PhUSE EU Connect 2018

Paper DS06

Best Practice for Explaining Validation Results in the Study Data Reviewer's Guide

Kristin Kelly, Pinnacle 21 LLC, Plymouth Meeting, PA, USA

Michael Beers, Pinnacle 21 LLC, Plymouth Meeting, PA, USA

ABSTRACT

In addition to providing the datasets, define.xml and annotated CRF (aCRF) in a clinical SDTM package, the Technical Conformance Guide (TCG) recommends to also include a Study Data Reviewer's Guide (csdrg.pdf) to convey any further information about the study data that cannot be described in the aCRF or the define.xml that would aid the reviewer in understanding the data. An important section of this document contains results of automated validation checks for conformance to SDTM and regulatory business rules that cannot be resolved due to data oddities. Each issue that remains should be listed with an explanation for why the specific validation check cannot be rectified. Dispositioning validation findings can be a challenge, as some validation messages are more confusing than others. Providing a vague, incorrect response or no explanation at all can affect reviewability of study data. This paper will focus on examples and best practices for providing comprehensive explanations of issues in the cSDRG.

INTRODUCTION

For tabulation data (e.g., SDTM) collected during a clinical trial, it is recommended to include a Clinical Study Data Reviewer's Guide (cSDRG) in the study package. Per the Technical Conformance Guide[1] (TCG), this document should be named 'csdrg.pdf' (where 'c' designates clinical) and provided as a PDF for each study in module 5 (m5) of the eCTD. A sponsor may choose to organize this document to suit their needs but there is a recommended template developed by PhUSE that is widely used.[2]

The cSDRG typically contains the SDTM version used as well as dictionary/terminology (e.g. MedDRA, CDISC) versions for that study. It should provide information about the trial design and any domain-specific information that cannot be described in the define.xml, as well as results from automated validation checks. The cSDRG should also contain content to explain instances where data collected on a CRF may not be present in the SDTM datasets. In addition, issues encountered during conduct of the study or creation of the submission deliverables should be explained.

It is important to provide enough detail in the cSDRG so that a reviewer will be able to easily review the data. Transparency about the study data enhances traceability throughout the submission. This paper will focus on the Data Conformance section (Section 4) of the cSDRG and provide examples of explanations that will increase transparency and reviewability of the data.

BACKGROUND

The FDA has their own tools based on Pinnacle 21 Enterprise that run automated checks on the SDTM data when it is received as part of a submission and before it is passed to a reviewer. This is done to identify data conformance issues earlier in an effort to save time downstream in the review process. The FDA has published business and conformance rules based on the SDTM standard and FDA data requirements[3,4]. These rules check the adherence to their expectations of the content and quality of the data. Because of this, sponsors should also be running validation tools that include the FDA validation rules prior to submission.

As stated above, the cSDRG should have a section for results of automated validation items that check for conformance to SDTM that cannot be resolved due to data oddities. Each issue that remains should be listed in this section with a comprehensive explanation for why the specific validation check cannot be rectified. This is to be transparent about the submitted data because the FDA will see the same validation results. Not providing a complete response for a validation check results in having a reviewer spend time investigating why a particular issue remains and may also signal to them that there may be a bigger problem that is being minimized.

VALIDATION RULE EXPLANATION ISSUES

There could be many reasons why validation issues are not fully explained by the sponsor. One reason could be that the person completing this section of the cSDRG does not know how to investigate why a particular check is firing.

Another could be that they are unsure about which issues should be fixed versus those that cannot be rectified due to an oddity in the collected data. For example, even though a check may have a severity of 'Warning' and not 'Error', this does not mean it is okay to leave it as is and simply provide an explanation. This could result in having programming issues that remain in the data as well as incorrect or incomplete responses in the cSDRG.

The following examples show a few of the situations described above. The rule IDs from Pinnacle 21 Community and Enterprise are provided in each example. Each example contains snippets from the table provided in the Data Conformance section of the cSDRG. Also, when the word 'reviewer' is used, please note that this could be any consumer of the data downstream. This person could be an analyst that reviews the data before/after submission, a clinical programmer, a medical reviewer, etc.

