PharmaSUG 2017 - Paper SS05

Good Data Validation Practice

Sergiy Sirichenko, Pinnacle 21, Plymouth Meeting, PA

Max Kanevsky, Pinnacle 21, Plymouth Meeting, PA

ABSTRACT

According to FDA and PMDA guidance for regulatory submissions, sponsors are expected to perform study data validation and explain any issues that were not fixed. This is a new requirement introduced only a few years ago, which is why the industry is still struggling to achieve full implementation. In many cases, data validation is not performed correctly with sponsors having the wrong interpretation of validation results, or having invalid, confusing, or non-relevant issue explanations. The major reason for these problems is a lack of explicit and detailed documentation about the data validation process. The purpose of this presentation is to address this gap. We will summarize the regulatory requirements for data validation, discuss basic concepts and methodology, provide instructions on correct installation and usage of Pinnacle 21 Community open source tool, and review best practices for how to explain issues in study data.

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).

Let's consider the following scenario. A pharma company prepared a submission package, validated datasets, completed Reviewer's Guides, and submitted the application to PMDA. However, a few days later the application was returned due to an issue that the Sponsor was not aware of. How could this happen?

A few years ago, PMDA introduced special rejection rules for submission data. Some of these rules are related to CDISC Control Terminology (CDISC CT). For example, the rules for flag variables like --DTHFL and --BLFL, which may be populated by only the term "*Y*" or a missing value. Pinnacle 21 uses CDISC CT files provided in ODM format, but extends them for validation purposes. Specifically, Pinnacle 21 introduced a new (Y) codelist with a single term '*Y*', which is a subset of the original CDISC CT (NY) codelist that contains 4 terms *'N*', '*NA', 'U'* and '*Y*'. The (Y) codelist is then used to validate flag variables instead of the (NY) codelist.

There are several things that might have gone wrong with the Sponsor above. One potential cause, is that the Sponsor didn't correctly configure the validation software. When the Sponsor downloaded Pinnacle 21 Community (P21C), the installation package included only the 5 most recent versions of CDISC SDTM terminology. The Sponsor might have needed an older version of CT and therefore decided to download the original CDISC CT files from the NCI website, instead of the extended CDISC CT files provided on Pinnacle 21 website. P21C would have validated the datasets using these files, but due to the missing extensions, the check for the flag variables would have not executed.

Another potential cause might be the Sponsor's decision to use an outdated version of P21C or OpenCDISC Validator. This is common as many sponsors believe that they should use the version of the Validator that was originally available at study start up. However, older versions of Validator do not include the PMDA rejection rules.

As a result, in both cases, some or all PMDA rejection rules were not executed leaving the Sponsor uninformed of the potential issues, resulting in Sponsor needing to fix and re-submit data to fit current regulatory expectations.

A NEED FOR GOOD DATA VALIDATION PRACTICE

Unfortunately, most regulatory submissions today have major problems with *Data Compliance*. Typical examples are

- Incorrect installation and configuration of Validator
- Incorrect execution of data validation
- Missing validation for define.xml
- Incorrect interpretation of validation results
- Lack of understanding of the validation process
- Invalid, incorrect, or irrelevant explanations for reported validation findings
- Lack of knowledge on how to evaluate and handle specific data issues

The major reason for these problems is lack of knowledge

- *Data Conformance* is a new concept and the industry is still working on its implementation
- Simplicity and ease of use of P21C Validator may be misleading in term of its correct use
- There is a lack of documentation and education resources for data validation

To fix these deficiencies we need to introduce education and training on "*Good Data Validation Practice"*, which will help answer the following questions:

- What is data validation?
- How to configure Validator?
- How to perform data validation?
- How to interpret validation results?
- How to evaluate risk of data issues?
- How to fix data errors?
- How to explain data issues?

The content of this paper will focus on the first question. It's extremely important to understand the concept of *Data Conformance*, the scope of data validation, its limitations, and practical implementation.

