PharmaSUG 2019 - Paper DS-319

Updates on validation of ADaM data

Sergiy Sirichenko, Pinnacle 21

ABSTRACT

Analysis data is critical for regulatory review process. It helps reviewers understand the details of performed analysis and reproduce results reported by sponsors. Clinical study analysis data is required to be submitted in CDISC ADaM format to both FDA and PMDA. Therefore, validation of analysis data for compliance with CDISC ADaM standard and additional business rules from regulatory agencies is an important step in preparation of study data for regulatory submissions.

In this presentation we will provide an overview of ADaM validation implemented by Pinnacle 21 and used by both FDA and PMDA. It will cover changes related to the new ADaM IG 1.1 standard and the updated version of validation rules from CDISC team. The presentation will also detail additional regulatory business rules, including data and define.xml consistency, validation of ADAM OTHER datasets, SDTM/ADaM traceability, and integrated data.

INTRODUCTION

All users of study data can be classified into two basic groups. People, organizations, and tools that utilize data (Consumers) vs. those that create data (Custodians).

Consumers of study data can be represented by statisticians, reviewers, data warehouses, automated analysis tools, sponsors, etc. Custodians of study data are data managers, programmers, EDC systems, vendors, etc. Often, the same people may represent both groups. For example, within the same project the same statistical programmers are consumers of SDTM data and custodians of ADaM data.

Consumers have specific requirements for particular business needs of study data. Custodians need to meet consumer's expectations.

A data Validator is a communication tool which provides executable business rules to both groups. It helps custodians to prepare high quality data and helps consumers to ensure quality of data.

Data Quality is defined by intended use and as absence of errors which matter. The challenge is that all users are different. Each user is interested only in data issues directly related to him/her. Also, they may have supplementary task-specific requirements. Sometimes there are conflicting requirements from different consumers of study data. For example, FDA reviewers prefer and ask for results utilizing Conventional US units, while PMDA requests SI units.

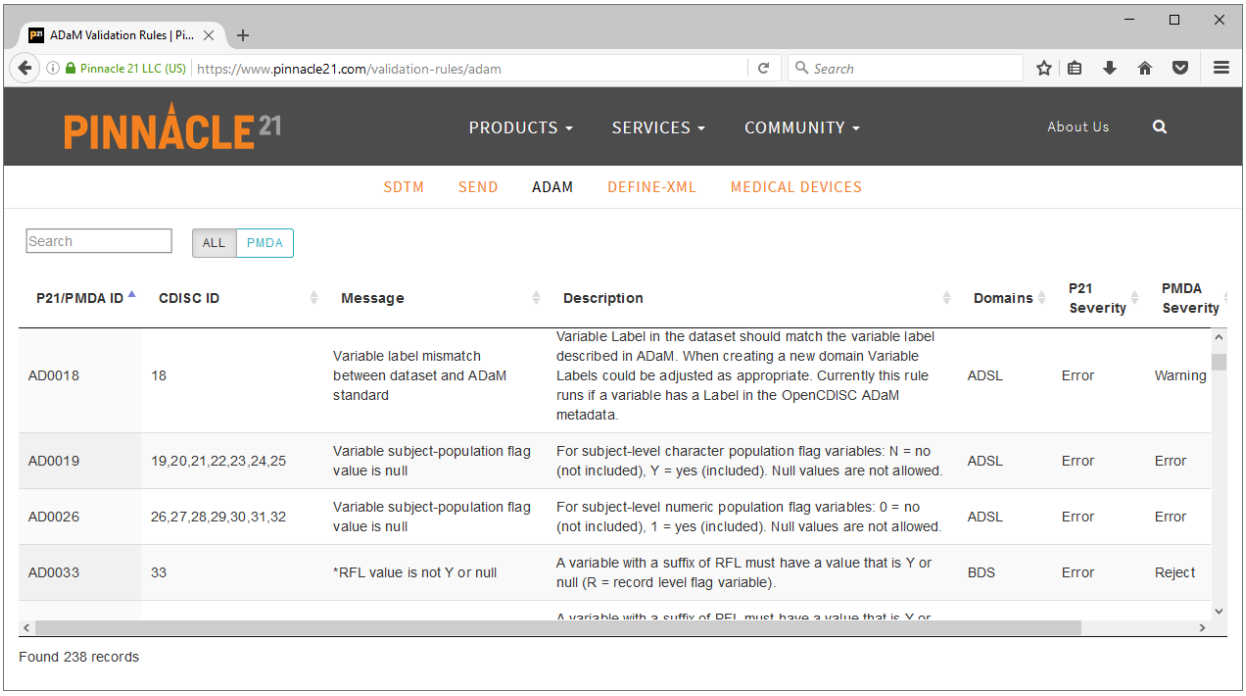
SOURCE OF DATA VALIDATION RULES

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 Conformance Rules" and "SDTMIG v3.2 Conformance Rules". These documents were created by extracting and summarizing all business rules incorporated in text of 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.
    - For example, Pinnacle 21 Validator has checks to ensure correct execution of validation process itself by users.
- Additional requests by users
    - Many companies have custom standards and additional business rules to support internal tools and processes

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.



