Paper SI14

Confusing Data Validation Rules Explained, Part 2

Michael Beers, Pinnacle 21, Blue Bell, PA, USA

ABSTRACT

The 2018 PhUSE US Connect paper 'Confusing Data Validation Rules Explained' by Pinnacle 21 discussed some of the more confusing data validation rules, including the possible sources of confusion, and clarifications on the intent and purpose of the validation rule. This paper will discuss a new set of confusing validation rules, and explain what the resulting validation issues would mean.

INTRODUCTION

Validation of study data is a critical step in preparing for a submission. The FDA's Study Data Technical Conformance Guide states that sponsors should either correct any discrepancies between study data and the standard or the business rules or explain meaningful discrepancies in the Reviewer Guide (i.e., nSDRG, cSDRG or ADRG) [1]. A clear understanding of the validation rules is necessary to accomplish this goal of correcting discrepancies in study data or providing a useful explanation in the Reviewers Guide.

Validation rules are developed by regulatory agencies and standards development organizations (SDO). These validation rules can be confusing, as study data, and the issues being checked for are complicated. Proper dispositioning of validation rules requires good working knowledge of clinical trial data, data standards, and regulatory requirements. Moreover, data standards and regulatory requirements change over time.

Another consideration is that sometimes the guidance between SDOs and regulatory agencies may be contradictory. Examples of this are how to map certain data (Planned Arm Code (ARMCD) for screen failures for example), and what is required to be submitted (variables requested by regulatory agencies vs core status set by CDISC…the EPOCH, study day (--DY) variables, etc.).

Study data validation confusion can be around the validation process (how to, when to, etc.), or the results of the validation. Clarifying best practices about the validation process (how to, when to, etc.) is not addressed in this paper.

Confusion associated with the results of validation can be due to:

- an unclear or complicated validation message,
- misconception of a certain concept, or
- misunderstanding of the purpose of the validation rule.

This paper will select a few of the confusing validation rules and try to clarify how to interpret the results of these validation rules.

Study data validation rules can generally be grouped under these types of rules: Data Quality, Controlled Terminology, Metadata, Regulatory Conformance, and Standards Compliance. Examples of confusing validation rules for each of these types will be discussed.

DATA QUALITY RULES

Data Quality validation rules, for the most part, identify issues with the collection of the data, deficiencies in the data, or issues that otherwise could potentially affect reviewability. Here are examples of data quality validation issues which may be confusing to sponsors:

DUPLICATE RECORDS IN – DOMAIN (SD1201)

Duplicate records, or perhaps more accurately stated as multiple observations/events collected at the exact same time point, can cause unanticipated results during analysis and can complicate analysis by causing issues with FDA review tools. This validation rule looks to identify these potential situations so that sponsors can correct the issue or provide an explanation as to how a reviewer might work around the issue. The difference between this rule SD1201 and rule SD1117 (Duplicate records), is that SD1201 is specific to Events domains, while SD1117 is specific to Findings domains.

Here are some of the common scenarios seen for this validation rule:

- Actual duplicate information, except for the sponsor assigned sequence variable (--SEQ).
- Records that are only differentiated by a sponsor-defined variable (--SPID).
- Records that are differentiated by the collection date (--DTC), instead of the start date (--STDTC) of the event.

- Records in the Disposition domain for multiple informed consent obtained.
- Records in the Medical History or Disposition domain for rescreened subjects.
- Multiple events, typically Medical History, that started in the same year/month, but exact date is unknown.

Sponsors tend to interpret this rule in a very strict meaning of the word 'duplicate'. Therefore, we tend to see not very useful explanations for these validation issues. A typical explanation from a sponsor for this validation rule looks something like:

"These are not true duplicates. Records are unique based on primary keys listed in the define.xml"

It is understood that there may be other variables that contribute to the keys of the sponsor's dataset, however many times these are sponsor-defined variables, such as Sponsor-Defined Identifier (--SPID). Stating that a variable such as --SPID is needed to uniquely identify a record is typically not useful information, unless perhaps the define.xml clearly and completely describes what this sponsor-defined variable contains.

A good disposition of this validation issue, if unable to be fixed, would be to provide a detailed explanation of why there appears to be multiple events collected for subjects at the same time point, which variable(s) would be needed to differentiate these records, and exactly what the variable(s) contains.

NO RECORDS FOR 'SCRNFAIL' SUBJECT ARE FOUND IN IE DOMAIN (SD1032)

This validation rule will fire if the subject's ARMCD equals 'SCRNFAIL', but no IE records exist for the subject.

