PHUSE US Connect 2021

Pinnacle 21: Improving data fitness score for e-submission

Renu Motwani, Bayer, Mississauga, Canada

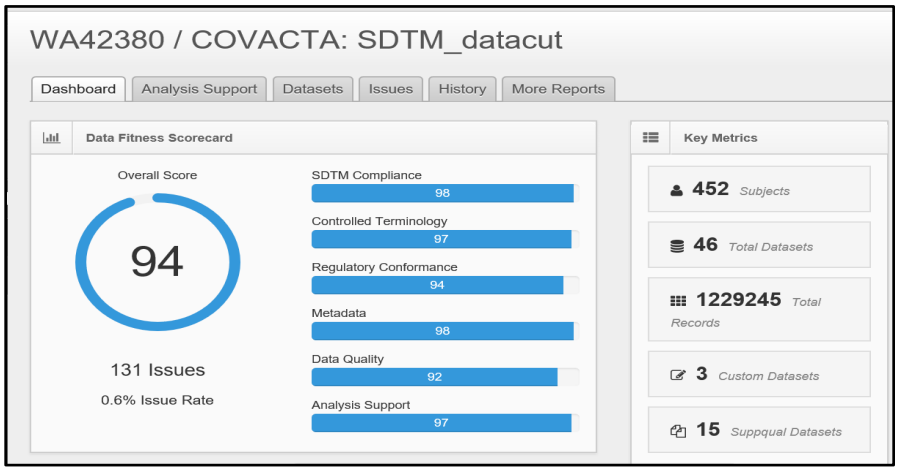
ABSTRACT

'Pinnacle 21 Enterprise' is used by regulatory agencies (FDA and PMDA) for validation of datasets. Pinnacle 21 Enterprise is designed to ease regulatory submission preparation, manage standards for datasets, and provide continuous compliance. The robustness of data is assessed with respect to six quality dimensions in Pinnacle 21 Enterprise. The fitness of data in terms of six quality dimensions including standard compliance, data quality, and more. The FDA does not base rejection purely on the data fitness score. The score gives an overall quick indication of data package quality across different functional areas and addressing potential issues before submission avoid costly delays. To improve validation score we need to make sure all essential components are available for complete study validation (data, metadata, control terminology, aCRF and RGz). The severity of issues will reduce your score, fixing rejects will greatly help increase your data score. Running validating data and specification (define.xml) brings scores up. We must ensure that data and specification have 1-1 match. Updating controlled terminology that has true values according to your study data also impacts the score and fixing various errors/warnings in P21.

INTRODUCTION

Today, Data Conformance is a standard and required part of regulatory submissions. Regulatory agencies expect sponsors to validate their study data before submission and either correct or explain discrepancies in the Reviewer's Guide. Mistakes in data validation may be costly as they can result in delays, unnecessary information requests, or even Refuse-to-File (RTF) or Refuse-to-Receive (RTR).

Is your data "fit for use"? Measure your "Data Fitness Score" with the same tool used by FDA and PMDA. See how your data performs across six quality dimensions, including standard compliance, data quality, and more. Pinnacle21 helps in viewing organizing and tagging various data issues, and assign tasks as needed. This is a great way to communicate and collaborate with CROs, developers, and other team members to identify and resolve issues quickly. Pinnacle 21, dashboard gives an overview of where your data stands for a specific study. It gives details on issues and even standard rules to follow to make your data fit for submission. Fixing rejects/ errors/warnings helps a lot in increasing the data score. In addition, addressing the issues with define.xml, metadata and control terminology also helps in building the score. Last but not the least, providing explanations to issues not able to fix, and closing the issues does help a lot with data score. These issues explanations are submitted in Reviewer's guide with submission package.



Display 1. Pinnacle 21 Dashboard -Six quality dimensions affecting data score

SDTM COMPILANCE

A very crucial part of the entire process is defining the SDTM mapping specification; it begins with annotations of the CRF according to the SDTM structures by analyzing the raw data. This process is time consuming, doing it first time right will save time & efforts.

