Majdoub Haloui is a seasoned Statistical Programmer at Merck & Co., Inc. with over twenty-four years of experience across the pharmaceutical, biotech, information technology, and engineering sectors. He excels in submission data standards and quality management, offering exceptional training, consultation, and mentorship to study teams across a wide range of therapeutic areas, including both drugs and vaccines. Majdoub' s expertise encompasses all phases of clinical development, from early to late stages, and he is skilled in fostering effective partnerships and collaborations.

Hemu Shere is a Principal Scientist at Merck & Co., Inc. with the Platforms, Process, and Inspections Group within Statistical Programming (SP). She is currently platform and tools lead with oversight for systems used by programming and statistics for the analysis and reporting of clinical studies. She has been with Merck for 17 years supporting SP and Information Technology and has also worked at Parexel as a statistical programmer.

Dileep Penmetsa is an Associate Principal Scientist at Merck & Co., Inc.in the Submissions Data Standards team within Statistical Programming. He has over 14 years of experience in Statistical Programming in the industry and has demonstrated expertise in CDISC data standards and submission data standards along with experience in successfully guiding study teams through the submission process, ensuring compliance with all relevant regulations and guidelines.

Surabhi Dutta is an Associate Principal Scientist at Merck & Co., Inc. working within the Platforms, Process, and Inspections Group in Statistical Programming. She holds the role of Business System Owner within BARDS and possesses a versatile background in Statistical Programming, Clinical Data Reporting, and Data Science. Currently pursuing a Ph.D. in Health Informatics, her research focuses on enhancing the usability and interoperability of real-world data through innovative applications of Artificial Intelligence and Machine Learning, with the goal of enabling evidence-based decision making.

MERCK
INVENTING FOR LIFE

Automating Consistency Checks Across Regulatory CDISC Submission Documents

Majdoub Haloui, Merck & Co., Inc., Rahway, NJ, USA
Hemu Shere, Merck & Co., Inc., Rahway, NJ, USA
Dileep Penmetsa, Merck & Co., Inc., Rahway, NJ, USA
Surabhi Dutta, Merck & Co., Inc., Rahway, NJ, USA

PHUSE SDE Collegeville PA

August 29th, 2024

Objective

To provide an in-depth overview of an automated tool designed to verify the consistency and accuracy of study data documentation prior to submission. The presentation will cover the tool's capability to identify issues such as missing links or discrepancies across SDTM and ADaM packages, its target users (SDTM and Analysis & Reporting programmers), and its applicability for both in-house and outsourced studies, with a discussion on potential limitations for certain checks. Additionally, the presentation will outline the technical architecture underpinning the system, offering insights into its underlying structure and functionality.

Agenda

- Introduction
- Consistency Checks Benefits
- Types of Checks
- High Level Architecture
- Technology
- Benefits of Using the Tool
- Conclusion

Introduction

- Quality deliverables are important for successful submissions to regulatory agencies.
- No commercially available tools to support **cross-checking** of information across **DRGs, define.xml and aCRF.**
  - Pinnacle 21 Enterprise (P21E) is a valuable tool for SDTM, ADaM and define.xml validation as well as define.xml creation but not for document content cross checking.
- A comprehensive review of submission documents is a massive undertaking especially when it is done manually.
  - Tedious nature of the work and field-by-field comparison may result in gaps and mismatches that can easily be overlooked.
  - Certain tasks are difficult and time consuming to do manually.

Consistency Checks Benefits

- Automation eliminates incomplete and inconsistent findings, reduces the burden on programming resources, and greatly improves the quality of the submission deliverables.

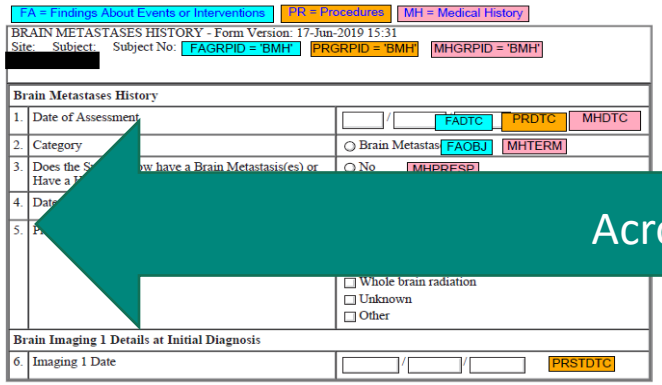
- Initiated a project to deliver a system that would aid programmers in providing quality submission documents that:
  - Checks for **consistency** and correctness across **data reviewer's guides**, **define.xml(2.0/2.1) files and aCRF**.
  - Identifies issues such as documentation errors and inconsistencies across documents, including missing hyperlinks or mismatches across SDTM and ADaM packages.
  - Streamlines the submission process with less manual cross-checking, increasing quality and reducing the time and effort within the submission process.