Common scenarios where this rule fires are:

- Records for inclusion/exclusion criteria for screen failure subjects are just not submitted for some unknown, and likely invalid, reason.
- Subjects that met inclusion/exclusion criteria, but failed randomization criteria. Many studies have separate randomization inclusion criteria, and sponsors tend to not map this to the IE domain, instead mapping it to SUPPDS.
- Reasons for screen failure have been mapped to the Disposition domain instead of the Inclusion/Exclusion Not Met domain. Often this occurs when the CRF is set up with a field to enter the criteria numbers that the patient failed, and the sponsor did not set up the mapping to match the numbers to the actual IE criteria and map to the IE domain (it is easier to just map the exact value collected to the DS domain).
- Subjects that met inclusion/exclusion criteria, but withdrew prior to randomization.
- Subjects are identified as screen failures, when that is not the most accurate ARM. For example, if the reason is that the cohort is full, or the sponsor closed enrollment, it is better to use Not Assigned.

Confusion also results due to the following causes:

- Regulatory guidance is less than clear on whether you must submit screen failure data or not. For example, the Technical Conformance Guide states that "Screen failures, *when provided*, should be included as a record in DM…" [1] Sponsors many times are not sure if they should submit screen failure data or not, and if so, which data is required.
- Changing CDISC guidance makes it less clear and inconsistent how screen failure subjects are identified. In SDTMIG v3.3, guidance was changed to have ARM/ACTARM mapped as null for screen failures (to match the guidance in the FDA's Technical Conformance Guide) with the addition of the new variable ARMNRS (Reason Arm and/or Actual Arm is Null) [2], where prior to SDTMIG v3.3 CDISC guidance was to map ARM/ACTARM as 'Screen Failure'. This results in different mapping of ARM/ACTARM, and inconsistent identification of screen failures, across the industry.
- CDISC guidance does not cover these situations listed above, such as how to map randomization inclusion criteria. If these common scenarios were accounted for in CDISC guidance, then there would be less room for interpretation of mapping.

While it may be somewhat common for some screen failure subjects to not have records in the Inclusion/Exclusion Criteria Not Met domain, it may become problematic when many subjects are missing this information, as it could affect the ability to determine possible bias in patient enrollment. It is recommended to verify that this failed criteria data is collected for screen failure subjects, and mapped appropriately to the Inclusion/Exclusion Criteria Not Met domain.

CONTROLLED TERMINOLOGY RULES

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

VARIABLE VALUE NOT FOUND IN EXTENSIBLE CODELIST WHEN VALUE-LEVEL CONDITION OCCURS (CT2005)

In the Implementation Guides provided by CDISC, many variables have a codelist assigned (at the variable-level). Occasionally however, a variable in a domain may collect different types of data, and those different types of data should use separate codelists. This results in the need for the creation of value-level metadata for that variable, and codelists being assigned depending on a specific where clause. An example of this is the Disposition domain. The

DSDECOD variable is subject to different codelists depending on whether DSCAT = DISPOSITION EVENT or PROTOCOL MILESTONE.

This validation rule will fire if a value in the dataset, for a variable with a CDISC-defined codelist when a value-level condition occurs, does not exactly match a value in the CDISC extensible codelist. While this rule may fire for domains such as Trial Summary, etc., the most common case of this rule firing is for the DSDECOD variable in the Disposition domain, when DSCAT = DISPOSITION EVENT. When DSCAT = DISPOSITION EVENT, the DSDECOD values should adhere to the NCOMPLT (Completion/Reason for Non-Completion) extensible codelist, and the rule will fire if the values in the data are not listed in this CDISC codelist.

Typically, the scenarios that cause this validation rule to fire are:

- A sponsor uses a value that has no corresponding match in the CDISC codelist. This is often due to the grey area between Protocol Milestones and Disposition Events. For example, a sponsor may have a value collected on the CRF of 'TRANSITIONED TO OPEN LABEL" or "CONTINUING", and the distinction between Protocol Milestone and Disposition Event is unclear.
- A sponsor uses a value that has a match in the CDISIC codelist but uses different casing for the value.
- A sponsor uses a synonym for a value in the CDISC codelist. This is often done when the sponsor wants to map the exact value collected on the CRF to the DSDECOD variable, instead of the correct approach of mapping it to a corresponding CDISC Submission value from the codelist.

Extending an extensible codelist when there is no corresponding value to use is an acceptable approach, however the other scenarios are not. Sponsors tend to generically dismiss this validation issue with an explanation such as: "Codelist is extensible."

