PharmaSUG 2018 - Paper SS-18

Confusing Data Validation Rules Explained

Michael Beers, Pinnacle 21

ABSTRACT

An important step in the submission of tabulation and analysis data to regulatory agencies is the validation of the data according to rules developed by regulatory agencies and standards development organizations (SDO).

Pinnacle 21 tool is commonly used throughout the industry to produce these validation results.

Errors and warnings in study data reported by the Validator must be corrected. If the issues won't be corrected, then they must be explained in the Reviewer's Guide.

This dispositioning of validation findings can be a challenge, as some business rules and validation messages are more confusing than others.

Unclear understanding of validation results can lead to important data issues not being corrected, or incorrect explanations in the Reviewer's Guide.

This presentation will walk through some of the more confusing of the validation rules and explain what the findings 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 SDTM 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 (SD1117)

Duplicate records 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.

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

- Actual duplicate information, except for the sponsor assigned sequence variable (--SEQ).
- Records with the same timing information for a test, but the results are different.
- Records with the same timing information for a test, but one of the records is NOT DONE.
- Records that are only differentiated by a sponsor-defined variable.

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:

"The keys defined by the check are not sufficient to identify a unique record for patient."

This is an actual explanation from a sponsor, and the actual keys listed in define.xml for this domain (Questionnaires) were: STUDYID, USUBJID, QSSEQ. The –SEQ variable contains a sequence number assigned by the sponsor and must be unique within a subject per CDISC guidance [2]. Using this variable as a key provides no useful information in regards to the structure of a domain.

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 observations 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.

ACTARMCD does not equal ARMCD (SD2236)

This validation rule will fire if the Actual Arm Code (ACTARMCD) does not equal the Planned Arm Code (ARMCD).

The confusion associated with this validation rule is with the intent of the rule. A typical explanation of this validation issue is that the implementation is correct per CDISC implementation guidance, as ACTARMCD is allowed to be different from ARMCD. While that may be so, the real purpose of this validation rule is to identify subjects who were not treated as planned in a study. A proper dispositioning of this validation issue would be to explain in detail why the subjects were not treated as planned, as this information is useful to a reviewer Stating that the two variables are allowed to be unequal per the SDTM Implementation Guide does not provide useful information to a reviewer.

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.

Value not found in xx extensible codelist (CT2002)

This validation rule will fire if a value in the dataset, for a variable with a CDISC-defined codelist, does not exactly match a value in the CDISC extensible codelist.

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

- A sponsor uses a value that has no corresponding match in the CDISC codelist.
- 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.
- A sponsor combines values that should be in split into separate SDTM variables (for example, values of LEFT/RIGHT combined with an anatomical location in a location variable (--LOC) instead of the laterality variable (--LAT)).

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, and to list all actually valid extended values, if possible.

Variable and Decode values do not have the same Code in CDISC CT (CT2003)

In CDISC controlled terminology, for paired variables (--TESTCD/--TESTCD for example), both values will have the same NCI Code value. Therefore, using the LB domain as an example, for a value of LBTESTCD, the value of LBTEST must be the linked value from the LBTEST codelist with the same NCI Code. For example, the LBTESTCD value of ‘GLUC’ has a NCI Code value of C105585, and the LBTEST value for that record must use the value from the LBTEST codelist with the same NCI Code (C105585), which is ‘Glucose’.

Many times, a sponsor will use a value from the CDISC Synonyms column for the paired value, that does not match exactly the CDISC Submission Value that should have been used.

The most concerning scenario is when there is the inability to be sure which test was performed. This can be seen in an example from a sponsor’s Laboratory Results (LB) data. The sponsor used LBTESTCD = CYTYRO (NCI Code: C74683) for records in the LB domain. Here is the issue:

- The LBTEST value that matches LBTESTCD = CYTYRO is Tyrosine Crystals (NCI Code: C74683), however the sponsor used LBTEST = Triple Phosphate Crystals (NCI Code: C74756).
- This calls into question which test was actually performed. Was it Tyrosine Crystals, or Triple Phosphate Crystals?
- For this validation issue, the sponsor provided this explanation: "As per Controlled Terminology, codelist for LBTEST is extensible". This of course provides no useful information and does not even address the actual validation issue.