Types of Checks

- System implements four types of checks:
  - **SDTM**: Ensures consistency across SDTM submission documents.
  - **ADaM**: Ensures consistency across ADaM submission documents.
  - **SDTM & ADaM**: Ensures consistency across SDTM & ADaM submission documents.
  - **All**: Executes all check types above.
- Who Benefits:

Role	Description
SDTM Programmer	Perform consistency checks across SDTM documents (cSDRG, aCRF, define.xml)
Analysis & Reporting Programmer	Perform consistency checks across ADaM documents (ADRG, define.xml)
Study Lead	Perform consistency checks for both SDTM and ADaM documents

SDTM Check Type Details


FA = Findings About Events or Interventions | PR = Procedures | MH = Medical History

BRAIN METASTASES HISTORY - Form Version: 17-Jun-2019 15:31

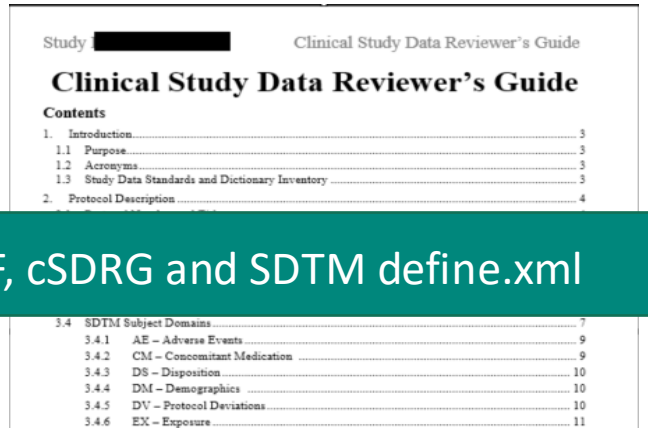
Site: Subject: Subject No: FAGRPID = 'BMH' PRGRPID = 'BMH' MHGRPID = 'BMH'

Brain Metastases History

1.	Date of Assessment	/ / FADTC PRDTC MHDTC
2.	Category	○ Brain Metastas FAOBJ MHTERM
3.	Does the S… w have a Brain Metastasis(es) or Have a …	○ No MHPRESP
4.	Date	
5.	P…	☐ Whole brain radiation ☐ Unknown ☐ Other

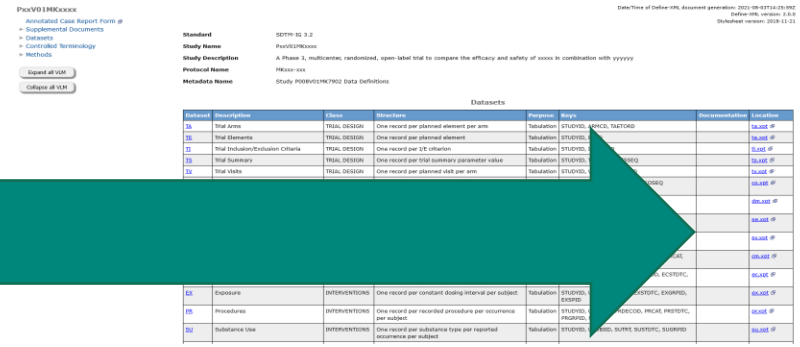
Brain Imaging 1 Details at Initial Diagnosis

6.	Imaging 1 Date	/ / PRSTDTC


Study [redacted] Clinical Study Data Reviewer's Guide

Clinical Study Data Reviewer's Guide

Contents


Datasets

Across aCRF, cSDRG and SDTM define.xml

- All SDTM subject domains referenced in cSDRG exist in the SDTM define.xml and their labels are consistent
- When a variable origin in the SDTM define.xml is set to CRF, that variable must be annotated on the right CRF page
- All domains listed under the question 'Were any domains planned, but not submitted because no data were collected?' in the cSDRG (section 3.1 Overview) are annotated on the CRF but do not exist in SDTM define.xml
- At least one of these columns should be checked for a domain/dataset " Efficacy, Safety or Other/ Baseline or other subject characteristics" in the 'Subject Domains' table in the cSDRG
- Empty cells checks across the cSDRG tables, SDTM CT (in define 2.1/cSDRG), SDTM IG Cross checks (define.xml/cSDRG)
- **Detailed check list is available in the backup slides**