A proper dispositioning of this validation issue would be to correct the implementation to use the valid controlled terminology value where possible. If valid extensions to the codelist were used, explain the issue in the reviewer's guide and list all actually valid extended values in the explanation. Also, if your study collects valid extensions to a CDISC codelist, it is recommended to submit a request to have them added to the CDISC codelist to avoid this situation in the future.

VALUE FOR --DECOD NOT FOUND IN WHODRUG DICTIONARY (SD1344)
VALUE FOR --CLAS NOT FOUND IN WHODRUG DICTIONARY (SD1345)
VALUE FOR --CLASCD NOT FOUND IN WHODRUG DICTIONARY (SD1346)

These are new validation rules, recently introduced to validate concomitant medications coding against the WHODrug Dictionary. The rules will fire if the value for CMDECOD/CMCLAS/CMCLASCD do not exist in the version of WHODrug that is selected in the configuration of the validation.

There are a few sources of confusion related to this validation rule:

- Not all WHODrug version formats are supported for validation. The FDA's Data Standards Catalog lists the B3 format as the only format supported [4], so that is the only format available to validate data against. If a sponsor uses an older format not listed as supported in the Data Standards Catalog, or uses an in-house proprietary medications dictionary, then validation will not be supported.
- Guidance for mapping of dictionary values greater than 200 characters is not provided in CDISC guidance. There is guidance on this mapping provided by Uppsala Monitoring Centre however, but this requires the sponsor to be aware of the existence of this separate document, which does not appear to be referenced in any CDISC Implementation Guide. This WHODrug document is titled "How to use WHODrug for compliance with CM domain in the CDISC SDTM standard" [3].
- The Concomitant Medications domain many times combines different types of collected medications data into the same domain. This often results in cases where the CM domain has some medication records that are coded to the WHODrug dictionary, and some records that are not coded, if it is not necessary for analysis.

It is recommended that sponsors use the WHODrug version format listed as supported in the FDA's Data Standards Catalog to code their concomitant medications data, refer to the document "How to use WHODrug for compliance with CM domain in the CDISC SDTM standard" [3] (since this provides specific mapping guidance which is lacking from CDISC Implementation Guides), and configure their validation to use the corresponding WHODrug dictionary to ensure any validation issues are caught prior to submission.

METADATA RULES

Metadata validation rules look for issues with the define.xml. These can be issues with the XML code, incorrect implementation of define.xml, or inconsistencies between the define.xml and the study datasets.

VARIABLE VALUE NOT FOUND IN USER-DEFINED CODELIST WHEN VALUE-LEVEL CONDITION OCCURS (SD1228)

Sometimes a variable may use a codelist for some values, but not all. An example of this is the laboratory test standard character result (LBSTRESC) variable. Some tests can have qualitative results (POSITIVE/NEGATIVE for example), while other tests will have quantitative results for which a codelist is not appropriate. The solution for this

scenario is to not assign a codelist at the variable-level in the define.xml, but instead use value-level metadata to assign a codelist to only the applicable tests.

In the define.xml, if a codelist is listed for a variable in the value-level metadata section, this validation rule SD1228 will check values in the data (where the value-level condition is met) against this specified codelist. If there are values in the data (for this value-level condition) that are not listed in the codelist in the define.xml, the rule will fire. This validation rule will only fire when the define.xml is included when the datasets are validated. Including the define.xml is an important step in the validation; however, it is common for sponsors to exclude the define.xml during validation, which results in the sponsor not correcting this issue, and not explaining why the issue remains.

Typically, there are four scenarios that cause this validation rule to fire:

- A value in the codelist in the define.xml is misspelled and does not exactly match the corresponding value in the dataset.
- The values in the codelist in the define.xml are in a different case than the values in the dataset.
- A value in the dataset is just missing from the codelist in the define.xml.
- The wrong codelist was accidentally assigned for a variable in the define.xml.

Many times, sponsors explain this issue in the reviewers guide as 'The codelist is extensible'. This suggests that the sponsor does not understand what the validation rule is checking, as this rule has nothing to do with the concept of extensibility.

Sponsors must ensure that there define.xml is complete, and that all codelists listed at the variable-level and also value-level, are complete and reflect the values in the data accurately.

VARIABLE VALUE IS LONGER THAN DEFINED MAX LENGTH WHEN VALUE-LEVEL CONDITION OCCURS (SD1231)

Sponsors should create value-level metadata in their define.xml when specifying metadata at the variable-level is not sufficient for describing the data. Value-level metadata should always be created for supplemental qualifiers, but often for other cases as well, such as for results for findings domains, parameters for the Trial Summary domain, etc. When creating value-level metadata in a define.xml, the metadata for each value-level condition must accurately reflect the actual data. This validation rule checks one of those pieces of metadata…the length attribute. The validation rule will fire when the values in the dataset, for that variable, when that value-level condition is met, are greater than the length specified in the define.xml.