STANDARDS AND DATA VALIDATION

P21C Validator (formerly OpenCDISC Validator) was created to help with implementation of CDISC standards. Standardized data allows automation through re-use of programming code, which can be enforced by data validation tools like P21C.

Today, some vendors can provide study data in SDTM format just one (1) week after the first subject first visit. It provides an opportunity to execute data management queries and clean collected data by standardized programming code.

With standardized data, FDA and PMDA also attain the capability to check and enforce data quality. And most importantly, it can be done early in a regulatory review process. So, any potential data issues can be identified, evaluated, and communicated with sponsor for resolution if needed. Such a process increases quality of submission data and helps to avoid costly delays in regulatory review later.

Data Conformance is a required part of Reviewer's Guide. Sponsor is expected to perform data validation before submission, document and explain all non-fixed reported data issues.

FDA DEFINITION FOR DATA VALIDATION

FDA's Study Data Technical Conformance Guide says that

> "*Study data validation* ***helps to ensure*** *that the study data are* ***compliant***, ***useful***, *and will* ***support meaningful review and analysis***"

Today we are observing a changing landscape in data validation with focus shifting from *Standards Compliance* to *Data Quality.*

Overall adaption of data standards is accomplished. CDISC standards are required by the regulatory agencies. Most studies submitted today follow SDTM, ADaM, and Define-XML standards. So, *Standards Compliance* can now be considered a commodity. CDISC standards are a pre-requisition for data processing, just like SAS® XPT format is a pre-requisite for submission.

Everybody understands that the major value of data is its content, not its format. Medicines are evaluated on their safety and efficacy, not the SDTM format. So, it's no surprise that regulatory reviewers are interested more in *Data Quality* and less on *Standards Compliance*.

TYPES OF DATA QUALITY ASSESSMENTS

In terms of practical implementation, there are 3 general types of data quality assessments. It's important to understand the difference between them.

P21 ERRORS

The original *Severity* in P21 Validator was a property of computational algorithm, rather than an estimation of *Risk* or *Impact* of reported data issue.

P21 *Errors* are *Checks* based on executable algorithm which produce issue reports with 100% confidence. For example,

- *Start Date is after End Date*
- *New terms in Non-Extensible CT* - Any new term is invalid in this case

P21 WARNINGS

P21 *Warnings* are *Reports* of potential data issues for manual review, which will decide if this is a real issue or not. For example,

- *New terms in Extensible CT* – New terms may be added. However, they should not overlap with existing standard terms. Such assessment cannot be automated and need manual review by a subject matter expert
- *Missing Units on Results* – Some assessments do not have units. Implementation of this business rule can filter out some common cases. However, at this point it cannot handle everything correctly. So, manual review is expected.

P21 DATA FITNESS REPORTS

P21 Data Fitness Reports are available only in Enterprise version. They represent additional diagnostics and a source of useful information. Here are some examples:

- *Death info reconciliation* – It's a list of all subject death information in study data. Sometimes a subject death is not collected in expected way with missing *DEATH* records in DM, AE, DS domains. However, subject may have information about their actual death stored in very unexpected locations like CO, SUPPAE, DV, SUPPDV or any other domains. Consistency in subject *DEATH* info across DM, DS, AE and other domains is also imported. Due to lack of standardization for this type of information, data quality assessments performed by data fitness reports with flexible study-specific structure are superior to regular checks.

- *Quality of MedDRA Coding* – This report helps Reviewers with evaluation of submitted MedDRA coding by matching collected text for Adverse Events with MedDRA PT, LLT coding.
- *Content of SUPPQUAL domains* - This report provides a quick and convenient familiarization with non-standard study-specific data.
- *Missing BASELINE* – There is *Missing Baseline Records* check which produces a list of subjects with this issue. An additional report includes all records for those subjects. It helps understand why the issue exist? For example, some subjects may not have assessments before dosing, some may not have Baseline Flag populated for existing records.