RULES THAT MAY INDICATE PROGRAMMING ISSUES – TO FIX OR NOT TO FIX?

Sometimes it can be difficult to determine which validation issues should be rectified versus those that are due to some issue with data collection that cannot be 'fixed' after database lock.

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

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

Whether or not something is due to programming really depends on the rule in the report and needs to be evaluated on a case-by-case basis.

SD1043 – INCONSISTENT VALUE FOR --TESTCD WITHIN –TEST

This rule is based on the FDA business rule, FDAB009, that states the following: 'All paired variables must have a one-to-one relationship. Examples include Short Name and Name of Test; Parameter Name and Parameter Code or Number; Variable Name and Variable Label, etc'[3]. It is also a business rule for SDTM.

It is sometimes mistaken as being a data oddity that cannot be rectified. When SDTM datasets are created, the values of –TESTCD and –TEST in Findings domains are assigned based on CDISC controlled terminology. There should be a one-to-one relationship between –TESTCD and –TEST. This means that there can only be one –TESTCD value for a given –TEST and vice versa. Because –TESTCD/--TEST values are assigned using controlled terminology, if this rule appears in the validation report, this indicates a programming issue that should be fixed. A typical response for this type of issue is shown in the example below.

Check ID	Diagnostic Message	Severity	Dataset	Count (Issue Rate)	Explanation
SD1043	Inconsistent value for --TESTCD within --TEST	Error	EG	2(4%)	As per data.

Table 1. Incorrect explanation for SD1043

This programming issue is also checked by additional rules for other paired variables such as QNAM/QLABEL, ARMCD/ARM or VISITNUM/VISIT. These rules should not be listed at all in the cSDRG because the values in the dataset should be updated accordingly.

CT2003 – VARIABLE CODE AND DECODE VALUES DO NOT HAVE THE SAME CODE IN CDISC CT

In CDISC controlled terminology, for paired variables (e.g.--TESTCD/--TEST), 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 an 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 the CDISC Submission Value that should have been used. The most concerning scenario is when there is the inability to determine 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.

The following are more details of the issue:

- The LBTEST value that matches LBTESTCD = 'CYTYRO' is 'Tyrosine Crystals' (NCI Code: C74683), but 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 the following explanation below:

Check ID	Diagnostic Message	Severity	Dataset	Count (Issue Rate)	Explanation
CT2003	Coded and Decoded values do not have the same Code in CDISC CT	Error	LB	1(4%)	As per Controlled Terminology, codelist for LBTEST is extensible.

Table 2. Insufficient explanation for CT2003

This provides no useful information and does not address the actual validation issue. These controlled terminology mismatches for paired variables should be corrected in the data and this rule should not be listed in the cSDRG similar to the previous example. If the issue is not updated the best case is that it is an incorrect implementation of controlled terminology and the worst case is that they provide discrepant information that leads to uncertainty in the data.

SD0063 – SDTM/DATASET VARIABLE LABEL MISMATCH

In this example, the explanation in the cSDRG is not complete. This rule, SD0063, checks that the variable label in the dataset matches the variable label defined in the SDTMIG. Though the PE domain is noted as having a variable with an incorrect label, it is not stated in the explanation which variable label does not match. This leads to a reviewer referring to their version of the validation report to find the details for this issue.

Check ID	Diagnostic Message	Severity	Dataset	Count (Issue Rate)	Explanation
SD0063	SDTM/dataset variable label mismatch	Warning	PE	1 (6.67%)	Label is fine.

Table 3. Insufficient explanation for SD0063

Table 4 shows a complete explanation provided in the cSDRG that mentions the variable in PE that is affected, provides the revised label and the reason the label was updated.

Check ID	Diagnostic Message	Severity	Dataset	Count (Issue Rate)	Explanation
SD0063	SDTM/dataset variable label mismatch	Warning	PE	1 (6.67%)	Per the SDTMIG, the variable label in the dataset should match the label in the SDTMIG. In SDTMIG v3.2, there are several labels that were erroneously defined to be >40 characters. The variable label for PESTRESC, 'Character Result/Finding in Standard Format', exceeds 40 characters. To avoid truncation, PESTRESC was assigned the label 'Character Result/Finding in Std Format' in the dataset so that it is <= 40 characters due to SAS restrictions.