Display 1. ADAM RULES WITH REFERENCES TO CDISC CHECK IDS ON PINNACLE21.COM

VALIDATION OF TABULATION AND ANALYSIS DATA

There are different stages observed in data standard development and adoption life-cycle. When initial version of the standard is released, its implementation starts very slow. A major driver for adoption of a particular standard by the industry is its requirement by the regulatory agencies. At some point, availability of standardized data allows development of standardized tools. These automated tools require high level of standard compliance and may also introduce addition business rules for study data. Value of standard reporting and analysis increases demand for high quality standardized data. Standardized data allows an extension of validation from standard compliance to data quality with focus on content.

Historically, manual review process for legacy data had lower expectations on data quality. Often, it was quicker for reviewer to fix or work around issues rather than request a fix from the sponsor. Data review

time is limited, while manual evaluation of study data quality is difficult. Standardized data allows early evaluation and enforcement of data quality.

There are three major groups of validation rules:

- Standard compliance
- Data Quality with focus on content
- Tool-specific to support automated processes

Weights of each group vary across standards. They are quite even for SDTM and SEND data which have rigid structure and robust standard control terminology. It makes development of new tools for standardized tabulation data simple. As a result, in addition to standard compliance rules there are many data quality and tool-specific business rules for SDTM/SEND data.

ADaM standard has flexible structure. ADaM looks more like Good Specification Practice compared to SDTM data standard. It introduces a challenge for developing standardized tools because structure of ADaM study data is less predictable. As a result, there are less tool-specific and data quality business rules for standardized analysis data. Today the focus of ADaM validation is still standard compliance rather than data quality and tool support as is for SDTM/SEND.

Therefore, the major source of ADaM validation rules is CDISC.

CDISC ADAM CONFORMANCE RULES

The first version of ADaM Conformance Rules based on ADaM IG 1.0 was published by CDISC team in 2010 and last updated in February 2019.

Check Number	IG Version	ADaM Structure Group	Machine-Testable Failure Criteria	Message Type	Guide	Section	Item	Cited Guidance
1	1.0	ADSL	ADSL dataset does not exist	Error	Model v2.1; ADaM IG v1.0	6; 2.3.1		Model v2.1, Section 6: ADSL and its related metadata are required in a CDISC-based submission of data from a clinical trial even if no other analysis datasets are submitted. ADaM IG v1.0, Section 2.3.1: ADSL and its related metadata are required in a CDISC-based submission of data from a clinical trial even if no other analysis datasets are submitted.
1	1.1	ADSL	ADSL dataset does not exist	Error	Model v2.1; ADaM IG v1.1	6; 2.3.1		Model v2.1, Section 6: ADSL and its related metadata are required in a CDISC-based submission of data from a clinical trial even if no other analysis datasets are submitted. ADaM IG v1.1, Section 2.3.1: ADSL and its related metadata are required in a CDISC-based submission of data from a clinical trial even if no other ADaM datasets are submitted.
2	1.0	ALL:SDTM	A variable is present in ADaM with the same name as a variable present in SDTM but the variables do not have identical labels	Error	Model v2.1; ADaM IG v1.0	4.1.2; 3	4 (General Variable Naming Conventions)	Model v2.1, Section 4.1.2: Any ADaM variable with the same name as an SDTM variable is required to be a copy of the SDTM variable, and its label, attributes, and values cannot be modified. ADaM adheres to the principle of harmonization known as "same name, same meaning, and same values." ADaM IG v1.0, Section 3, Item 4 (General Variable Naming Conventions): Any ADaM variable whose name is the same as an SDTM variable must be a copy of the SDTM variable, and its label, meaning, and values must not be modified. ADaM adheres to a principle of harmonization known as "same name, same meaning, same values."