In order to be CDISC compliant, raw datasets must be mapped from the structure used in your clinical data management system (or another database) to the CDISC SDTM structure. This can be a long, complex and daunting process.

It's really important to have a good understanding of SDTM domains and their structure. The SDTM Implementation Guide (SDTMIG) is there to help with this. It gives a detailed overview of SDTM specifications and metadata for all SDTM domains. It includes guidance for producing SDTM datasets. If you make yourself familiar with the SDTMIG before you start mapping, it'll make the process much smoother.

Process to follow to have SDTM compliant data:

- Identify all the datasets you want to map.
- Identify all the SDTM datasets that correlate with those datasets.
- Get the dataset metadata.
- Get the SDTM dataset metadata that corresponds to Step 3.
- Map the variables in the datasets identified in Step 1 to the SDTM domain variables.
- Create custom domains for any other datasets that don't have corresponding SDTM datasets.

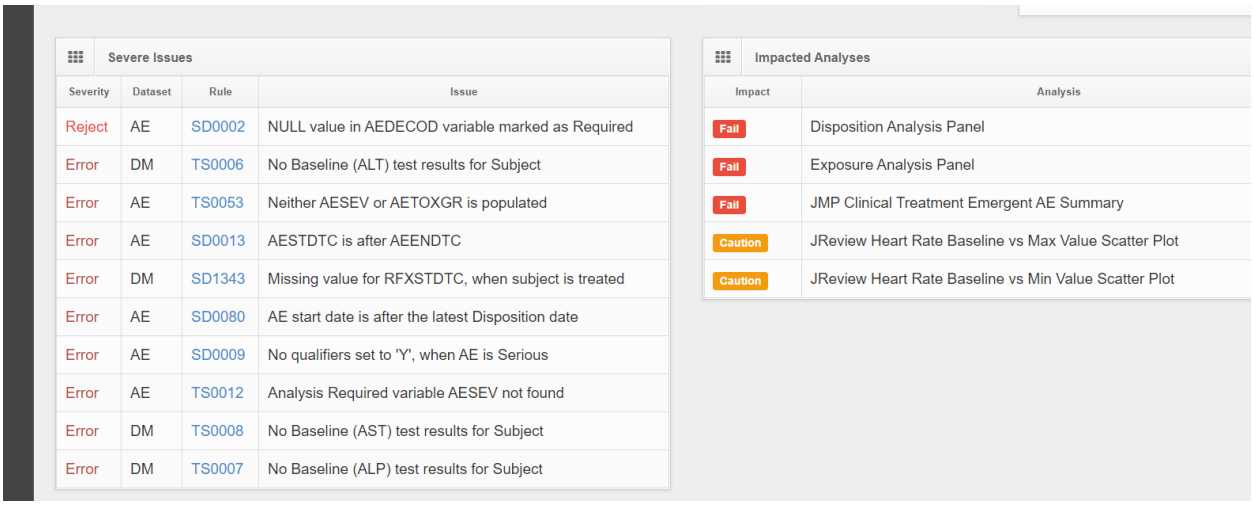
Some typical SDTM mapping scenarios that have created data issues in Pinnacle 21

- You need to rename some variables to be able to map to the corresponding SDTM variable. For example, if the original variable is GENDER, it should be renamed SEX to comply with SDTM standards. Variable attributes must be mapped as well as variable names. Attributes like label, type, length and format must comply with the SDTM attributes.
- The format that a value is stored in is changed. However, the value itself does not change. For example, converting a SAS date to an ISO 8601 format character string.
- Sometimes multiple variables must be combined to form a single SDTM variable. A non-SDTM variable might need to be split into 2 or more SDTM variables to comply with SDTM standards.
- Some SDTM variables are obtained by deriving a result from data in the non-SDTM dataset. For example, instead of manually entering a patients age, using the date of birth and study start date to derive it instead.
- Some variable values need to be recoded or mapped to match with the values of a corresponding SDTM variable. This mapping is recommended for variables with a code list attached that has control terminology that can't be extended. You should map all values in the controlled terminology, and not just the values present in the dataset. This would cover values that are not in the dataset currently but may come in during future dataset updates.
- There are situations where the structure of the non-CDISC dataset is completely different to its corresponding SDTM dataset. In such cases you need to transform its structure to one that is SDTM-compliant. For example, the Vital Signs dataset. When data is collected in wide form, every test and recorded value is stored in separate variables. SDTM requires data to be stored in lean form. Therefore, the dataset must be transposed to have the tests, values and unit under 3 variables. If there are variables that can't be mapped to an SDTM variable, they would go into supplemental qualifiers.