Table 4. Complete explanation for SD0063

In this example, the sponsor chose to make the label meaningful rather than truncating to 40 characters in length. Although SD0063 fired due to an error in the SDTMIG in this case, there could be other times when the wrong label is assigned for a variable. Thus, it is important to explain why it is acceptable for a label mismatch to remain, as in this case.

CT2002 – VARIABLE VALUE NOT FOUND IN EXTENSIBLE CODELIST

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 five 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 CDISC 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 (e.g, values of LEFT/RIGHT combined with an anatomical location in a location variable (--LOC) instead of the laterality variable (--LAT)).
- A sponsor collects OTHER, with a specify field, but in the SDTM variable concatenates OTHER with the specified value (e.g., OTHER: specified value).

Extending an extensible codelist when there is no corresponding value to use is an acceptable approach; however, the other scenarios are not. In this example, the sponsor originally did not map CMDOSU (Dose Units) values to the CDISC UNIT codelist, e.g. 'MG' in the data was not mapped to the value of 'mg' from the UNIT codelist. Sponsors tend to generically dismiss this validation issue with an explanation such as the following:

Check ID	Diagnostic Message	Severity	Dataset	Count (Issue Rate)	Explanation
CT2002	CMDOSU value not found in UNIT codelist	Warning	CM	1125 (35%)	Codelist is extensible.

Table 5. Insufficient explanation for CT2002

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 valid extended values, if possible. In this example, the same sponsor went back and mapped all possible values to the UNIT codelist and for those that could not be mapped, these are explained in the cSDRG.

Check ID	Diagnostic Message	Severity	Dataset	Count (Issue Rate)	Explanation
CT2002	CMDOSU value not found in UNIT codelist	Warning	CM	3 (<0.1%)	This rule fired for 'OTHER', 'DROPS', and 'UNK' values in CMDOSU. These values could not be mapped to CDISC CT but the UNIT codelist is extensible.

Table 6. Appropriate explanation for CT2002

SD0013 – --STDTC IS AFTER –ENDTC

This example illustrates a circumstance where the reason for a rule appearing in the validation results is due to a data oddity and not a programming error.

Check ID	Diagnostic Message	Severity	Dataset	Count (Issue Rate)	Explanation
SD0019	--STDTC is after --ENDTC	Error	AE	2 (6.67%)	This check fired for 2 records in AE where AESTDTC is after AEENDTC: USUBJID = XYZ-008, AESTDTC = 2011-09-14, AEENDTC = 2011-09-12 USUBJID = XYZ-300, AESTDTC = 2011-05-23, AEENDTC = 2011-05-19 For the two records flagged, the start and end dates were missed during data cleaning before database lock.

Table 7. Complete explanation for SD0013

The sponsor provided a comprehensive explanation that included details of the records that were flagged in AE by this check as well as the reason why the validation result remains.

RULES THAT REQUIRE INVESTIGATION – SEARCH AND RESCUE

There are some rules that appear in the validation results that require some investigation to create a comprehensive response. This could include referencing the protocol, the aCRF, and multiple domains. The following examples show scenarios where more examination of the data is required.

SD1117 – DUPLICATE RECORDS

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 it.

The following 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 with a very strict meaning of the word 'duplicate'. Therefore, it is common to see incorrect explanations in the cSDRG for this rule. A typical explanation from a sponsor for this validation rule might be something like:

Check ID	Diagnostic Message	Severity	Dataset	Count (Issue Rate)	Explanation
SD1117	Duplicate records	Warning	QS	324 (32.5%)	The keys defined by the check are not sufficient to identify a unique record for patient.

Table 8. Incorrect explanation for SD1117

This is an actual explanation from a sponsor, and the actual keys listed in define.xml for this domain, Questionnaires (QS) 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[5]. It acts as a surrogate for the variables that create the natural keys in the data so using this variable as a key provides no useful information regarding 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 --SPID (Sponsor-Defined Identifier). 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.