Inconsistency Checks- Example

SDTM define.xml

CMDOSE	Dose per Administration	float	Record Qualifier	8		CRF Annotated Case Report Form [24]
CMDOSU	Dose Units	text	Variable Qualifier	11	Dose Units in Concomitant Medications [26 Terms]	CRF Annotated Case Report Form [2478]
CMDOSFRM	Dose Form	text	Variable Qualifier	19	Dose Form in Concomitant Medications [20 Terms]	CRF Annotated Case Report Form [28]

aCRF

Discrepancy

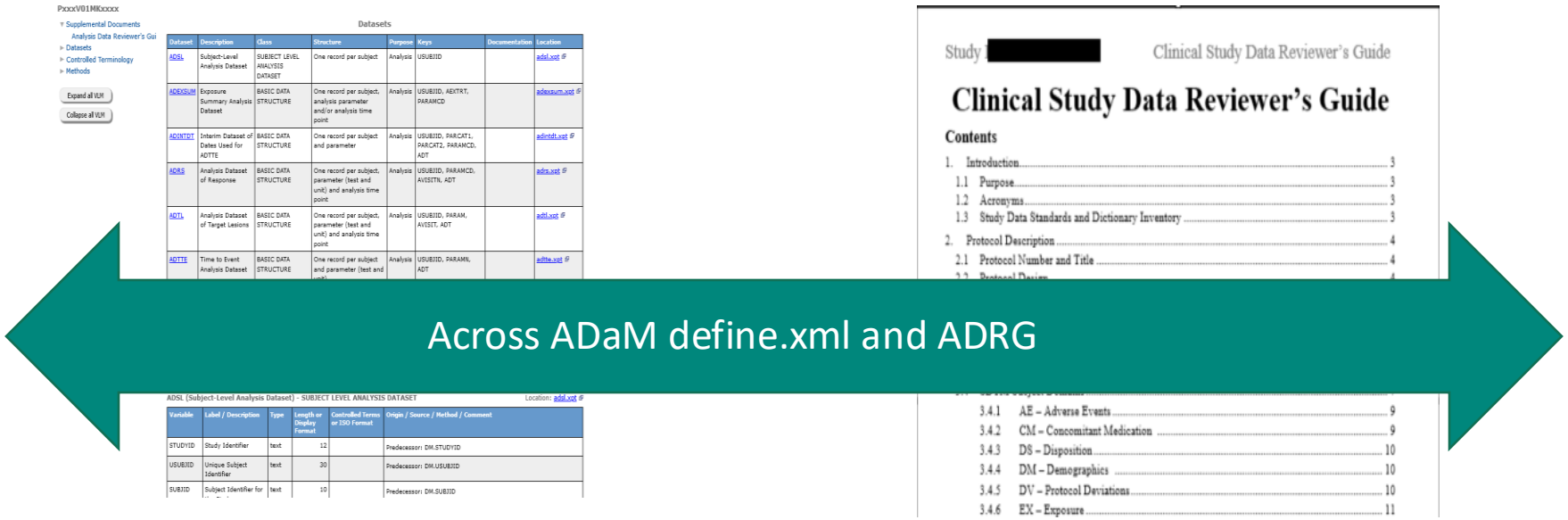
CM = Concomitant Medications | FA = Findings About Events or Interventions | RS = Disease Response

Header Text: [redacted]
Visit:
Form Version: [redacted]
Site No:
Subject No:
Generated By: [redacted]

Form: [redacted]
Form Status:
Site Name:
Subject Initials:
Generated Time (GMT): 31/Mar/2022 14:19

CMGRPID = 'ODB' | FAGRPID = 'ODB' | RSGRPID = 'ODB'

ADaM Check Type Details



- All ADaM datasets referenced in ADRG section 5.2 Analysis Datasets exist in the ADaM define.xml and their labels and structure are consistent
- All analysis datasets and descriptions listed in the ADaM Programs table (ADRG Section 7.1) are consistent with the Analysis datasets in section 5.2 Analysis Datasets
- Empty cells checks across the ADRG tables, ADaM CT (in define 2.1/ADRG), ADaM IG Cross checks (define.xml/ADRG)
- **Detailed check list is available in the backup slides**

Inconsistency Checks- Example

ADRG

5.2 Analysis Datasets

Dataset Dataset Label	Class	Efficacy	Safety	Baseline or other subject characteristics	PK/PD	Primary Objective	Structure
ADSL Subject Analysis Dataset	ADSL			X			Multiple record per subject
ADAE Adverse Event Analysis Dataset	OCCDS						One record per subject per each AE recorded in SDTM AE plus an additional records from SDTM.FA severity changes are recorded