Common scenarios where this issue fires are:

- The sponsor creates a define.xml for an ongoing study, then later more data is collected with values that are longer than the length originally specified in the define.xml, but the sponsor does not go back and update the length in the define.xml.
- The sponsor is completely unaware that this validation issue exists, because when validating the datasets, the define.xml was incorrectly excluded from the validation.
- The sponsor thinks the finding is a false-positive, due to hidden characters such as leading spaces. When the sponsor investigates the issue by viewing the values in the data, many times they will not see that leading spaces are included, causing the validation issue.

When creating value-level metadata in a define.xml, care must be taken to ensure that the metadata accurately reflects the actual data. As a define.xml is a document created to reflect the data, there should be no reason to have an inaccurate define.xml. Therefore, the recommendation is to always fix all define.xml related validation issues, so that a clean error-free define.xml is provided as part of the submission. This can be done by properly validating the define.xml with the data, in order to run the data vs. define.xml cross-check rules, in addition to validating the define.xml by itself to identify any XML related issues.

REGULATORY CONFORMANCE RULES

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.

MISSING EPOCH VALUE, WHEN A START OR OBSERVATION DATE IS PROVIDED (SD1339)

In Confusing Data Validation Rules (Part 1), validation rules SD1077/SD1140 (FDA/PMDA Expected Variable Not Found) were discussed. Since the time of that paper, a new validation rule has been introduced to go beyond just checking for the presence of the EPOCH variable, and now also check for the population of the EPOCH variable. This validation rule SD1339 will fire if a complete (not partial) start date or observation data is provided, yet the EPOCH variable is null.

This new validation rule, as well as SD1077, is based on the FDA Technical Conformance Guide, which states that the variable EPOCH should be included for clinical subject-level observation (e.g., adverse events, laboratory, concomitant medications, exposure, and vital signs). This will allow the reviewer to easily determine during which phase of the trial the observation occurred (e.g., screening, on-therapy, follow-up), as well as the actual intervention the subject experienced during that phase [1]. Providing and populating the EPOCH variable for each domain helps in

analysis by allowing easy determination of which phase of a trial an event or observation occurred in, without having to compare dates to make these determinations.

Previously, just checking for the presence of the EPOCH variable (which SD1077 did), only covered one portion of this statement from the TCG. With the addition of SD1339, we are now more fully checking the guidance in the TCG.

The confusion associated with these EPOCH-related validation rules is due to seemingly conflicting information between regulatory agency 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 and populate the variable.

A common example of a sponsor explanation for this validation rule is: "EPOCH is a permissible variable as per SDTM IG 3.1.3, hence not included."

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, it should be obvious that just 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.

Note that for historical data, for which an EPOCH may not be appropriate, it is recommended to just explain this situation in the reviewer's guide.

MISSING EXTTIND TRIAL SUMMARY PARAMETER (SD2273)
MISSING NCOHORT TRIAL SUMMARY PARAMETER (SD2274)
MISSING OBJSEC TRIAL SUMMARY PARAMETER (SD2275)
MISSING PDPSTIND TRIAL SUMMARY PARAMETER (SD2276)
MISSING PDSTIND TRIAL SUMMARY PARAMETER (SD2277)
MISSING PIPIND TRIAL SUMMARY PARAMETER (SD2278)
MISSING RDIND TRIAL SUMMARY PARAMETER (SD2279)
MISSING SDTIGVER TRIAL SUMMARY PARAMETER (SD2280)
MISSING SDTMVER TRIAL SUMMARY PARAMETER (SD2281)
MISSING THERAREA TRIAL SUMMARY PARAMETER (SD2282)

Pinnacle 21 users will soon notice these new validation issues in each of their studies. These new validation rules were introduced to check for missing Trial Summary parameters that are listed in the FDA's Technical Conformance Guidance as 'Desired' [1].

The FDA's Technical Conformance Guide is increasingly used to guide the industry on which Trial Summary parameters should be included. This had previously been done in the appendices of the SDTM Implementation Guides, but now is controlled by the regulatory agencies. In the FDA's Technical Conformance Guide, appendices are provided which list Trial Summary parameters, as well as an 'FDA Notes' column which specifies some implementation guidance, and a column named 'FDA Desired', for which each parameter is noted as 'Y' or 'Conditional'. These new validation rules were created to check for the existence of these parameters listed as FDA Desired = 'Y' in the sponsors Trial Summary domain.