DATA QUALITY

Once we receive the data, we have to make sure that required variables are populated and permissible values are included to make the study data compliant. There is definitely lot of communication initially with data managers to make sure we have data collected correctly. If further clarification or correctness needed with data we need to communicate with sites to provide explanations or get the data corrected.

It is very important that we analyze issues on the basis of severity: Reject, Error & Warning.



Severe Issues

Severity	Dataset	Rule	Issue
Reject	AE	SD0002	NULL value in AEDECOD variable marked as Required
Error	DM	TS0006	No Baseline (ALT) test results for Subject
Error	AE	TS0053	Neither AESEV or AETOXGR is populated
Error	AE	SD0013	AESTDTC is after AEENDTC
Error	DM	SD1343	Missing value for RFXSTDTC, when subject is treated
Error	AE	SD0080	AE start date is after the latest Disposition date
Error	AE	SD0009	No qualifiers set to 'Y', when AE is Serious
Error	AE	TS0012	Analysis Required variable AESEV not found
Error	DM	TS0008	No Baseline (AST) test results for Subject
Error	DM	TS0007	No Baseline (ALP) test results for Subject

Impacted Analyses

Impact	Analysis
Fail	Disposition Analysis Panel
Fail	Exposure Analysis Panel
Fail	JMP Clinical Treatment Emergent AE Summary
Caution	JReview Heart Rate Baseline vs Max Value Scatter Plot
Caution	JReview Heart Rate Baseline vs Min Value Scatter Plot

Display 2. P21 report showing rejects/error

REJECTS: Are the critical issues that prevent review and automation processes. These should be fixed to make the data compliant. Below are few FDA rejection rules, we need to make sure that these are met while package submission.

Rule	P21	FDA eCTD
Trial Summary (TS) dataset must be included in every submission	SD1115	1734
Demographics (DM) dataset must be included in every submission	SD1020	1736
ADaM Subject level (ADSL) dataset must be included in every submission	AD0001	1736
FDA eCTD submissions must include a define.xml file for each study in Module 4 (nonclinical) and Module 5 (clinical)	DD0101	1736

Display 3. rejection rules from FDA

Errors: These are the issues reported, which needs to be fixed to make the data compliant. If not, then to be addressed in reviewers guide explaining the reason for not fixing.
Some of the examples:

-Start Date is after End Date

-Missing SDTM required variable

Warnings: These are the potential issues requiring manual review. Some examples:

-Terms added to Extensible CT

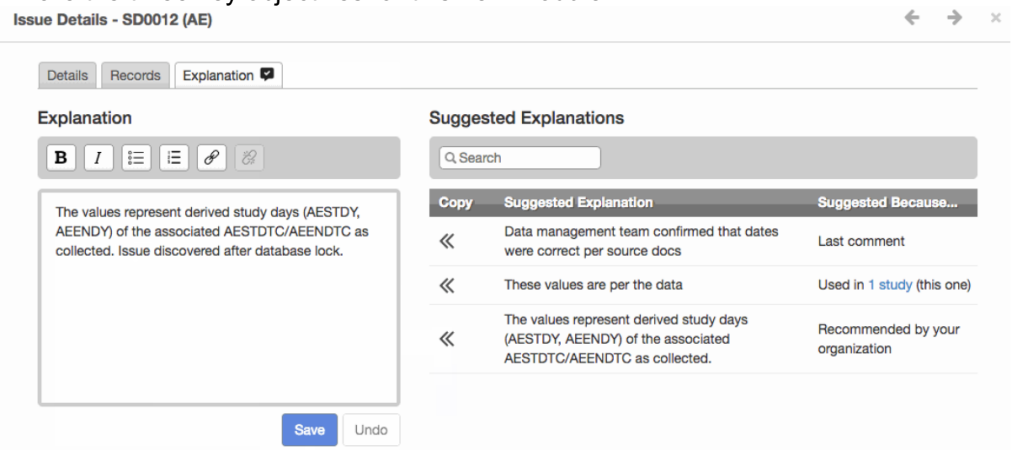
- Missing Units on Results
-False-Negative and False-Positive messages
False-Negative means that an issue exists but is not reported.
For example, Permissible variables with missing values for all records, downgraded from a ‘warning’ to ‘notice’ in SDTM.

