PharmaSUG 2022 - Paper SS-152

Explanations in the Study Data Reviewer's Guide: How's It Going?

Kristin Kelly, Pinnacle 21 LLC

ABSTRACT

In 2018, the paper 'Best Practice for Explaining Validation Results in the Study Data Reviewer's Guide' was presented at the PharmaSUG conference. Its focus was to provide sponsors with some best practices for explaining validation results in the Study Data Reviewer's Guide (cSDRG) as well as guidance for interpreting some of the more confusing validation rules in Pinnacle 21. This paper will focus on metrics gathered across industry since that time of the quality of explanations as well as best practices that will aid in improving them even more.

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 metrics collected across industry to provide a gauge of how well sponsors have improved when providing explanations.

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]. CDISC also has published rules for SDTM that are included in the Pinnacle 21 rule sets. These rules check the adherence to expectations of the content and quality of the data. Because of this, sponsors should also be running validation tools that include both the FDA and CDISC 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.

INDUSTRY METRICS FOR RULES AND EXPLANATIONS

The industry metrics collected for rules and explanations have been sourced through P21 Enterprise. The

Enterprise version has functionality that allows users to enter explanations directly into the software. When the study datasets, define.xml and other components for submission are finalized, the user can then export the cSDRG from P21 Enterprise with the explanations populated in the Conformance section.

HIGH-LEVEL METRICS

Some high-level metrics for this sample are the following:

Total Number of Issue Explanations	150,128
Total Number of Unique Sponsors/CROs	90
Total Number of Unique Rules	734

Based on this, out of the 150K explanations collected across 90 sponsors/CROs, more than 700 unique rules needed to be explained in the cSDRG.

RULE HIGHLIGHTS

Of the 734 unique rules that have explanations applied, CT2002 is the most explained rule (33,478). This is 3.5 times the number of explanations for the second most frequently appearing rule, SD1076 (9,391). The top 20 rules that account for only 2.6% of the total unique rules account for ~60% of total explanations.

Rule	Rule Message	Explanation Counts	Percentage
CT2002	Variable value not found in extensible codelist	33478	22.24%
SD1076	Model permissible variable added into standard domain	9391	6.24%
SD1082	Variable length is too long for actual data	4475	2.97%
SD1078	Permissible variable with missing value for all records	4398	2.92%
SD1339	Missing EPOCH value, when a start or observation date is provided	4091	2.72%
SD1117	Duplicate records	3893	2.59%
SD0058	Variable appears in dataset, but is not in SDTM model	2931	1.95%
SD0022	Missing Start Time-Point value	2845	1.89%
SD0021	Missing End Time-Point value	2622	1.74%
AD1012	Secondary custom variable is present but its primary variable is not present	2609	1.73%
SD1230	Variable datatype is not %Variable.@Clause.DataType% when value-level condition occurs	2568	1.71%
SD1231	Variable value is longer than defined max length %Variable.@Clause.Length% when value-level condition occurs	2559	1.70%

Rule	Rule Message	Explanation Counts	Percentage
SD1203	--DTC date is after RFPENDTC	2054	1.36%
AD0018	Variable label mismatch between dataset and ADaM standard	1985	1.32%
SD1201	Duplicate records in Events domain	1720	1.14%
SD0026	'Missing value for --ORRESU, when --ORRES is provided	1585	1.05%
SD0065	USUBJID/VISIT/VISITNUM values do not match SV domain data	1536	1.02%
SD0037	Value for variable not found in user-defined codelist	1534	1.02%
SD1124	Missing value for --REASND, when --STAT is 'NOT DONE'	1517	1.01%
SD1202	--STDTC date is after RFPENDTC	1501	1.00%

Table 1. Top 20 most explained rules

DATASET HIGHLIGHTS

There were 1,693 unique datasets in the sample. The top 20 datasets accounted for 60% of the explanations in the cSDRG.

