDTM datasets
  - define.xml
- Our solution had to:
  - Scalable
  - Efficient
  - Easy to use

Proposed Solution

- R and Shiny were chosen for implementation
  - R is an open-sourced programming language for statistical computing and graphics with a vast community of users that contribute add-on packages for various tasks.
  - Shiny is an R package that makes it easy to build interactive web apps straight from R
- Utilizing the {xml2} package, the exported comments (XFDF file) could be parsed without data loss
- Shiny allows the creation of an intuitive UI for programmers
  - Doesn't require users to know R – Ease of use
  - Utilizing RStudio Connect, study programmers can utilize the tool without needing to install additional software – Scalable
  - Minimizes manual effort of study programmers - Efficient

Proposed Solution

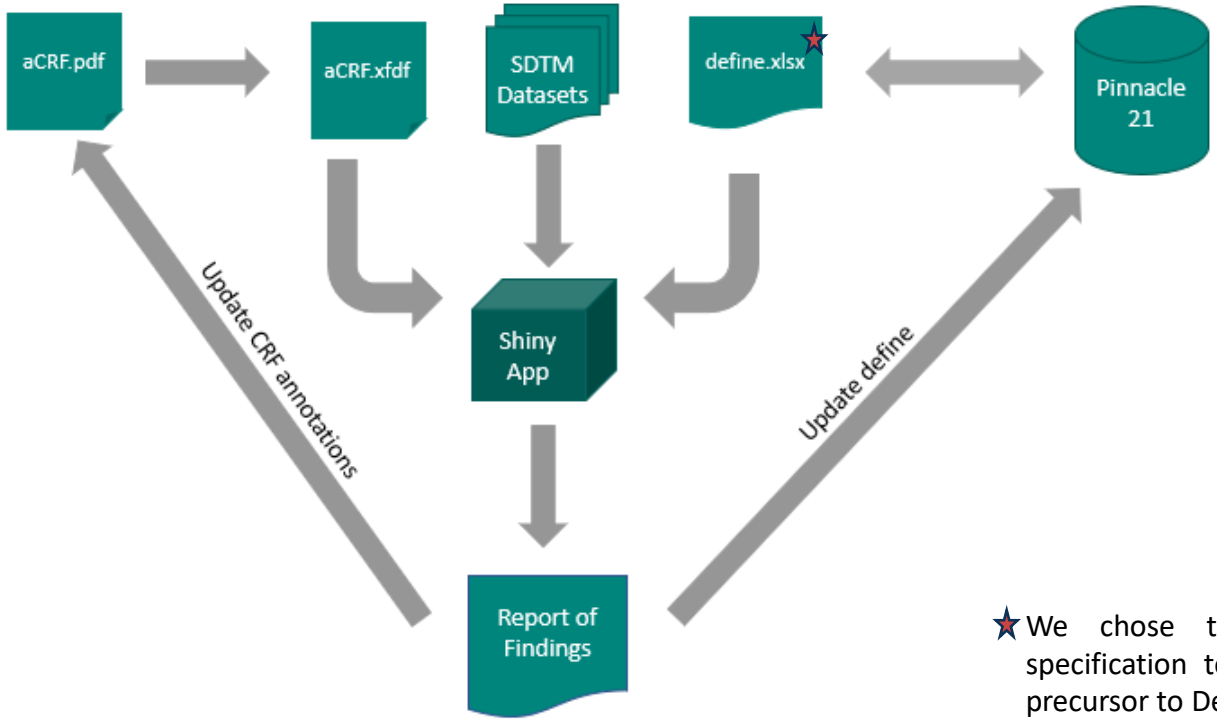
Extracting aCRF annotations

- Comments (annotations) in PDF documents can be exported to an XFDF file
- An XFDF file is an XML Forms Data Format that stores information usable by a PDF file
- To create an XFDF file, the following is done:

Export comments to a data file

1. From the options menu ••• in the comments list, choose **Export All To Data File**.
2. Name the file and choose **Acrobat FDF Files (*.fdf)** or **Acrobat XFDF Files (*.xfdf)** for the file type.
3. Specify a location for the file, and then click **Save**.