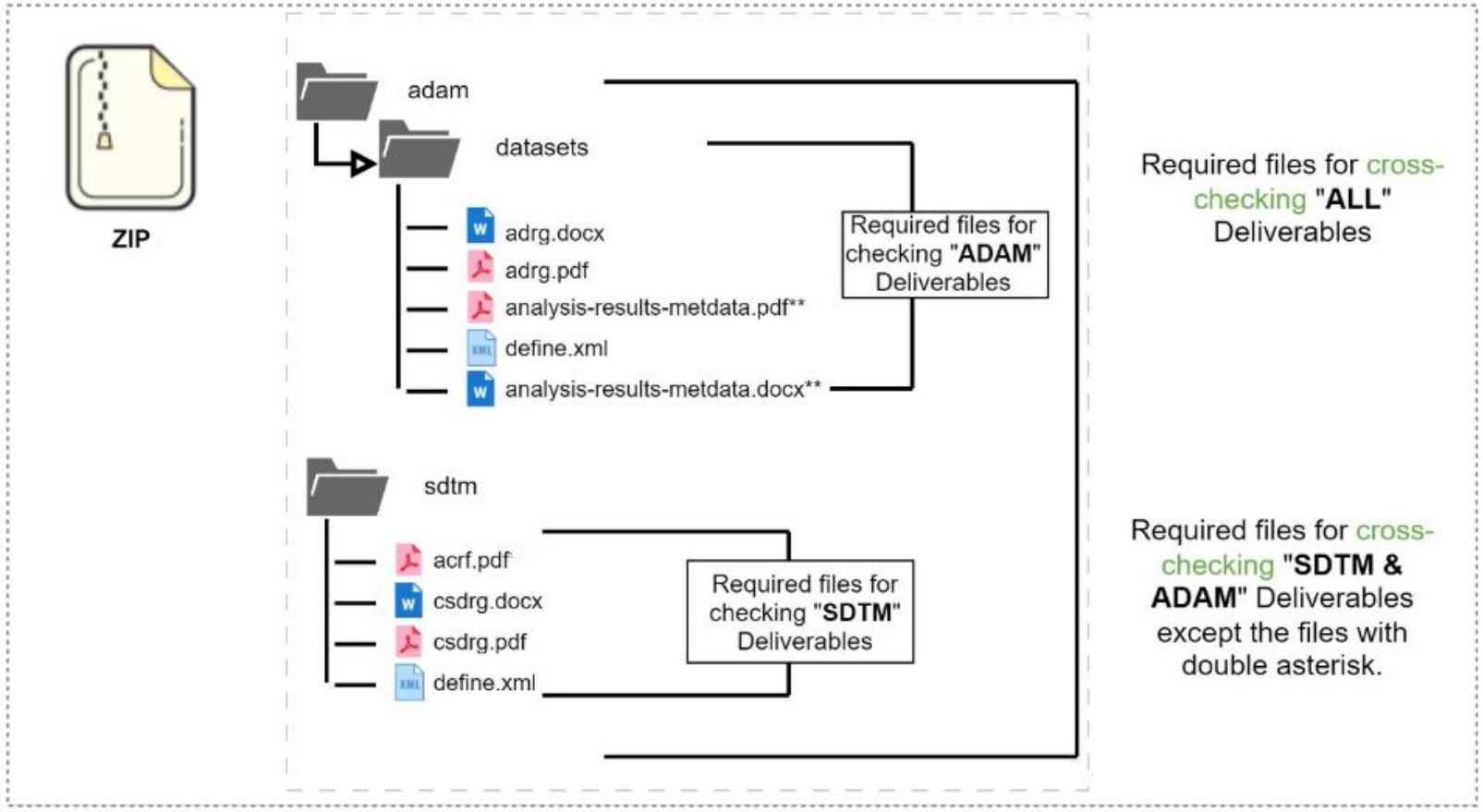
Discrepancy

ADaM define.xml

Dataset	Description	Class	Structure	Purpose	Keys
ADSL	Subject-Level Analysis	SUBJECT LEVEL ANALYSIS DATASET	one record per subject	Analysis	USUBJID
ADLB	Analysis Dataset for Lab Grade	BASIC DATA STRUCTURE	One record per subject, parameter (test and unit) and analysis time point	Analysis	USUBJID, PARAMCD, AVISIT ADT
ADQSADAS	ADAS-Cog Analysis	BASIC DATA STRUCTURE	One record per subject per parameter per analysis visit per analysis date	Analysis	USUBJID, PARAMCD, AVISIT ADT