SD1077 – FDA/PMDA EXPECTED VARIABLE NOT FOUND

The FDA Technical Conformance Guide states that the variable, EPOCH, should be included for clinical subject-level observations (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, without having to compare dates to make these determinations.

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

A common example of a sponsor explanation for this validation rule is the following:

Check ID	Diagnostic Message	Severity	Dataset	Count (Issue Rate)	Explanation
SD1077	FDA Expected variable not found	Warning	AE	1 (100%)	EPOCH is a Permissible variable as per SDTM IG 3.1.3, hence not included.

Table 9. Incorrect explanation for SD1077

Although CDISC lists this variable as a Permissible variable, it does not seem appropriate to disregard regulatory guidance and justify it based on the fact that CDISC specifies these as Permissible variables.

Furthermore, simply including the EPOCH variable is not enough; sponsors should be sure to populate the EPOCH variable for each record possible and list a clear and complete derivation for this variable in the define.xml.

SD1118 – NEITHER --STDTC, --DTC NOR --STDY ARE POPULATED

In this example study, the rule, SD1118, appeared in the validation report for the DS domain and was explained in the cSDRG.

Check ID	Diagnostic Message	Severity	Dataset	Count (Issue Rate)	Explanation
SD1118	Neither DSSTDTC, DSDTC nor DSSTDY are populated	Warning	DS	28(33.33%)	The date was not collected for these subjects.

Table 10. Incomplete explanation for SD1118

The information available in the 'Details' tab for this rule in the report showed 28 records flagged in the DS domain where DSDECOD = 'COMPLETED' but no timing information was populated in DSSTDTC, DSDTC, or DSSTDY. No further investigation was done. To a reviewer, this may indicate a problem because there was no date collected for study completion especially when there is no further explanation in the reviewer's guide.

If further examination had been done, it could have been determined if the records flagged indicated a larger issue or not. In the protocol, 28 subjects were expected to be enrolled in the study. In the DM domain, there were 28 records indicating that 28 subjects were enrolled. All subjects had been randomized and treated as indicated in

ARMCD/ARM, ACTARMCD/ACTARM, and Reference Start Date/Time (RFSTDTC) (defined as first dose as noted in the define.xml) was populated.

In the DS domain, all subjects had two records each where DSDECOD = 'COMPLETED' for two different visits (VISIT = 'FOLLOW-UP/EARLY TERMINATION' and 'STUDY COMPLETION') but DSSTDTC was populated only where VISIT = 'STUDY COMPLETION'. The records for one subject in DS are provided below to illustrate this. Only the relevant variables are included.

USUBJID	DSSEQ	DSTERM	DSDECOD	VISITNUM	VISIT	DSSTDTC
ABC-123-008	1	INFORMED CONSENT OBTAINED	INFORMED CONSENT OBTAINED	1	SCREENING	2015-04-16
ABC-123-008	2	COMPLETED	COMPLETED	14	FOLLOW-UP/EARLY TERMINATION	
ABC-123-008	3	COMPLETED	COMPLETED	15	STUDY COMPLETION	2015-09-23

Table 11. DS Dataset Example

In the aCRF, there was a CRF for both the 'Follow-up' and 'Study Completion' visits and both were completed for all subjects which reflects why each subject had 2 completion records in DS where the VISIT values were different. Since the study was a Phase I study and all subjects completed it, the term 'COMPLETED' was collected twice but only the CRF for VISIT = 'STUDY COMPLETION' collected the date. If some time had been taken to investigate the issue, the explanation in the cSDRG may have looked like the following:

Check ID	Diagnostic Message	Severity	Dataset	Count (Issue Rate)	Explanation
SD1118	Neither DSSTDTC, DSDTC nor DSSTDY are populated	Warning	DS	28(33.33%)	This check fired for all 28 subjects enrolled where DSDECOD = 'COMPLETED' and DSSTDTC was null. There were two records for each subject where DSDECOD = 'COMPLETED' for the VISITs = 'FOLLOW-UP/EARLY TERMINATION' and 'STUDY COMPLETION'. DSSTDTC was collected on the CRF for the 'STUDY COMPLETION' visit only but the term 'COMPLETED' was collected twice for both visits. In the protocol, these two visits are defined as one visit. The records in DS where DSSTDTC is populated indicates that all completion information was collected for each subject, accordingly.