False-Positive means that an issue is reported, but it is not a true issue. An example of a False Positive record is a duplicate record : This could not actually be a duplicate as P21 checks are only based on the key variables as per the check but we might have more variables which need to be looked at to see if it is a true duplicate within the dataset.

AUTOMATED ISSUE EXPLANATIONS IMPROVE CONSISTENCY

P21 Enterprise also improves compliance and productivity with automated issue explanation suggestions.

- The new Issue Explanations module improves validation issue management and provides greater consistency for reviewer's guides. It offers suggested issue explanations derived from templates uploaded by authorized users or from other studies in the same project.
- Shown below, the module provides a selection of either organization recommended explanation, the last comment, or explanation from another study. Increased transparency, consistency and efficiency are the three key objectives for this new module.



Display 4. explanation tab in P21 for providing reason or automated issue explanation

CONTROL TERMINOLOGY

Controlled Terminology validation rules identify discrepancies between the values a sponsor used in their data compared to allowable values of controlled terminology lists.

Incorrect or missing codelists create errors/warnings which should be either resolved or provide valid explanations in reviewers guide. There are instances where sponsors populate codelists only for variables that have standard CDISC Control Terminology (AEACN), but do not create study specific codelists. For example, for Category (--CAT), Subcategory (--SCAT), or EPOCH variables.

Suppqual domains are typically described using value level metadata but sponsors often leave out codelists for supplemental qualifiers that have controlled terminology. We recommend creating codelists only for variables where data was collected, derived or assigned based on a list of pre-specified terms. For example, if CMDOSU is collected using values from a drop-down menu in EDC system, it should reference a codelist in Define.xml file. However, if CMDOSU was collected as free text, a codelist is not necessary as it will result in presence of several hundred unique terms. We believe that in most cases study data codelists with more than 30-40 terms are impractical and are never used. It is strongly recommended in creating a separate codelist for each variable, instead of collapsing all codelist to several different variables, for e.g. all units variable codelist for several variables.

Sometimes the issue is reported for valid standard terms and their NCI codes, the possible explanations is an incorrect version of CDISC Control Terminology used for validation. For example, Error DD0028: Term/NCI Code mismatch in Codelist 'Laboratory Test Code'. A reported term 'CALB' and its NCI Code

'C125942' are actually correct standard term and its corresponding NCI Code in CDISC Control Terminology. In this case, we need to have a valid explanation in reviewer's guide to address and close this issue in Pinnacle 21.

Another example, DD0024: Invalid Term in Codelist 'No Yes Response' reporting 'N', 'U' or 'NA' terms which are standard terms in CDISC CT. Actual issue is due to invalid utilization of non-relevant terms for Flag variables like DTHFL, --BLFL or --PRESP where these 'N', 'U' and 'NA' terms are not applicable according to SDTM IG documentation. SDTM Flag variables are assigned to CDISC SDTM CT codelist (NY) 'No Yes Response' which includes 4 terms. However, SDTM IG specifies that these variables may only have either 'Y' or a missing value. For validation purpose, original NCI files of CDISC CT are extended by Pinnacle 21 with additional codelists including a subset of (NY) codelist limited to a single term 'Y'.