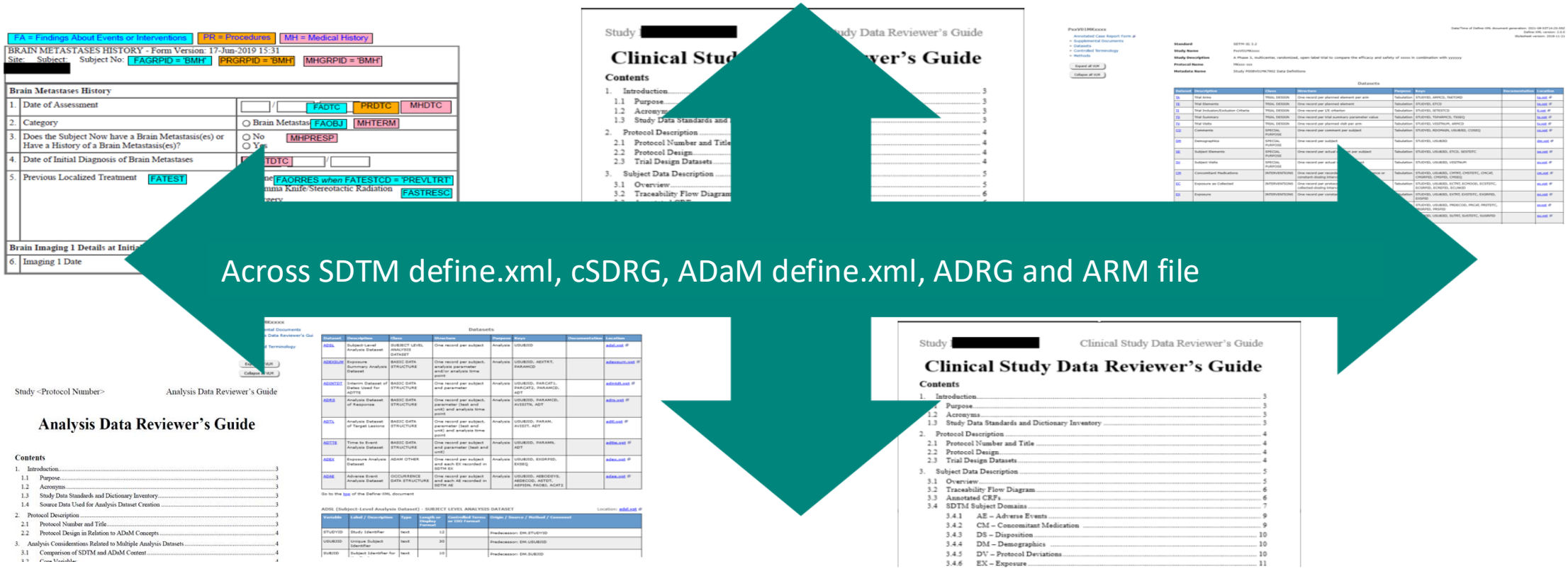
Discrepancy

Input Files Per Check Type



Latest release has functionality that includes checks for .xpt files and program files(.txt) and ARM document.

SDTM & ADaM check Type Details



- SDTM version, study name, protocol number/title, MedDRA/WHODD versions are consistent across cSDRG, ADRG and SDTM/ADaM defines.
- **Detailed check list is available in the backup slides**

Inconsistency Checks- Example

1.3 Study Data Standards and Dictionary Inventory

Standard or Dictionary	Versions Used
SDTM	SDTM 1.4/SDTM IG v3.2
Controlled Terminology	CDISC SDTM CT 2018-06-29
Data Definitions	define.xml v2.0
Medications Dictionary	WHO Drug Dictionary GLOBALB3Mar21
Medical Events Dictionary	MedDRA v24.0
LOINC	NA
TAUG	NA
Other standards	Adverse Events in AE domain are graded based on NCI CTC… version 4.03 Lab values are graded based on NCI CTCAE version 4.03

SDTM define.xml

External Dictionary	Dictionary Version
MED-RT	2021-04-05
MedDRA	24.0
NDF-RT	2021-04-05
SNOMED	2021-03-01
UNII	2021-04-12
WHODD	GLOBALB3Sep20

1.3 Study Data Standards and Dictionary Inventory

Standard or Dictionary	Versions Used
SDTM	SDTM v1.3 / SDTM IG v3.1.3
SDTM Controlled Terminology	CDISC SDTM CT 2016-03-25
ADaM	ADaM Model Document 2.1 ADaM Implementation Guide v1.1 ADaM Basic Data Structure for Time-to-Event Analysis v1.0 ADaM Structure for Occurrence Data (OCCDS) v1.0
ADaM Controlled Terminology	CDISC ADaM CT 2020-03-27
Medications Dictionary	Merck Proprietary Dictionary
Medical Events Dictionary	MedDRA v22.1
Data Definitions	define.xml v2.0, define.pdf
TAUG	NA
Other standards	Adverse Events in AE domain are graded based on NCI CTCAE version 4.0

Discrepancy

ADaM define.xml

External Dictionary	Dictionary Version
MedDRA	23.1

All Check Type Details

- All checks in the above check types.

Inconsistency Checks- Example

3.3 SDTM Subject Domains

Dataset – Dataset Label	Efficacy	Safety	Other	SUPPXX	Related Using RELREC	Observation Class
AE – Adverse Events		X		X	CM, DD, FA, HO, PR	Events
BS – Biospecimen				X		Findings
CM – Concomitant Medication	X			X	AE	Interventions
CO – Comment						Special-Purpose
DD – Death Details				X	AE	Findings
DM – Demographics			X	X		Special Purpose
DS – Disposition		X		X		Events
DV – Protocol Deviations				X		Events
EG – ECG Test Results				X		Findings
EX – Exposure		X		X		Interventions
FA – Findings About				X	AE, CM	Findings
HO – Hospitalizations				X	AE	Events
IE – Inclusion / Exclusion				X		Findings
LB – Laboratory Data		X		X		Findings
MH – Medical History		X		X		Events
MI – Microscopic Findings				X		Findings
PC – Pharmacokinetic Concentrations				X		Findings
PF- Pharmacogenomic Findings				X		Findings
PR – Procedures				X	AE	Interventions
QS – Questionnaire				X		Findings
RELREC - Related Records						Special-Purpose