The recommendation is to provide all Trial Summary parameters that are listed as FDA Desired = 'Y', even if they do not apply to the study. In the cases where the parameters do not apply to the study, TSVAL should be left as null, and TSVALNF (Null Flavor) should be used to specify the reason that TSVAL is null.

STANDARDS COMPLIANCE RULES

Standards Compliance validation rules look for issues with how study data was mapped to CDISC Standards. These can be implementation inconsistencies, discrepancies, deficiencies per CDISC guidance, or traceability issues between SDTM and ADaM data.

TRACEABILITY RULES NOT EXECUTED DUE TO MISSING DM DATASET (AD1024)
TRACEABILITY RULES NOT EXECUTED DUE TO MISSING AE DATASET (AD1025)
TRACEABILITY RULES NOT EXECUTED DUE TO MISSING EX DATASET (AD1026)

One of the most important concepts in regards to clinical trial data is traceability. These three new validation rules (AD1024, AD1025, and AD1026) are specific to the validation of ADaM data, and have been created to ensure that a sponsor is including the correct SDTM datasets in their ADaM validation so that the following existing validation rules will be run, to check traceability between SDTM and ADaM datasets:

- The combination of STUDYID and USUBJID value does not exist in the SDTM DM domain (AD0053)
- SDTM.EX is present but neither ADSL TRTSDT nor TRTSDTM are present (AD0061)
- SDTM.EX is present but neither ADSL TRTEDT nor TRTEDTM are present (AD0061A)
- For the same USUBJID, the ADSL.AGE != DM.AGE (AD0204)
- For the same USUBJID, the ADSL.AGEU != DM.AGEU (AD0205)
- For the same USUBJID, the ADSL.SEX != DM.SEX (AD0206)
- For the same USUBJID, the ADSL.RACE != DM.RACE (AD0207)`

- For the same USUBJID, the ADSL.SUBJID != DM.SUBJID (AD0208)
- For the same USUBJID, the ADSL.SITEID != DM.SITEID (AD0209)
- For the same USUBJID, the ADSL.ARM != DM.ARM (AD0210)
- Record key from SDTM AE is not traceable to ADaM ADAE (not enough ADAE recs) (AD0253)
- Record key from ADaM ADAE is not traceable to SDTM.AE (extra ADAE recs) (AD0258)
- For the same USUBJID, the ADSL.ACTARM != DM.ACTARM (AD0367)

These validation rules will not run, however, unless the sponsor knows to include three SDTM datasets in their ADaM validation: DM, AE, and EX. This has been a problem in the past, that sponsors were not including these SDTM domains in their ADaM validation, which led to traceability issues not being seen by the sponsor, and not being explained in the reviewer's guide. However, these SDTM domains are always included as part of the automated validation that occurs at the regulatory agencies, and so any traceability issues are seen by the agencies. These three validation rules (AD1024, AD1025, AD1026) were introduced to alert the user that these SDTM domains need to be included in the ADaM validation.

Note that while these three SDTM domains must be included in the ADaM validation, they should not be included in the ADaM datasets folder in the eCTD for the actual submission to the regulatory agencies.

CONCLUSION

A clear understanding of validation rules, and the issues that are identified by these rules, is critical to providing high quality standardized study data. Confusion regarding the validation rules can lead to important issues not being corrected, regulatory agencies not receiving the information they need, and data issues not being explained sufficiently. Therefore, it is important for all stakeholders to develop a clear understanding of the validation issues that may be present in their data. This paper intends to provide clarity for some of the more confusing study data validation rules.

REFERENCES

[1] FDA. "Study Data Technical Conformance Guide" October, 2019. https://www.fda.gov/media/131872/download

[2] Study Data Tabulation Model Implementation Guide: Human Clinical Trials. Clinical Data Interchange Standards Consortium (CDISC) Submission Data Standards (SDS) Team. Version 3.3. November 2018

[3] Uppsala Monitoring Centre. "How to use WHODrug for compliance with CM domain in the CDISC SDTM standard". 2017.
https://www.who-umc.org/media/2940/how-to-use-whodrug-for-compliance-with-cm-domain-in-the-cdisc-sdtm-standard-march-2017.pdf

[4] FDA. Data Standards Catalog v6.1 (September 9, 2019). https://www.fda.gov/media/85137/download

CONTACT INFORMATION

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

Michael Beers
Pinnacle 21 LLC
mbeers@pinnacle21.com