Proposed Solution – Workflow Diagram



★ We chose to use the Excel specification template which is a precursor to Define XML in place of the define.xml file because this is the file that users interact with most frequently

Development of Application

Programming

- Define 2 functions:
  - getSUPP: reads and combines all SDTM XPT files into a single dataframe
  - getComments: reads the acrf.xfdf file and parse the annotations
- Combine output from the 2 functions
- Perform 7 checks on combined dataframe
- Produce 2 final dataframes:
  - Inconsistencies found in the SDTM define or data
  - Inconsistencies found in the aCRF
- Display results in UI
- Download findings to Excel, optional

UI

Define.xml SUPPQUAL QC | QNAM Discrepancy Checker | Instructions | Configuration

Please Upload Pinnacle 21 Define Excel Spec

Browse... MK9999-001.xlsx

Upload complete

Select .xfdf file to upload

Browse... MK9999-001.xfdf

Upload complete

CPI Path to SDTM data

<PATH TO DATA>

Generate

Download Report of Origin Issues

Show 10 entries

Search:

Annotation Issues (Generated using: MK9999-001.xlsx and MK9999-001.xfdf)

	Dataset	QNAM	Comment	Origin	Issue	page	Issue_Count
1	SUPPAE	RELPR	QVAL when QNAM = 'RELPR' in SUPPAE	CRF	Annotated on CRF but not found in Dataset	3	1

Showing 1 to 1 of 1 entries

Previous 1 Next

Show 100 entries

Search:

Define Issues (Generated using: MK9999-001.xlsx and MK9999-001.xfdf)

	Dataset	QNAM	QLABEL(Dataset)	QLABEL(Define)	QORIG(Dataset)	QORIG(Define)	Issue	Issue_Count
1	SUPPAE	AEACNBLD	Action Taken With Blinded Study Med	Action Taken With Blinded Study Med	CRF		QORIG does not match origin in Define	1
2	SUPPAE	AECLINT	Clinical Interest	Clinical Interest	CRF		QORIG does not match origin in Define	1
3	SUPPAE	AEDURDD	Duration of Adverse Event Diff of Dates	Duration of Adverse Event Diff of Dates	Derived		QORIG does not match origin in Define	1