5.2 Analysis Datasets

Dataset Label	Class	Efficacy	Safety	Baseline or other subject characteristics	PK/PD	Primary Objective	Structure
ADSL Subject-Level Analysis Dataset	ADSL			X			One record per subject
ADAE Adverse Event Analysis Dataset	OCCDS		X				One record per subject per each AE recorded in SDTM AE
ADBASE Baseline and Subgroup Analysis	OTHER			X			One record per subject
ADCM Concomitant Medications Analysis Dataset	OCCDS		X				One record per subject per each CM recorded in APERIOD

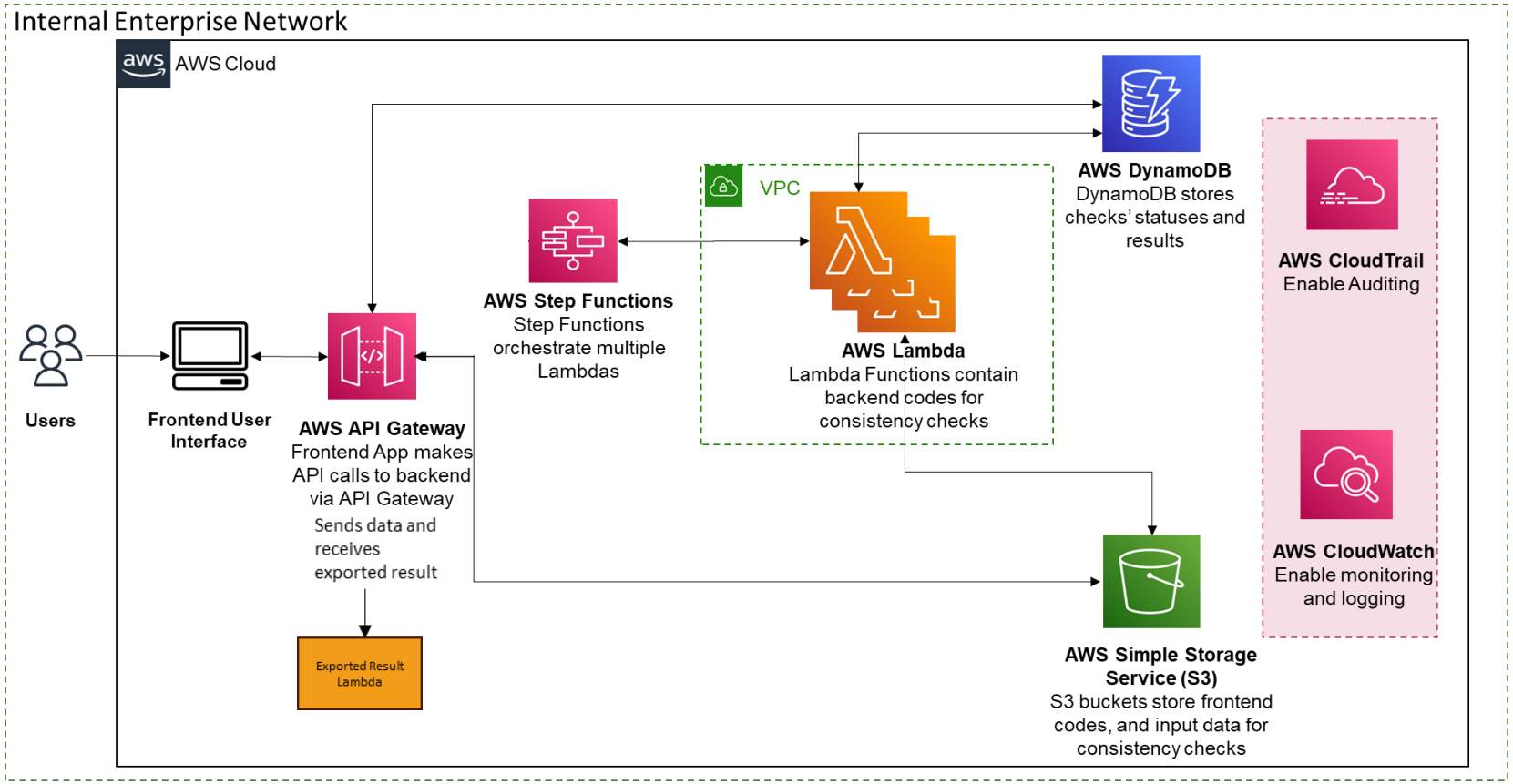
Discrepancy

7.1 Analysis Dataset Creation Programs

Analysis datasets are created from SDTM domains.