SAMPLE from NCI CT					
Code	**Codelist Code**	**Codelist Extensible (Yes/No)**	**Codelist Name**	**CDISC Submission Value**	**NCI Preferred Term**
C66767		No	Action Taken with Study Treatment	ACN	CDISC SDTM Action Taken with Study Treatment Terminology
C49503	C66767		Action Taken with Study Treatment	DOSE INCREASED	Dose Increased
C49504	C66767		Action Taken with Study Treatment	DOSE NOT CHANGED	Dose Not Changed
C49505	C66767		Action Taken with Study Treatment	DOSE REDUCED	Dose Reduced
C49501	C66767		Action Taken with Study Treatment	DRUG INTERRUPTED	Drug Interrupted
C49502	C66767		Action Taken with Study Treatment	DRUG WITHDRAWN	Drug Withdrawn
C48660	C66767		Action Taken with Study Treatment	NOT APPLICABLE	Not Applicable
C17998	C66767		Action Taken with Study Treatment	UNKNOWN	Unknown
SAMPLE from Sponsor CT Library					
C_CODE	**CODELIST_CODE**	**EXTENSIBLE**	**CODELIST_DESCRIPTION**	**SUBMISSION_VALUE**	**DECODE**
C49503	C66767		Action Taken with Study Treatment	DOSE INCREASED	Dose Increased
C49504	C66767		Action Taken with Study Treatment	DOSE NOT CHANGED	Dose Not Changed
C49505	C66767		Action Taken with Study Treatment	DOSE REDUCED	Dose Reduced
C49501	C66767		Action Taken with Study Treatment	DRUG INTERRUPTED	Drug Interrupted
C49502	C66767		Action Taken with Study Treatment	DRUG WITHDRAWN	Drug Withdrawn
C48660	C66767		Action Taken with Study Treatment	NOT APPLICABLE	Not Applicable
C17998	C66767		Action Taken with Study Treatment	UNKNOWN	Unknown

Display 5: Example of NCI CT and Sponsor CT Library (column headers do not need to be the same)

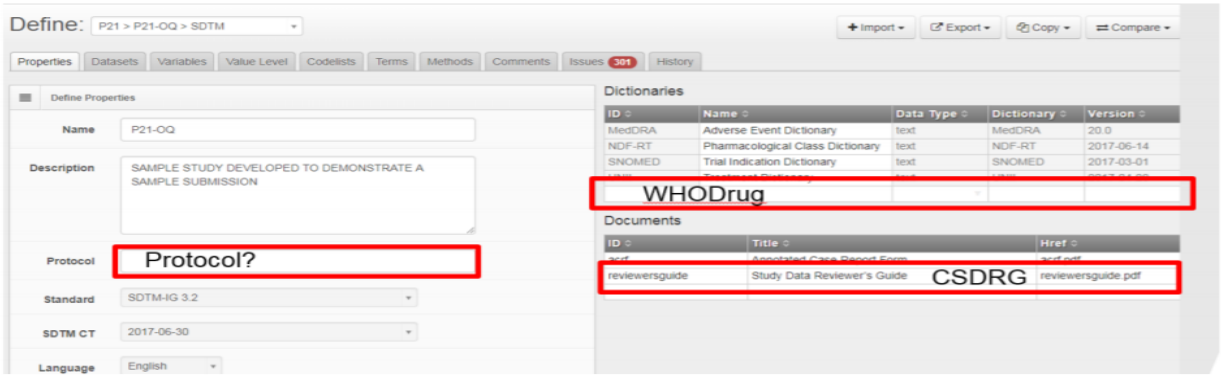
METADATA

Metadata validation rules look for issues with the define.xml, or inconsistencies between the define.xml and the study datasets.

The data definition file (Define.xml) is the most important part of the electronic dataset submission for regulatory review. It plays a critical role in automating processes as a major source of machine-readable study metadata. Some small errors in define.xml file may be critical for executing automated processes. If the version is incorrect, then a validation process will produce both false-positive and false-negative results.