These controlled terminology mismatches for paired variables should be corrected, as in the best case they just point to an incorrect implementation of controlled terminology, but in the worst case they provide discrepant information that leads to uncertainty in the data.

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.

Value for variable not found in user-defined codelist (SD0037)

This validation rule will only fire when the define.xml is included when the datasets are validated. This 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.

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.

Invalid Term in codelist 'No Yes Response (Yes only)' codelist (DD0024)

This validation rule fires when a variable should only have a value of ‘Y’ or null, per CDISC implementation guidance, but in the define.xml that variable references a codelist that contains other values.

A common reason for this issue is that a sponsor will create one No/Yes codelist, and have many variables reference it, regardless if all of the values of that variable apply.

An example of this is the Subject Death Flag (DTHFL) variable in the Demographics domain. This variable, per CDISC guidance, should be ‘Y’ or null. However, it is common for sponsors to reference an No/Yes codelist in the define.xml for this variable, that contains values of ‘N’, ‘U’, etc. By referencing a codelist with these other values, it becomes unclear if the sponsor is using these values that aren’t allowed.

The solution to this issue is, for variables where only values of ‘Y’ are allowed, to reference a separate codelist with only this value.

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.

FDA/PMDA Expected variable not found (SD1077/SD1140)

The FDA Technical Conformance Guide 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.

This validation rules looks 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 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 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.

SDTM COMPLIANCE RULES

SDTM validation rules look for issues with how study data was mapped to SDTM. These are SDTM implementation inconsistencies, discrepancies, or deficiencies per CDISC guidance.

USUBJID/VISIT/VISITNUM values do not match SV domain data (SD0065)

The validation looks to make sure all visits in the SDTM datasets for a subject, with the exception of NOT DONE records, are included in the Subject Visits (SV) domain.

There seems to be four typical scenarios that cause this validation issue to exist:

- Incorrect implementation of the SV domain, where scheduled visits in the other SDTM datasets are just not mapped to the SV domain when they should have been.
- Unscheduled visits in the other SDTM datasets, for some reason, are not mapped to the Subject Visits (SV) domain.
- There are values of the VISIT variables in the SDTM datasets that a sponsor feels are not actual visits. These can be log pages (such as concomitant medications, etc.), drug dispensation visits, etc.
- Dates are missing for records from a certain visit for a subject, and therefore the records are not mapped to the SV domain. If all records for that visit for that subject are missing dates, the visit will not map to the SV domain.

Scenarios associated with incorrect implementation (the first two) should be corrected. Both scheduled and unscheduled visits should be used to generate the SV domain.

For the scenario where a sponsor doesn't feel a value populated in the VISIT variable is an actual visit, perhaps the VISIT variable is not the appropriate variable to use.

For the scenario where a record exists for a visit, but a date is missing, there is no guidance around whether or not these records should contribute to the generation of the SV domain. This is a case where due to the lack of guidance on an issue, the implementation is done inconsistently across the industry. The industry would benefit from an SDO or the regulatory agencies providing guidance on how this situation should be handled.

ECDOSE is not greater than 0 when ECOCCUR does not equal 'N' and ECDOSTXT is null (SD1247)

This is a new (at the time of writing this paper) CDISC SDTM conformance rule (CG0100), and an example of a validation rule with a complicated message.

The Exposure As Collected (EC) domain, as defined in the SDTM Implementation guide version 3.2 [2], represents the protocol-specified study treatment administrations, and should be used in cases where administrations are collected units different from the protocol-specified units (for example collected in tablets but protocol-specified unit is mg, etc.).

In the EC domain, the standard way to represent a missed dose is to set the Occurrence variable (ECOCCUR) equal to 'N'. Using a dose (ECDOSE) equal to 0 is not an acceptable alternative for representing missed doses [2].

This validation rule looks to identify those records where it appears that this invalid alternative method of representing missing doses is potentially being used. A record will be flagged when EXDOSE equals 0 and EXOCCUR does not equal 'N'.

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, 2017.
https://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM384744.pdf

[2] Study Data Tabulation Model Implementation Guide: Human Clinical Trials. Clinical Data Interchange Standards Consortium (CDISC) Submission Data Standards (SDS) Team. Version 3.2. November 2013

CONTACT INFORMATION

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

Michael Beers
Pinnacle 21 LLC
mbeers@pinnacle21.net