Analysis Dataset	Description	SAS Program Name	Input Dataset	
			Tabulation	Analysis
adsl.xpt	Subject-Level Analysis Dataset	adsl0gen.txt	dm.xpt, suppdm.xpt, ex.xpt, suppex.xpt, dd.xpt,	N
adae.xpt	Adverse Event Analysis Dataset	adae0gen.txt	ae.xpt, suppae.xpt, cm.xpt, suppcm.xpt, relrec.xpt	adsl.xpt
adcm.xpt	Concomitant Medications Analysis Dataset	adcm0gen.txt	cm.xpt, suppcm.xpt, relrec.xpt	adsl.xpt

High Level Architecture



Information Extraction/Natural Language Processing (NLP)

- Information Extraction
  - Annotations from acrf.pdf
  - Fields from define.xml
  - Headers from DRGs
  - Tables from DRGs
  - Unstructured content from DRGs
- Document parsing in Python
  - define.xml (lxml)
  - DRGs (Python-docx,lxml)
  - acrf.pdf (Fitz)
- Exporting results in excel format
  - xlsx writer

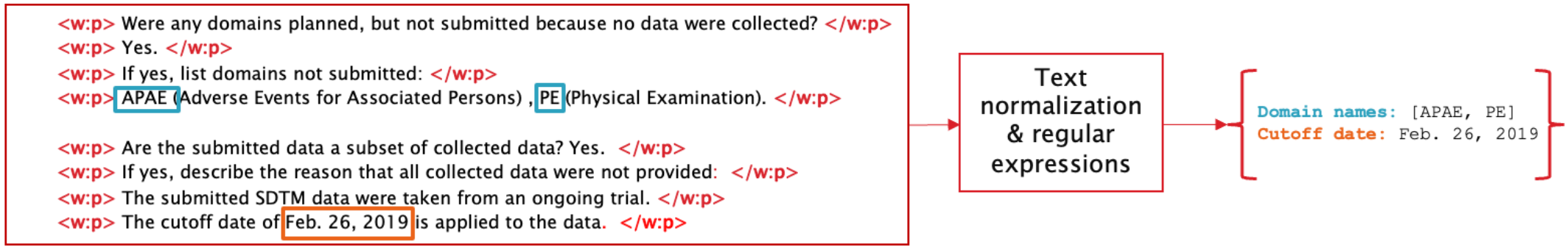
6. Data Conformance Summary

6.1 Conformance Inputs

Specify the software name and version for the analysis datasets	**Yes, Pinnacle 21 Enterprise v4.2.1 Validation Engine Version 2010.1**
Specify the version of the validation rules (i.e. CDISC, FDA) for the analysis datasets	**CDISC ADaM Conformance rule v2.0**
Specify the software name and version for the define.xml	**Pinnacle 21 Enterprise v4.2.1**
Specify the version of the validation rules (i.e. CDISC, FDA) for the define.xml	**FDA Validation rule v1.3**
Provide any additional compliance evaluation information	**Not Applicable**

Information Extraction/NLP

- Natural Language Processing (NLP)
  - Rule-based approach for setting up the checks/lambdas
  - Text normalization (extra spaces, quotes/special characters)
  - Pattern matching using regular expressions



Benefits of Using the Tool

- Valuable addition to the process as it improves document **quality** and **saves time** compared to manual checking.
- Tasks that took **1-2 days** per study can now be completed in **5 minutes** .
- Especially useful when multiple studies are included in a single submission, resulting in more time savings.
- Programmers can focus on other tasks, **increasing productivity and efficiency**.
- Reduces the risk of human error and ensures consistent and reliable results.
- Also, useful when validating **outsourced studies/partners and external vendors'** deliverables.
- Manual checking cannot be thoroughly performed every time updates are made to deliverables due to time constraints, but the tool solves this issue.

Conclusion

- Greatly enhances the quality and speed of the document review process while reducing the risk of errors and non-compliance.
- Use of the tool can lead to significant improvements in the document review process, making it an essential component of any efficient and effective process.
- Innovative technologies, such as natural language processing and text normalization techniques, enable the transformation of unstructured data into structured data.