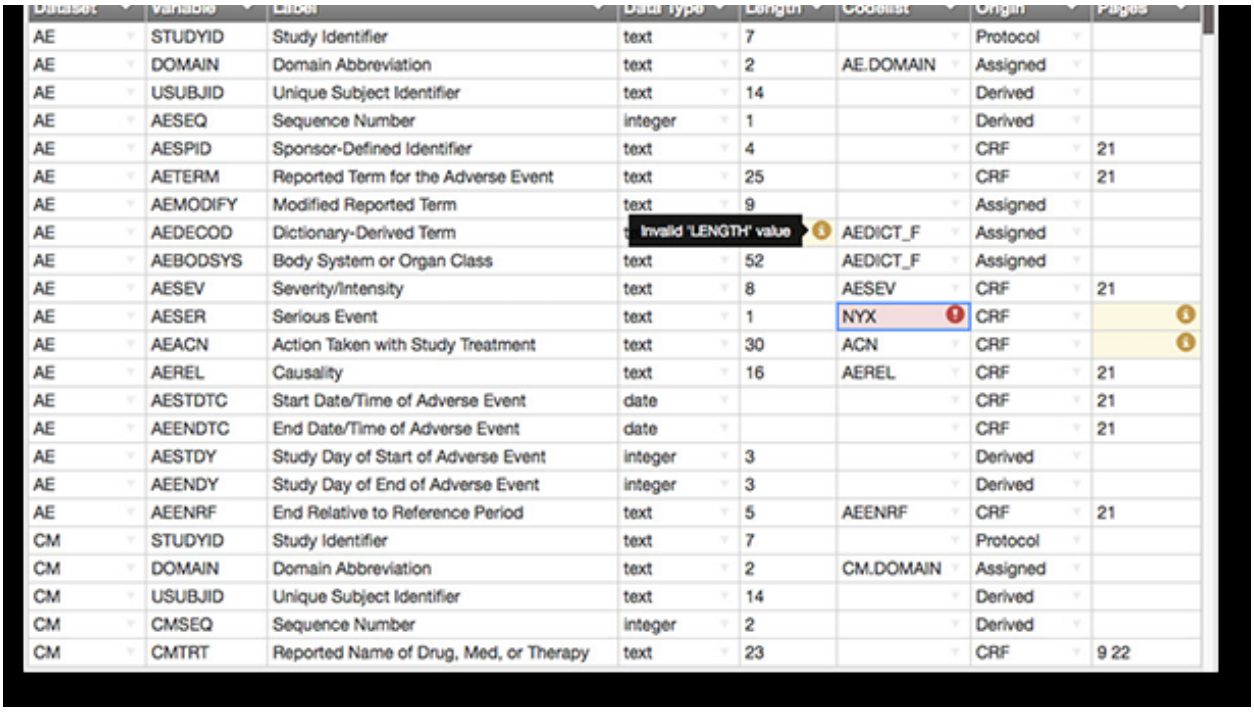
PMDA requires sponsors to fix validation Reject issues and explain all validation Errors. If study metadata is invalid or missing, then some compensatory actions are taken. We have to make sure we have the correct version of SDTM IG, SDTM CT and MedDRA entered during validation. For example, FDA DataFit uses define.xml file as a machine-readable source of MedDRA version. Incorrectly implemented (e.g., like Comment instead of Dictionary) MedDRA version also means missing machine-readable metadata. Missing decimal points in MedDRA version (e.g., "19" instead of "19.0") is PMDA Rejection criteria. Such issue must be fixed by Sponsor before proceeding further. A sufficiently documented Define file offers significant benefits. It provides detailed specification for datasets, variables,

codelists, data origins, and derivations, which allow reviewers to interpret submission data faster and move through the process more quickly.



Display 6. Pinnacle 21 enterprise DEFINE tab

Some of the most common issues while reviewing define.xml for validation are Missing Origin, and sometimes Origin=" CRF", but no reference to particular page(s) number mentioned. Sometimes, Inconsistency between origin and derivation is also noticed for example Origin="CRF Page" and Computation Method populated. In addition, it is also noticed that Origin="Derived" without detailed derivation algorithm These issues are highlighted in yellow or red when we go issue tab on define.xml. These needs to be filled in order for define.xml to be in good standing.