MAJOR IMPLEMENTATION CHALLENGE

Data Quality is defined by *intended use* and as *absence of errors which matter.*

ALL USERS ARE DIFFERENT

A challenge is that all users are different. Each user is interested only in data issues directly related to him/her. For example,

- A Data Warehouse Specialist may be interested only in issues that prevent data from loading into the data warehouse. Examples are standard compliance, data consistency, and study metadata.
- Data Analyst may be interested in issues related to uploading into analytic tools and running specific analysis. Examples are data consistency, standard compliance, and missing data.
- Regulatory Reviewer needs study data which can support meaningful review and analysis. In this case, content related issues are more important than standards compliance.

LEVEL OF DETAILS

While each user cares only about particular subset of data issues, there are also different expectations on level of details. For example,

- Project Manager may be interested in overall data quality for the study during its preparation for regulatory submission. *Is this study good enough? What is Data Quality Score for this study?*
- Programmer usually works on an issue level fixing programming errors.
- Data Manager handles data issues on data point level.

USER-SPECIFIC “SEVERITY”

A definition of Severity may also be different across users. FDA and PMDA have separate list of business rules where Severities are different between agencies. A recent version of FDA validation checks does not include Severity at all. PMDA also has a special Rejection Severity.

For most users, including FDA and PMDA, Severity is an assessment of Risk or potential impact of a data issue for a specific use. However, even within the regulatory agencies there are many different users with different needs and assessments of Severity.

P21 Severity is a property of check algorithm.

USER-SPECIFIC MESSAGE AND DESCRIPTION

P21 checks are designed by Programmers for Programmers. So, the messages and descriptions assume that users have good knowledge of CDISC standards. However, these messages and descriptions are not optimal for non-technical audience like Medical Reviewers. Clinical people do not understand technical language of P21 messages and need different reviewer-friendly text.

For example, instead of "*Inconsistent LBSTRESU*" a Medical Reviewer wants to see "*Conversion of collected results into the same units was not done for all records*".

"*Missing Seriousness Criteria for SAE*" should be presented as "*No seriousness qualifiers (AES*) were collected for all Serious Adverse Events (AESER=Y)*"

Data Management-friendly messages are expected to be data point level and include all diagnostics data and tracking info needed for issue resolution or generating DM queries. For example,

- *"Subject ABC-001 has potentially invalid Visit Dates: VISIT 2: 2017-02-19, VISIT 3: 2016-02-27"*
- *"Seriousness Criteria are missing for 'Myocardial Infarction' Adverse event of subject ABC-001 started on 2017-02-27 and flagged by Investigator as Serious"*

SCOPE OF P21C VALIDATION CHECKS

It's important to understand who is the *final user* for your study data. In our corner of the industry, the final users are FDA and PMDA Reviewers. Regulatory agency users are foremost interested in automating the review process, which means that they expect "minimal standard compliance" and review specific data elements.

Last year FDA announced their plan for a 2-steps data validation process. The first step will perform validation of standard compliance, which cover eCTD, CDISC, etc. In a case of major violations, submission data will be returned to Sponsor to fix and resubmit.

A first step in this direction was completed in October 2016 when FDA introduced the eCTD Technical Rejection Criteria [1]. The requirements state that submission data must include:

- define.xml
- Demographics (dm.xpt) and Trial Summary (ts.xpt) datasets in Tabulation data
- Subject Level Analysis Dataset (adsl.xpt) in Analysis data

At beginning of 2017 almost a half (45%) of "standardized" submissions failed these very basic requirements.

While most regulatory submissions are currently utilizing CDISC standards, their implementation is often quite loose. For example, Sponsor might create a high quality Define.xml file with a name of "*define2.xml".* Formally this would trigger the Rejection criteria because the expected file name is "*define.xml*". Why would such a minor inconsistency in file name cause a rejection? Isn't the Reviewer still able to open and read the contents of the file?