Name : Majdoub Haloui
Affiliation: Merck & Co., Inc.
Email: majdoub.haloui@merck.com

Name : Hemu Shere
Affiliation: Merck & Co., Inc.
Email: hemu.shere@merck.com

Name : Dileep Penmetsa
Affiliation: Merck & Co., Inc.
Email: dileep.kumar.penmetsa@merck.com

Name : Surabhi Dutta
Affiliation: Merck & Co., Inc.
Email: Surabhi.dutta@merck.com

Majdoub Haloui
Hemu Shere
Dileep Penmetsa
Surabhi Dutta

Backup Slides

SDTM Check Type Details

cSDRG and SDTM define Checks:

- All SDTM subject domains referenced in the cSDRG are present in the SDTM define.xml and have consistent labels.
- The largest CRF page number referenced in the SDTM define.xml must be equal to or less than the last page number of the aCRF.pdf.
- When the origin of a variable in the SDTM define.xml is set to CRF, it must be annotated in the aCRF and the CRF page number referenced in the define must match the CRF page number where the variable is annotated.
- All domains listed under the question 'Were any domains planned, but not submitted because no data were collected?' in the cSDRG (section 3.1 Overview) are annotated in the CRF but do not exist in SDTM define.xml.
- At least one of the columns in the "Subject Domains" table in the cSDRG should be checked for efficacy, safety, or other/baseline subject characteristics.
- The study name, Protocol Number, and title are consistent across the cSDRG and the define.
- The MedDRA, WHODD and LOINC versions are consistent across the cSDRG and the define.

Hyperlink Checks:

- Annotated Case Report Form (aCRF) link is present in the SDTM define.xml bookmark and links to acrf.pdf.
- Study Data Reviewer's Guide link is present in the SDTM define.xml bookmark and links to csdrg.pdf.
- Links to additional supplemental documents.

ADaM Check Type Details

ADRG and ADaM define Checks:

- All ADaM datasets referred to in Section 5.2 of the ADRG (Analysis Datasets) are present in the ADaM define.xml, and their labels and structure are consistent.
- At least one of these columns should be checked for a dataset “Efficacy, Safety, Baseline or other subject characteristics, or PK/PD” in the ‘Analysis Datasets’ table in the ADRG.
- All analysis datasets and descriptions listed in the ADaM Programs table (ADRG Section 7.1) are consistent with the Analysis datasets in section 5.2 Analysis Datasets.
- The study name is consistent across the ADRG and the define.
- The protocol Number and title are consistent across the ADRG and the define.
- MedDRA, WHODD versions are consistent across the ADRG and the define.

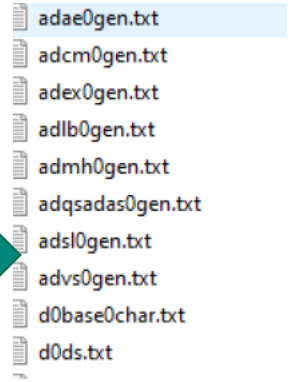
Hyperlink Checks:

- Analysis Data Reviewer's Guide link is present in the ADaM define.xml bookmark under ‘Supplemental Documents’ and links to the adrg.pdf.
- Analysis Results Metadata link is present in the ADaM define.xml bookmark under ‘Supplemental Documents’ and links to analysis-results-metadata.pdf.
- Links to other supplemental documents.

Analysis Results Metadata (ARM) File

2. Analysis Results Metadata Summary

Table Reference	Table Title	SAS Program Name (programs)	Input Datasets
[Refer to Citation Construct Job Aid] *Citations in square brackets cannot straddle two pages*	Disposition of Subjects	d0dispositon0asat081102gen.txt *Do not wrap filename*	adsl.xpt *Use ENTER after each dataset. Do not wrap filename*



3. Analysis Results Metadata Details

Table Reference: [Refer to Citation Construct Job Aid] *NOTE: Citations in square brackets cannot straddle two pages and cannot contain carriage return.*	
Analysis Result	*<include analysis result of interest>*
Analysis Parameter (s)	*<add useful parameters that is the focus of the analysis result>*
Analysis Variable (s)	*<include specific variables that will be analyzed>*
Analysis Reason	*<define the reason for including the analysis>*
Analysis Purpose	*<define the purpose of the analysis>*
Data References (incl. Selection Criteria)	*<add selection criteria that describes the analysis subset used to derive a result for a display>*
Documentation	*<Optional section intended to guide reviewers to external references such as protocol>*
Programming Statements	*<add program code used to perform the specific analysis and show how the specific analysis result was created>* *Example: [SAS Version 9.4],* [PxxxVxxMKxxxx: programs- *d0disp0msih0endogen.txt*]

- Cross check Program name and extension in the ARM file with ADaM programs directory (ADaM programs zip file).
- Check Input datasets name and extension in the ARM file with ADaM programs directory dataset names.
- Cross check Table Reference with Programming Statements in section 3.0

SDTM & ADaM check Type Details

- **Checks across cSDRG and ADRG:**
  - SDTM and SDTM IG versions.
  - SDTM Controlled Terminology version.
  - TAUG and Other Standards.
  - Data cutoff date.
  - Dataset used for Efficacy and/or Safety.
  - Pinnacle 21 software version and Validation Engines.

- **Checks across cSDRG, ADRG, SDTM and ADaM defines:**
  - Study name.
  - Protocol Number and title.
  - MedDRA and WHODD versions.