Table 12. Complete explanation for SD1118

If the preparer of the conformance section of the cSDRG had taken more time looking into the issue and explaining it, this would have saved a reviewer from having to do the 'search and rescue' work themselves. Providing this level of detail can also prevent a reviewer from thinking that this could signal a major oversight in how the data was collected, i.e., DSSTDTC not being collected for study completion.

BEST PRACTICES FOR PROVIDING RESOLUTIONS IN THE REVIEWER'S GUIDE

In the previous section, the examples were provided to convey how useful the explanations can be to a reviewer if the preparer of the cSDRG has some knowledge of why the rule was triggered, can determine if it can be rectified, and understands how to examine further than just the records flagged by the validation rule.

The following are a few best practices for those preparing the Data Conformance section of the cSDRG that will aid in the development of complete and correct explanations:

1. Consider keeping a log of specific checks that typically appear from study to study with the explanations provided. Keep in mind that explanations may vary across studies and should be reviewed prior to adding to an cSDRG. Use best judgement before just copy and pasting.
2. Different stakeholders review the P21 reports such as programmers, data managers, and publishers. As an organization, each group should create a knowledge base to aid each group in determining what may be a

programming issue, what may be a data collection issue, etc.

3. Be sure to review both the 'Rule Message' and the 'Rule Description' in conjunction with the 'Details' tab in the report for a particular rule. Taking all three into account clarifies why the rule fired.
4. If possible, refer to the records that were flagged in the physical datasets. It is helpful to see the entire record with all the relevant qualifiers and timing variables that may not appear in the report.
5. Refer to the protocol, define.xml, and aCRF if necessary. Sometimes just reviewing the dataset is not sufficient.
6. Be transparent! In explaining issues, provide as much detail as necessary, e.g. USUBJID, Variable values flagged, VISIT information, etc. Assure that the information provided is sufficient as to not bring questions from a reviewer. Construct the explanation in a way that it can be understood without looking in the 'Details' tab.
7. Remember that the SDTM standard is contained in two documents, the SDTM and the SDTMIG! Except for a few FDA-specific rules (e.g. adding EPOCH to domains), the validation rules were developed based on the information in BOTH documents. HINT: Section 4 contains many of the conventions that are being checked by Pinnacle 21 Community/Enterprise tools. Keep these documents handy when reviewing P21 reports and authoring explanations for validation issues.

CONCLUSION

Per the Technical Conformance Guide, it is important to include a Reviewer's guide with the SDTM tabulation data to help further explain study conduct as well as results from automated validation tools. Because each issue that remains is to be explained, it is important that the person reviewing the validation reports has a clear understanding of the validation rules so that correct and comprehensive explanations can be provided. This conveys transparency about the study data that increases the reviewability. The more detailed an explanation, the less questions will be asked, which in turn, could shorten the duration of a review. Though the cSDRG is one small piece of a submission, it is an integral part of facilitating review in an effort to get drugs to patients faster.

REFERENCES

[1] U.S. Food & Drug Administration, *Study Data Technical Conformance Guide* (current version at time of access – v4.0/ October 2017), Accessed March 15, 2018 - https://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM384744.pdf

[2] PhUSE, *Study Data Reviewer's Guide* (current version at time of access – v1.2/ January 2015), Accessed March 15, 2018 - http://www.phusewiki.org/wiki/index.php?title=Study_Data_Reviewer%27s_Guide

[3] U.S. Food & Drug Administration, *FDA Business Rules* (current version at time of access – v1.3/December 2017), Accessed March 15, 2018 - https://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm

[4] U.S. Food & Drug Administration, *FDA Validator Rules* (current version at time of access – v1.2/December 2017), Accessed March 15, 2018 - https://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm

[5] 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:

Kristin Kelly
Pinnacle 21, LLC
kkelly@pinnacle21.com

Michael Beers
Pinnacle 21, LLC
mbeers@pinnacle21.com