Yes, Reviewers as humans can easily understand that "define2.xml" is a Define-XML file. However, automated processes in between the Sponsor and the Reviewer, like the Clinical Trial Repository (Janus) data warehouse, cannot.

Another example is when the Sponsor specifies the version of MedDRA dictionary as a Comment instead of the special "ExternalCodeList" element in Define-XML file. It's easy for the Reviewer to find this information, but it's not the case for computers. Even a little mistake like a missing dot character can break programming code. The similar is about input data.

FDA said that their "minimal compliance" criteria will be defined and published later this year. They will be driven by need to automate processes utilized within the Agency. Therefore, Rejection rules from FDA are expected to be different compared to PMDA Rejection rules.

On the top of standard compliance there are also review-specific business rules. For example, Epoch (EPOCH) and Baseline Flag (--BLFL) variables are Permissible or are not present in some standard domains in SDTM Implementation Guide. However, this information is very useful to select records during study data analysis. That's why EPOCH and --BLFL variables are requested by FDA and PMDA.

In some cases, FDA even overrides CDISC standards. For example, SDTM's Arm (ARM) and Actual Arm (ACTARM) are Required variables and must have non-missing values for all study records. However, FDA Study Data Technical Conformance Guide asks to populate a missing value in ARM variable for all Screen Failure subjects and a missing value in ACTARM variable for all non-treated subjects. Instead of SDTM IG recommended terms "*Screen Failure", "Not Assigned"* and *"Not Treated"* programmers should use a missing value for regulatory submission to FDA.

To complicate matters even farther, PMDA requirements are different from FDA. Sponsors are expected to follow SDTM IG examples using "*Screen Failure", "Not Assigned" and "Not Treated"* terms. A missing value in ARM or ACTARM variables is a violation of PMDA Rejection rules. Therefore, implementation of missing values in ARM *for Screen Failure* subjects requested by FDA will be rejected by PMDA until its corrected according to PMDA business rules.

Some review-specific business rules may not be obvious or not clear for most users. For example, SD2236 check may produce validation messages that "*ACTARMCD does not equal ARMCD*". Here is no violation of SDTM compliance. It's quite a common case in study conduct. An actual goal of this check is to report all subjects who received wrong study treatment and provide explanations in Reviewer's Guide documentation. It facilitates a communication between Sponsor and Reviewer on study-specific data issues in advance and saves time by avoiding potential information requests from Reviewers after an initial submission.

Complexity of data validation is increasing. New Rejection rules from FDA and PMDA raised liability for correct implementation of *Data Conformance*. P21 Community helps many users. However, to ensure its right utilization it's important to understand the scope of the P21C project.

P21 Community is a free personal desktop tool for SAS Programmers to QC their work for regulatory submissions. For everything else, there are other tools including P21 Enterprise.