Display 7: shows the tabs in Pinnacle21 when working with Define.xml issues. These issues show rejects/errors/warnings which are related to study metadata or specs/eCRF uploaded to Pinnacle 21.

Uploading specs in Pinnacle 21 or creating define.xml and fixing most of the issues under DEFINE can bring up the data score.

REGULATORY CONFORMANCE

Regulatory Conformance validation rules look for violations of the guidance provided by regulatory agencies (FDA and PMDA) in their Technical Conformance Guides. These include missing datasets or variables requested by the regulatory agencies, and implementations inconsistent with the regulatory guidance.

According to the regulatory requirement, project team must ensure the Pinnacle 21 report is 'clean' before e-submission. 'Clean' means no error in the PINNACLE 21 report, and the remaining warnings can be explained clearly in the Reviewers' Guide section 4.2 Issues Summary. Addressing most of the warnings and errors caused by mapping issue or the raw data issue is very important. For mapping issues, programmers can update mapping logic and mapping program to correct them. That is under programmers' control. However, for raw data issues, it requires cross function collaboration to solve them. Some may need site correction and database to be updated. It normally takes quite long time to have the issue solved.

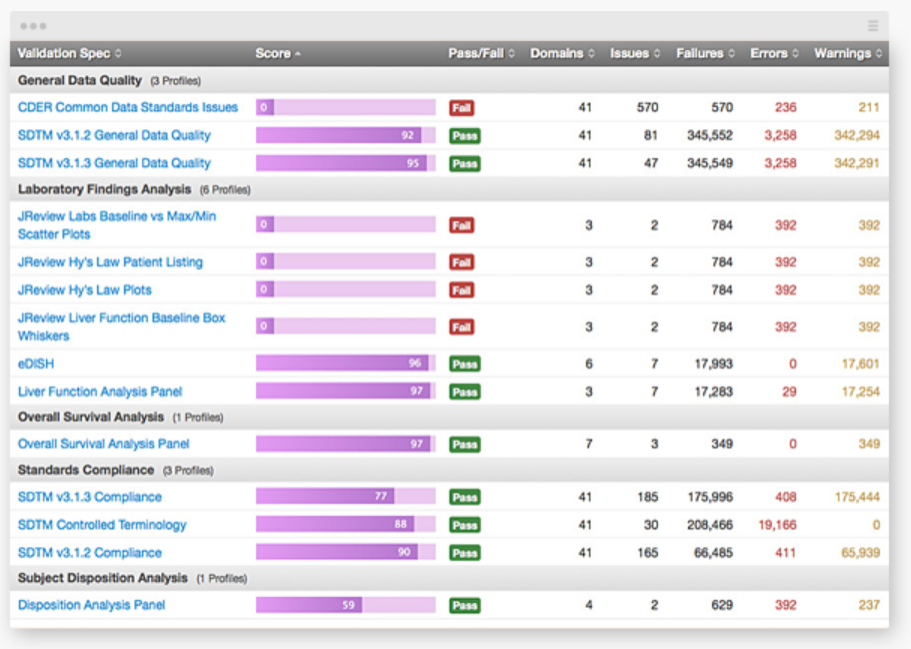
It is also mentioned in regulatory conformance check, to make sure that the EPOCH variable is provided in the appropriate domains. The confusion associated with this validation rule is due to seemingly conflicting information between the regulatory agencies (i.e, FDA and PMDA) and CDISC guidance. While regulatory agency guidance states that the EPOCH variable should be provided, CDISC guidance lists this as a Permissible variable, which some sponsors interpret as meaning that it is not necessary to include the variable.

Display 8. regulatory guidance states the EPOCH variable should be provided

Although CDISC lists this variable as a Permissible variable, it does not seem appropriate to disregard regulatory guidance and justify it based on the fact that CDISC specifies these as Permissible variables. Furthermore, simply including the EPOCH variable is not enough; sponsors should be sure to populate the EPOCH variable for each record possible and list a clear and complete derivation for this variable in the define.xml.