Display 2. ADaM Conformance Rules v2.0

New format of ADaM rules from CDISC team includes:

- *Check Number* – for traceability across versions of rules
- *IG Version* – reference to version of standard
- *ADaM Structure Group* – reference to applied datasets. For example, ADSL or ALL:SDTM
- *Machine-Testable Failure Criteria* – validation message
- *Message Type* – Error, Warning or Note based on capability of business rule to avoid false-positive

messages

- *Guide* – reference to source document of rule
- *Section* – reference to section within a source guide
- *Item* – additional granularity of source
- *Cited* Guidance – text from source document

New conformance checks for ADaM IG 1.1 include 271 business rules. Most of them are Errors. There are also 18 Warnings and 11 Notes messages.

Conformance rules for ADaM IG 1.0 were updated. For example, CDISC ADaM Checks v1.3 had 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 an argument for allowing missing Severity, however, in such a case the old ADaM checks would produce False-Positive Errors. Rule #282 is fixed in a new version by adding null value.

Another example is 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 were R2A1LO, R2A2LO, … But, apparently R2ALO is also a legal name. This rule is removed in the new version.

Most previous rules for Analysis Adverse Events and Time-To-Event were removed.

VALIDATION OF ADAM DATA FROM PINNACLE 21

Pinnacle 21 Validator was created to help the industry with implementation of CDISC standards and preparation of study data for regulatory submissions. The Pinnacle 21 tool is utilized by FDA and PMDA for validation of submissions and is available for all users in the industry.

Community version of the Validator is based on officially published rules from CDISC, FDA and PMDA. P21 Community 2.2.0 covers ADaM IG 1.0 utilizing the latest rules from CDISC ADaM published in 2015. It also has additional rules like consistency checks between define.xml and actual study data, missing checks from ADaM team and the agency-specific rules. PMDA published a list of their official rules with reference to Pinnacle 21 implementation and versions of software. FDA did not publish their official list of rules. Both agencies are using Pinnacle 21 Enterprise (a.k.a FDA DataFit) for validation of analysis data.

Validation of new CDISC ADaM conformance rules v2.0 has been supported in Pinnacle 21 Enterprise as a draft implementation a year before the official release of rules from CDISC in February 2019.

New Pinnacle 21 Community 3.0 includes validation ADaM data based on CDISC conformance rules 2.0. There are also several major changes in general approach for validation of ADaM data.

GENERAL UPDATES ON ADAM VALIDATION

- Rules for correct execution of validation are introduced
- Partial support for custom ADaM variables is introduced

UPDATES ON ADAM IG 1.0

- There is minor tuning according to a updated ADaM conformance rules from CDISC
- Additional validation for ADAM OTHER datasets is introduced

UPDATES ON ADAM IG 1.1

- The first version of ADaM IG 1.1 rule from CDISC team is implemented
- ADaM rules from CDISC are expended to other applicable datasets

VALIDATION PROCESS RULES

When validating ADaM datasets, some SDTM domains should be available for execution of data traceability checks. Unfortunately, despite of educational efforts of Pinnacle 21 team, most users are still not aware of these requirements and do not include AE, DM and EX domains in validation of ADaM datasets. As a result, data traceability rules are not executed. Since of them are *Errors*, correct validation of these rules is extremely important for PMDA submissions, where independent validation and results reconciliation are performed by the Agency. Incorrect execution of ADaM validation may result in delay of submission process in Japan.

Rule ID	Message
AD0053	The combination of STUDYID and USUBJID value does not exist in the SDTM DM domain
AD0061	SDTM.EX is present but neither ADSL TRTSDT nor TRTSDTM are present
AD0204	For the same USUBJID, the ADSL.AGE != DM.AGE
AD0205	For the same USUBJID, the ADSL.AGEU != DM.AGEU
AD0206	For the same USUBJID, the ADSL.SEX != DM.SEX
AD0207	For the same USUBJID, the ADSL.RACE != DM.RACE
AD0208	For the same USUBJID, the ADSL.SUBJID != DM.SUBJID
AD0209	For the same USUBJID, the ADSL.SITEID != DM.SITEID
AD0210	For the same USUBJID, the ADSL.ARM != DM.ARM
AD0253	Record key from SDTM AE is not traceable to ADaM ADAE (not enough ADAE recs)
AD0258	Record key from ADaM ADAE is not traceable to SDTM.AE (extra ADAE recs)
AD0367	For the same USUBJID, the ADSL.ACTARM != DM.ACTARM

Table 1. Examples of ADaM data traceability rules in Community 3.0

To avoid such problems, Community 3.0 and Enterprise 4.0.1 introduce new rules which require a presence of additional SDTM datasets AE, DM and EX for validation of ADaM data.