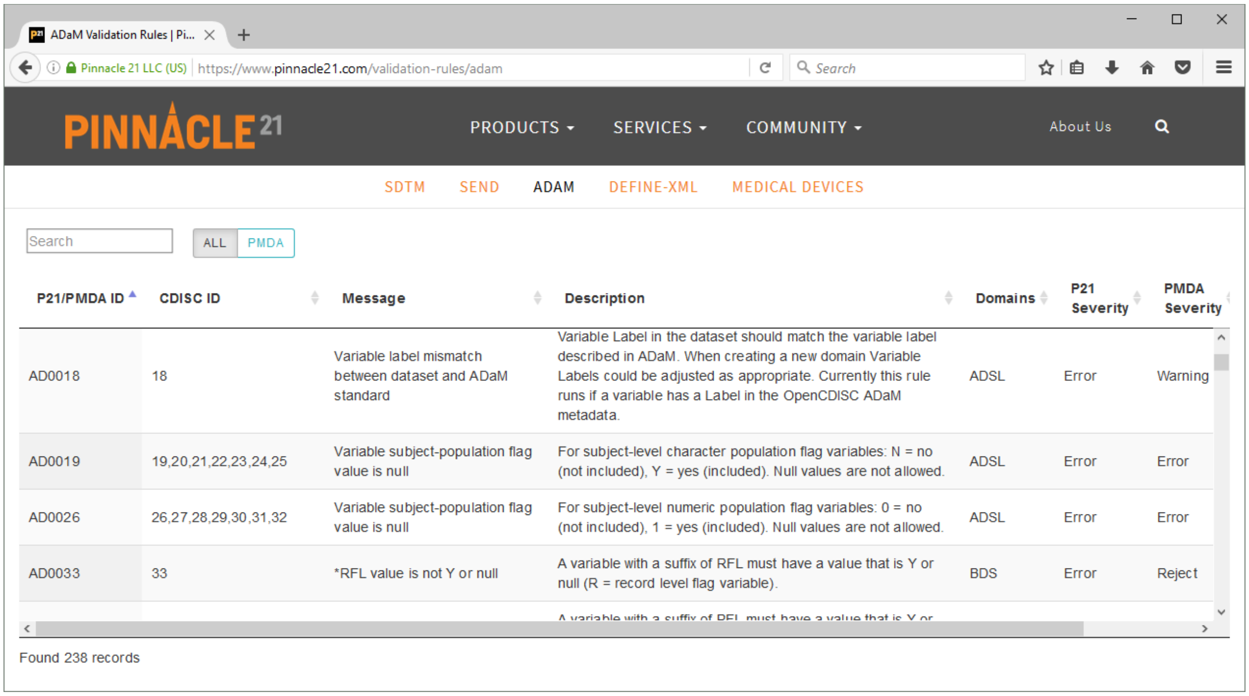
Implemented as an open-source project using Java and XML technology, P21 Community allows easy customization and integration. However, this flexibility was not intended for submission use case. An incorrect installation on a server or modification of original configuration files or validation engine code may results in incorrect validation reports which are not compatible with the Agencies business rules.

SOURCE OF DATA VALIDATION CHECKS

There are many stakeholders for *Data Conformance*. Development of data validation checks is a continuous process, which will mature but will never stop. Major sources of new business rules are

- Standards specifications like SDTM Model and IG.
    - Especially the new validation specific documents developed by CDISC like "CDISC ADaM Validation Checks" and "SDTMIG v3.2 Conformance Rules". These documents were created by extracting and summarizing all business rules incorporated in text Standard Model and Implementation Guide documents. SEND team is also working on an explicit set of validation checks for the Standard for Exchange of Nonclinical Data
- FDA/PMDA business rules available on agencies' websites
    - They represent high-level review and process related requirements published by the Agencies for public use
- Data management checks
    - If any business rules are applicable to most studies, they may be promoted to industry-wide rules
- Tool specific requirements
    - All tools and analysis have expectations on input data. Checking these requirements in advance allows to know if there are any potential problems with specific application to upload data or obtain meaningful results.
- Additional requests by users

There are no “exact” implementations of any specific business rules due to multiple stakeholders and programming limitations. There are cases when multiple business rules may be collapsed into a single executable check or otherwise. Several different checks are needed to cover a single business rule. P21 Validation reports have a special attribute “Publisher ID”. This rule traceability information can also be found on https://www.pinnacle21.com/validation-rules with a reference to IDs of the specific publisher.


ADaM Validation Rules | Pi...
Pinnacle 21 LLC (US) | https://www.pinnacle21.com/validation-rules/adam
Search
PINNACLE 21
PRODUCTS
SERVICES
COMMUNITY
About Us
SDTM
SEND
ADAM
DEFINE-XML
MEDICAL DEVICES
Search
ALL
PMDA