Output

The Output

- Option to export findings to Excel Workbook

- Findings categories:
  - Issues with the Define
  - Issues with the aCRF

Sample Output: Define Issues

A	B	C	D	E	F	G	H
Dataset	QNAM	QLABEL(Dataset)	QLABEL(Define)	QORIG(Dataset)	QORIG(Define)	Issue	Issue_Count
SUPPAE	AEACNBLD	Action Taken With Blinded Study Med	Action Taken With Blinded Study Med	CRF		QORIG does not match origin in Define	1
SUPPAE	AECLINT	Clinical Interest	Clinical Interest	CRF		QORIG does not match origin in Define	1
SUPPAE	AEDURDD	Duration of Adverse Event Diff of Dates	Duration of Adverse Event Diff of Dates	Derived		QORIG does not match origin in Define	1
SUPPAE	AEDURDDU	Duration Units	Duration Units	Derived		QORIG does not match origin in Define	1
SUPPAE	ELEMENT	Description of Element	Description of Element	Derived		QORIG does not match origin in Define	1
SUPPAE	ETCD	Element Code	Element Code	Derived		QORIG does not match origin in Define	1
SUPPAE	RELBLD	Relationship to Blinded Study Med	Relationship to Blinded Study Med	CRF		QORIG does not match origin in Define	1
SUPPAE	SMB	Study Medication - Blinded	Study Medication - Blinded	Assigned		QORIG does not match origin in Define	1
SUPPAE	SPDYRLEP	Stop Day Rel to Epoch	Stop Day Rel to Epoch	Derived		QORIG does not match origin in Define	1
SUPPAE	STDYRLEP	Start Day Rel to Epoch	Start Day Rel to Epoch	Derived		QORIG does not match origin in Define	1
SUPPAE	AEEPRELI	Epi/Pandemic Related Indicator	Epi/Pandemic Related Indicator	Assigned		QORIG does not match origin in Define	1
SUPPCM	CMDRUG	Encoded Drug Name	Encoded Drug Name	Assigned		QORIG does not match origin in Define	1
SUPPCM	ELEMENT	Description of Element	Description of Element	Derived		QORIG does not match origin in Define	1
SUPPCM	ETCD	Element Code	Element Code	Derived		QORIG does not match origin in Define	1
SUPPCM	SPDYRLEP	Stop Day Rel to Epoch	Stop Day Rel to Epoch	Derived		QORIG does not match origin in Define	1
SUPPCM	STDYRLEP	Start Day Rel to Epoch	Start Day Rel to Epoch	Derived		QORIG does not match origin in Define	1
SUPPCM	WCLAS01	Who Medication Class_01	Who Medication Class_01	Assigned		QORIG does not match origin in Define	1
SUPPCM	WCLAS02	Who Medication Class_02	Who Medication Class_02	Assigned		QORIG does not match origin in Define	1
SUPPCM	WCLSCD01	Who Medication Class Code_01	Who Medication Class Code_01	Assigned		QORIG does not match origin in Define	1
SUPPCM	WCLSCD02	Who Medication Class Code_02	Who Medication Class Code_02	Assigned		QORIG does not match origin in Define	1
SUPPCM	WCLAS03	Who Medication Class_03	Who Medication Class_03	Assigned		QORIG does not match origin in Define	1
SUPPCM	WCLAS04	Who Medication Class_04	Who Medication Class_04	Assigned		QORIG does not match origin in Define	1
SUPPCM	WCLSCD03	Who Medication Class Code_03	Who Medication Class Code_03	Assigned		QORIG does not match origin in Define	1
SUPPCM	WCLSCD04	Who Medication Class Code_04	Who Medication Class Code_04	Assigned		QORIG does not match origin in Define	1
SUPPDM	BRTHDTI	Imputed Date of Birth	Imputed Date of Birth	Derived		QORIG does not match origin in Define	1
SUPPDM	MSUBJNUM	Merck Subject Number	Merck Subject Number	CRF		QORIG does not match origin in Define	1

Define Issues | aCRF Issues

Sample Output: aCRF Issues

A	B	C	D	E	F	G
Dataset	QNAM	Comment	Origin	Issue	page	Issue_Count
SUPPAE	RELPR	QVAL when QNAM = 'RELPR' in SUPPAE	CRF	Annotated on CRF but not found in Dataset	3	1

Define Issues | aCRF Issues

Discussion

R Shiny Best Practices & Resource Commitment

- Using file uploads:
  - Avoid uploading files when the output from the app is being sent to regulatory agencies
  - Avoid using file uploads for large amounts of data
- Time commitment:
  - Developing and fully validating a Shiny app can be resource intensive
  - Other tools should be considered for more immediate solutions

Benefits of Shiny

- Allows end users to benefit from the functionality of R without installing new software
- Allows reviewers or programmers with minimal programming experience to benefit from the R package
- Provides an elegant and intuitive way to interact with the programs

Implementation of Consistency of aCRF

- Consistent annotation style for the aCRF is essential to performing these metadata checks and hence important to establish standard conventions
- Internally, the following VLM annotation style has been implemented:
  - QVAL when QNAM = “VALUE” in SUPPxx (i.e. QVAL when QNAM = 'AECLINT' in SUPPAE)

Conclusion

- The proposed tool is
  - Scalable as it doesn't need users to install additional software locally
  - Easy to use
  - Eliminates manual checks and helps the study teams to identify & address discrepancies efficiently
- The quest to create harmonized and accurate metadata is not over.
- We hope the tool and checks presented in this presentation inspire other organizations to incorporate in their workflow.
- We hope some of the checks can be incorporated into existing tools used by the industry so we can continue to streamline our processes.

Questions