Rule ID	Message
AD1024	Traceability rules not executed due to missing DM dataset
AD1025	Traceability rules not executed due to missing AE dataset
AD1026	Traceability rules not executed due to missing EX dataset

Table 2. List of ADaM data validation prcoess rules in Community 3.0

ADAM OTHER

P21 Community version 2.2.0 and older versions of Enterprise identify dataset as CDISC Basic Data Structure (BDS) if any of PARAM, AVAL, ADT, ASTDT, AENDT variables are present. However, ASTDT and AENDT variables are not good candidates for this task. Therefore, CDISC Occurrence Data Structure

(OCCDS) and ADaM Other (ADAM OTHER) was incorrectly validated as BDS datasets producing notorious false-positive messages like "*Required variable PARAMCD is not present*".

Fixing this bug resulted in the validation being skipped for OCCDS and ADAM OTHER.

Surprisingly, Pinnacle 21 clients started requesting that this bug fix is reversed to support its "*off-label*" functionality to perform at least some validation for OCCDS and ADAM OTHER datasets.

Therefore, a new special validation was introduced for ADAM OTHER:

- ADAMOTHER is dataset named "AD**", but not ADSL, ADAE, BDS or OCCDS
- ADSL "Core" variables were assigned to ADAM OTHER as Permissible
- Some global rules were applied
    - For example, AD0041: *"*DT does not have the ADaM required SAS Date format*", AD0092: *"Inconsistent value for TRTPN*", etc.

ADAM IMPLICIT RULES

Conformance rules from CDISC are limited to text in standard documents like ADaM Model and Implementation Guide. Sometimes, there are implicit assumptions that some variables and related business rules may be applicable in other datasets. However, this is out of scope of CDISC project.

For example, ADaM IG 1.1 documentation has 41 Treatment variables defined for ADSL datasets. There are also 13 Treatment variables defined for BDS datasets. While ADaM Structure for Occurrence Data document has only 3 Treatment variables specific for OCCDS:

- DOSEON - Treatment Dose at Record Start
- DOSCUMA - Cumulative Actual Treatment Dose
- DOSEU - Treatment Dose Units

"*The treatment variable used for analysis must be included. Typically this would be TRTP, TRTA, TRTxxP, or TRTxxA. See the ADaM Implementation Guide version 1.1 [2] for more details on these variables. Additional dosing variables may also be included*" (ADaM OCCDS v1.0)

Assumption is that ADSL variables can be utilized across all other analysis datasets as Core variables. As well as, BDS Treatment variables may be and actually are used in OCCDS datasets.

Limitation of official CDISC conformance rules to only 3 specific Treatment variables and ignoring other 54 potential Treatment variables are artificial.

As the first step all ADSL variables and related conformance rules are extended to BDS, OCCDS and ADAM OTHER datasets.

ADAM RULES REMOVED BY CDISC

Official CDISC ADaM rules are aimed to be "programmable" and avoid false-positive messages. As a result, many previous rules were removed in the new version of conformance checks.

For example, there is no formal rule to identify Analysis Adverse Events (ADAE) dataset. Therefore, all ADAE rules were removed.

Pinnacle 21 Validator assumes 'adae.xpt' file name for ADAE dataset. In practice, it is true for 99% studies. Validation of ADAE includes traceability checks and is very important for reviewers. Therefore,

we expect that PMDA will keep them as their official rules for ADaM. FDA also emphasizes value of ADAE validation.

Now CDISC team is considering an option to bring these rules back in the future releases when a new version Define-XML standard will include a new SubClass Element. It allows to formally specify in define.xml file ADAE as “Adverse Events Analysis Dataset” sub-class of Occurrence Data Structure and apply relevant conformance rules.

The similar issue is about Time-To-Event (TTE) data. All previous conformance rules for TTE were removed, because there is no formal rule to identify TTE datasets.

Pinnacle 21 identifies TTE datasets by presence of CNSR (Censor) variable. While it is not 100% accurate algorithm, a value of validating TTE analysis data is more important than potential risk of false-positive messages.