P21/PMDA ID	CDISC ID	Message	Description	Domains	P21 Severity	PMDA Severity
AD0018	18	Variable label mismatch between dataset and ADaM standard	Variable Label in the dataset should match the variable label described in ADaM. When creating a new domain Variable Labels could be adjusted as appropriate. Currently this rule runs if a variable has a Label in the OpenCDISC ADaM metadata.	ADSL	Error	Warning
AD0019	19,20,21,22,23,24,25	Variable subject-population flag value is null	For subject-level character population flag variables: N = no (not included), Y = yes (included). Null values are not allowed.	ADSL	Error	Error
AD0026	26,27,28,29,30,31,32	Variable subject-population flag value is null	For subject-level numeric population flag variables: 0 = no (not included), 1 = yes (included). Null values are not allowed.	ADSL	Error	Error
AD0033	33	*RFL value is not Y or null	A variable with a suffix of RFL must have a value that is Y or null (R = record level flag variable).	BDS	Error	Reject

Found 238 records

Display 1. ADaM checks with references to CDISC check IDs on pinnacle21.com

TRUE AND FALSE “FALSE-POSITIVE” VALIDATION MESSAGES

Data Validation is a diagnostics test. Like other diagnostics tests there are 4 possible cases:

Issue exists, and is reported	Issue does not exist, but it is reported
Issue exists, but it is not reported	Issue does not exist, and it is not reported

True Positive	False Positive
False Negative	True Negative

Figure 1. Contingency table

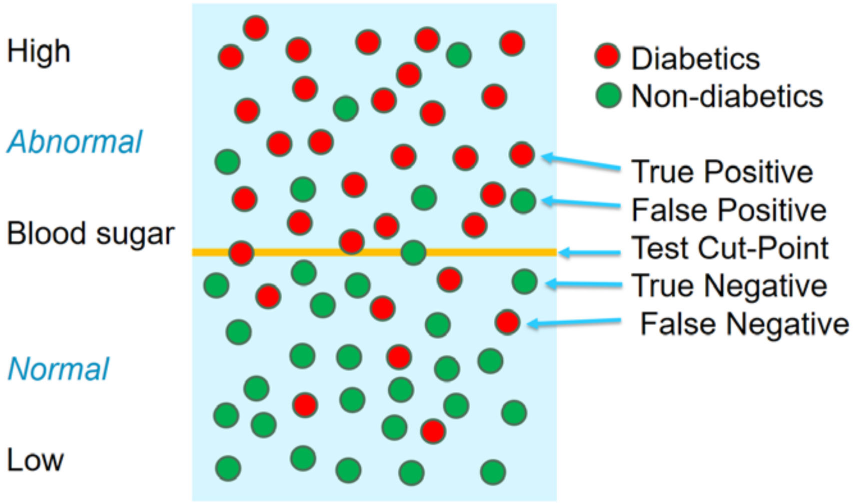


Figure 2. Diagnostics tests

All P21 Warning checks produce both False-Positive and potentially False-Negative messages due to nature of these business rules. An implementation may be tuned by better programming algorithms and simply moving of *Test Cut-Point.* Algorithms' tuning yields different specificities and sensitivities in detection of real data errors. In most cases a decrease of false-positive validation messages may be compromised by increasing number of non-reported actual issues in data.

What is more important? Different users have different expectations.

Minimizing False-Positives validation messages is desired when confirmatory diagnostics for Warnings is expensive. For example, extra work for data vendors may increase their expenses. Missing planned timelines for a submission is another common case, when false-positive messages are too costly.

Minimizing False-Negatives validation results is critical when penalty for non-fixed errors is high. For example, non-reported violations of Rejection FDA/PMDA rules will delay submission review. Low quality data may compromise study results.

FALSE-NEGATIVES

Almost nobody complains about False-Negative validation messages mostly due to missing validation messages in contrast to False-Positive issues. It's challenging to perform diagnostics to actual issues which are not reported. However, a data issue is still an issue regardless if it is reported or not.

It is important to remember that a Sponsor is responsible for regulatory submission including quality of study data. Sponsors can outsource work, but Sponsors may not outsource their liability to data vendors.