ANALYSIS SUPPORT

Analysis support tab helps in evaluating which section has passed or failed overall. It gives overview of what section needs attention with regards to score and provides total number of errors/warnings with that category.
This tab provides overall data quality as per SDTM versions and score on various analysis (lab, overall analysis and subject disposition) that have issues to be addressed.



Validation Spec	Score	Pass/Fail	Domains	Issues	Failures	Errors	Warnings
General Data Quality (3 Profiles)							
CDER Common Data Standards Issues	0	Fail	41	570	570	236	211
SDTM v3.1.2 General Data Quality	92	Pass	41	81	345,552	3,258	342,294
SDTM v3.1.3 General Data Quality	95	Pass	41	47	345,549	3,258	342,291
Laboratory Findings Analysis (6 Profiles)							
JReview Labs Baseline vs Max/Min Scatter Plots	0	Fail	3	2	784	392	392
JReview Hy's Law Patient Listing	0	Fail	3	2	784	392	392
JReview Hy's Law Plots	0	Fail	3	2	784	392	392
JReview Liver Function Baseline Box Whiskers	0	Fail	3	2	784	392	392
eDiSH	96	Pass	6	7	17,993	0	17,601
Liver Function Analysis Panel	97	Pass	3	7	17,283	29	17,254
Overall Survival Analysis (1 Profiles)							
Overall Survival Analysis Panel	97	Pass	7	3	349	0	349
Standards Compliance (3 Profiles)							
SDTM v3.1.3 Compliance	77	Pass	41	185	175,996	408	175,444
SDTM Controlled Terminology	88	Pass	41	30	208,466	19,166	0
SDTM v3.1.2 Compliance	90	Pass	41	165	66,485	411	65,939
Subject Disposition Analysis (1 Profiles)							
Disposition Analysis Panel	59	Pass	4	2	629	392	237

Display 9. Analysis support issues tab

CONCLUSION

Development of validation checks is a continuous process. There are always new standards, versions of standards, and business rules to be implemented. There are also frequent modifications of existing business rules. Pinnacle 21 is a great tool to check the data fitness and modify or correct any mapping related or data collected issues. To increase data score, we need to make sure we have study specification created correctly, use correct version of Control terminology and MedDRA. The severity of issues will reduce your score, fixing rejects will greatly help increase your data score. Running validating data and specification (define.xml) brings scores up. Following SDTM guide and it is highly recommended in creating a separate codelist for each variable and following regulatory compliance while validating data.

This paper shared experience with the most common data fitness issues observed across many sponsors and regulatory submissions. This also provided recommendations on how to ensure high quality submission data by evaluating the risk or potential impact of each issue and how each can be corrected or documented in reviewers guide. As a parting note, keep in mind that as more new tools are developed at FDA more requirements for higher quality data will emerge. Creating submissions data that is fit for use requires a continuous improvement process.

REFERENCES

- Data standards catalog
  https://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm
- Providing Regulatory submissions
  https://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm
- Define.xml schema validation
  https://www.cdisc.org/system/files/all/standard_category/application/pdf/definereport_v1_0.pdf
- FDA Business rules/Validator rules
  https://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm
- SDTM-Metadata submission guidelines
  https://www.cdisc.org/standards/foundational/study-data-tabulation-modelimplementation-guide-sdtmig/metadata-submission

- SDTM IG 3.2
  https://www.cdisc.org/standards/foundational/sdtmig
- Define-XML-2.0-Specification
  http://www.phusewiki.org/wiki/images/8/89/Define-XML-2-0-Specification.pdf

USEFUL LINKS

- https://www.pinnacle21.com/
- http://www.fda.gov/forindustry/datastandards/studydatastandards/default.htm
- http://cdisc.org/standards-and-implementations

CONTACT INFORMATION

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

Name: Renu Motwani
Company: Bayer
E-mail: renu.motwani@bayer.com