REDUCING NOISE

False-positive validation messages (FPVM) may be annoying and disruptive. There are three major sources of FPVM:

- Bugs - Reporting bugs to software developers will speed identification and resolution
- Warning messages are expected to have FPVM - Tuning of algorithms and better reporting will help
- User misinterpretation of real issues as FPVM - Education and knowledge sharing

Pinnacle 21 team has started utilizing the industry metrics to identify and resolve most common and notorious false-positive validation messages

Rule ID	Message	Affected studies	Average Issue Rate
CT2002	Variable value not found in extensible codelist	70.6%	13.8%
AD0047	Required variable is not present	56.5%	32.5%
AD0018	Variable label mismatch between dataset and ADaM standard	53.9%	12.3%
AD0198	Neither AVAL nor AVALC are present in dataset	48.9%	100.0%
AD9999	Dataset not validated	40.2%	100.0%
DD0084	Referenced File is missing	32.7%	100.0%
SD1231	Variable value is longer than defined max length %Variable.@Clause.Length% when value-level condition occurs	25.0%	59.1%
AD0124	Inconsistent value for PARCATy within a unique PARAMCD	23.9%	21.3%
DD0060	Define.xml/CDISC variable Label mismatch	23.9%	100.0%
AD0154	Multiple baseline records exist for a unique USUBJID, PARAMCD, BASETYPE	22.6%	31.9%
SD0037	Value for variable not found in user-defined codelist	22.6%	38.7%
AD0149	Inconsistent value for AVALC	21.7%	2.7%
DD0085	Missing Define XSL	21.4%	100.0%
AD0196	Required Variable value is null	17.2%	11.5%
AD1012	Secondary variable is present but its primary variable is not present	16.8%	41.9%
SD1229	Variable value is null when value-level condition occurs	15.8%	21.9%
SD1230	Variable datatype is not %Variable.@Clause.DataType% when value-level condition occurs	15.4%	61.4%
AD0225	Calculation issue: PCHG != (AVAL - BASE)/BASE * 100	12.6%	20.4%
AD0141	Inconsistent value for PARAM within a unique PARAMCD	11.9%	5.4%
CT2001	Variable value not found in non-extensible codelist	11.9%	50.9%

Table 3. Metrics on ADaM validation – Top 20 issues

The most commonly reported issue CT2002 is due to presence of extra terms in extensible CDISC ADaM Control Terminology. This is an example of Warning message. In most cases, extension of standard CT is valid. However, there are still some cases when new terms were incorrectly added as synonyms of existing standard terms. For example, in 'Derivation Type' CT a term 'SCREENING' may be a synonym of standard term 'SOCF'. 'MAX' is invalid new extended term because it was used instead of standard term 'MAXIMUM'.

AD0047 '*Required variable is not present*' is useful check. However, it is also quite notorious for producing false-positive messages. A major reason was a bug for incorrect identification of OCCDS datasets as BDS.

Another common case for firing AD0047 rule was its request for presence of TRTP variable in BDS dataset according ADaM IG 1.0 specifications, when some sponsors believe that TPTP variable is not needed when TPT01P and TPT02P are already utilized.

AD0018 '*Variable label mismatch between dataset and ADaM standard*' rule represents #3 top issue in ADaM data. In some cases, this very useful check may produce false-positive messages due to incorrect

interpretation of custom variables as a standard variable defined by wildcard characters. For example, ANLzzFL (Analysis Flag zz) is a standard ADaM variable in BDS datasets. When checking for correct implementation of standard variable label, AD0018 algorithm replaced 'zz' fragment in standard label metadata by representation of 'zz' characters in variable name. Unfortunately, the old algorithm did not see a difference between digits only per ADaM standard and any characters. Therefore, if a programmer introduced a new variable ANLTTEFL, then AD0018 expected to see 'Analysis Flag TTE' label. However, this case is out of scope of ADaM specifications and should not be validated. In new versions of P21 Validator, AD0018 algorithm is adjusted to validation only variables where wildcard characters are numbers per ADaM specs.

However, there are still some problems with correct implementation of ADaM standard.

For example, "*Inconsistent value of PARCATy within a unique PARAMCD" issue is present in 24% studies. "Inconsistent value for PARAM within a unique PARAMCD*" is in 12% studies.

7 out of 20 most common issues in analysis data are related to problems with correct implementation of define.xml file.

IMPROVING DIAGNOSTICS

Another area for potential enhancement of data validation is improving diagnostics of reported issues.

For example, some Inconsistency checks report only a new inconsistent value without a reference to original base value.

AD0018 "*Variable label mismatch between dataset and ADM standard*" report does not clarify expected standard label.

CONCLUSION

Validation of analysis data for compliance with CDISC ADaM standard and additional business rules from regulatory agencies is an important step in preparation of study data for regulatory submissions.

New version of Pinnacle 21 Validator includes validation for both ADaM IG 1.0 and ADaM IG 1.1 and extends CDISC conformance rules for needs of regulatory submissions.

CONTACT INFORMATION

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

Sergiy Sirichenko
Pinnacle 21 LLC
+1.570.817.6137
sergiy@pinnacle21.com