There are two major sources for False-Negative issues:

- Incorrect algorithm for existing checks
- New checks for implementation

For example, the check "*Date is after RFPENDTC*" was introduced in OpenCDISC v1.4.1 and removed in v2.0, which was limited to FDA official business rules. However, those checks are still a part of P21 Enterprise utilized by PMDA and FDA. Most studies prepared by P21 Community Validator have a problem with this business rule. An issue is not identified, not reported, and not explained by Sponsor in Reviewer's Guides.

FALSE-POSITIVES

P21 *Warnings* are not *Checks*, but *Reports* for review. So, they are expected to produce some False-Positive messages. However, there is an additional level of complexity in diagnostics of validation

Warnings. As it was pointed out before, *Data Quality* is defined by *intended use*. Therefore, there are many user-specific issues which may be confusing and considered as False-Positive by some other users: "*If I don't care about it, then it's not an issue!"*

For example, for "*Inconsistent LBSTRESU*" issue a Sponsor provided explanation that "*During Unplanned Visits lab tests assessments were done by local labs utilized different units for same tests. Conversion of results from local labs to standard units was not done because this data was not used in analysis. Sponsor considers this issue as not important and decided to not fix it."* The obvious problem in this example is that FDA and PMDA Reviewers may have different interpretation about importance of lab data during Unplanned visits. Usually, study subjects had unscheduled visits due to safety problems like Adverse Events. Therefore, analysis of data from Unplanned visits may be critical for evaluation of study drug safety.

Programming bug in data validation algorithm is a bug to be fixed. Here are some recommendations and P21 plans for handling False-Positive messages:

- Report bugs to P21 on forum or via email
- Check a list of known bugs on P21 website which will be available soon
- Use *Auto-Update* functionality for patch releases
- P21 is planning to introduce BETA releases with new checks. It will allow to identify incorrect implementation in advance and give extra time for industry to implement new business rules.

There is another special case of False-Positive checks due to invalid original business rules. For example, CDISC ADaM Checks v1.3 still have inconsistencies:

- *#279: AESEVN is not equal to 1, 2, 3, or null*
- *#282: ASEVN is not equal to 1, 2, or 3*
- *#281: There is more than one value of AESEVN for a given value of AESEV (AESEVN & AESEV 1:1 map)*

There may be some discussions if missing Severity is allowed? However, in a case of missing Severity current ADaM checks will produce False-Positive Errors.

Another example is a rule #190: *A variable with a prefix of R2A and a suffix of LO has y fragment appended after R2A that is not a single-digit integer [1-9*]. Expectations are R2A1LO, R2A2LO, … But apparently R2ALO is also legal name.

CONTINUOUS CHECKS' TUNING IS EXPECTED

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.

For example, when FDA asked to populate a missing value for ARMCD instead of SCRNFAIL and NOTASSGN for Screen Failure and Not Assigned subjects, this effected many other checks that used these values to filter out these subjects. Introducing a new implementation of standardized data requires updates to all relevant checks.

SUMMARY

As a summary, we would like to emphasize that

- *Data Conformance* is an important and required process for regulatory submission
- Most studies have deficiencies in data conformance
- Major reason for existing problems in data validation is a lack of knowledge
- P21 has started "*Good Data Validation Practice*" efforts
- Data Quality is defined by *Intended Use*
- All users are different
- P21 Community Validator is a personal desktop tool for SAS programmers to QC their work for FDA/PMDA submissions
- P21 *Errors* are *Checks*, *Warnings* are *Reports* to review

REFERENCES

1. FDA Technical Rejection Criteria for Study Data, 2016, Accessed October 31, 2016.
https://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/FormsSubmissionRequirements/ElectronicSubmissions/UCM523539.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: Sergiy Sirichenko
Company: Pinnacle 21 LLC
Work Phone: 908-781-2342
E-mail: ssirichenko@pinnacle21.net

Name: Max Kanevsky
Company: Pinnacle 21 LLC
Work Phone: 267-331-4431
E-mail: mkanevsky@pinnacle21.net