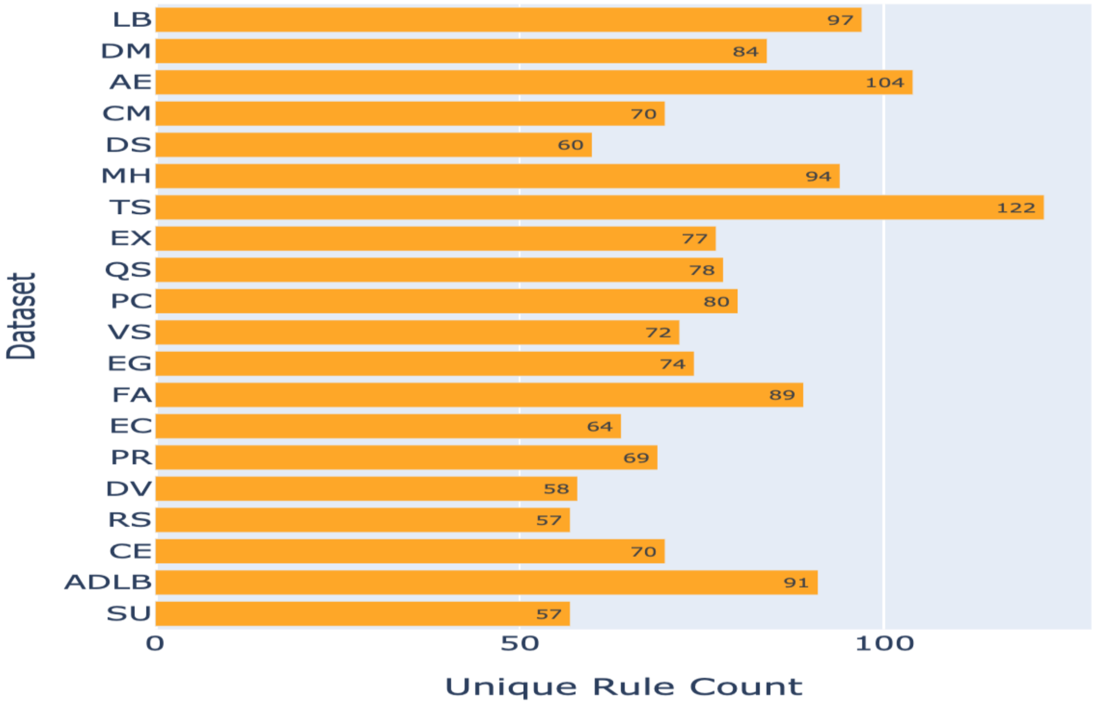

Table 2. Number of Unique rules per Dataset

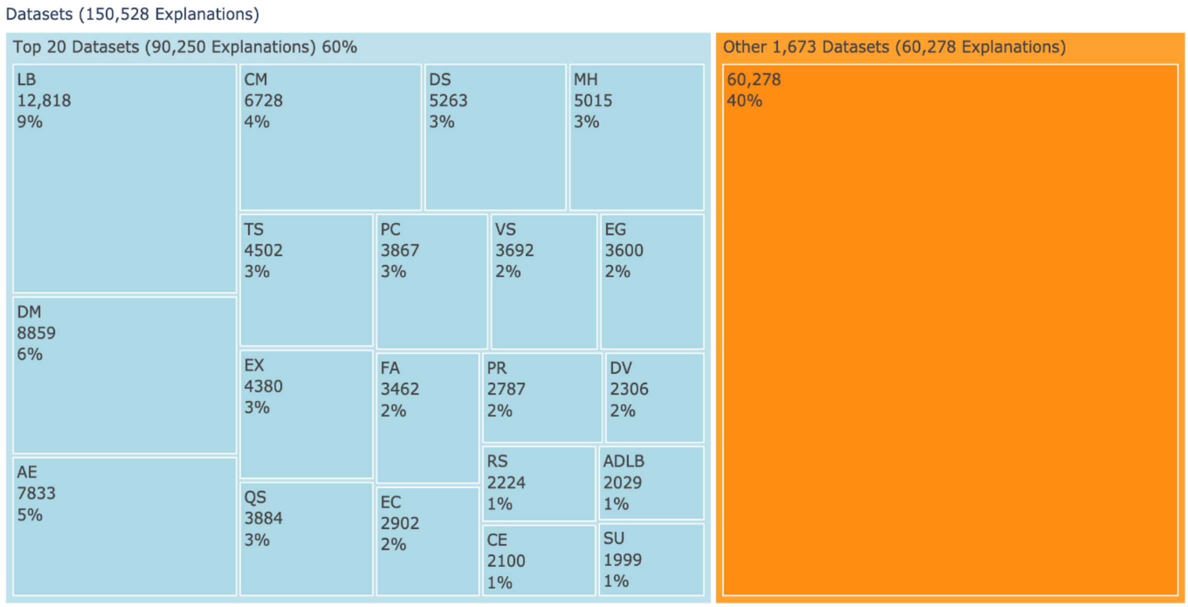

Table 3. Distribution of Issue Explanations by Dataset

TEXT ANALYSIS OF ISSUE EXPLANATIONS

The statistics reported are based on explanations that have been cleaned. In Natural Language Processing (NLP), ‘cleaned’ text typically refers to strings which have had all punctuation, digits, whitespace, and stop words (common English words such as ‘the’, ‘is’, ‘an’ etc.) removed. It is then tokenized (split into a list of words). These transformations allow focus on more meaningful text.

Some high-level metrics of text is shown in the table below.

Total Number of Unique Words (across explanations)	18,721
Median Number of Words per Explanation	15
Total Number of Explanations that Contain ‘False Positive’	4,790

One of the more interesting metrics from the text analysis was the number of explanations that contained the text, ‘False Positive’ at nearly 3.2% of total explanations. In validation, a ‘false positive’ is when a rule is triggered but no issue exists in the data, i.e., a rule fired when it shouldn’t have. At times, the rule is truly a false positive and the rule is refined in order to reduce or eliminate instances when it fires erroneously. But most of the time, the rule is working as expected and the explanation just states that it was a ‘false positive’ message and that there is no issue in the data.

METRICS FOR FALSE POSITIVE MESSAGES

So which rules are being explained as ‘False Positives’? The top 5 rules that are being dispositioned as ‘false positive’ messages are shown in the table below:

Rule ID	Rule Message	Explanation Counts
SD1230	Variable datatype is not %Variable.@Clause.DataType% when value-level condition occurs	1624
SD1231	Variable value is longer than defined max length %Variable.@Clause.Length% when value-level condition occurs	1153
CT2002	Variable value not found in extensible codelist	553
SD1082	Variable length is too long for actual data	97
SD1212	--STRESN does not equal --STRESC	39

Each rule with typical explanations will be discussed in the following sections.

'False Positives' for SD1230 and SD1231

Rule ID	Rule Message	Typical Explanation
SD1230	Variable datatype is not %Variable.@Clause.DataType% when value-level condition occurs	The issue is confirmed to be false positive in P21E. The issue occurs when LBSTRESC has null value in parameter value level metadata.

SD1230 will be triggered if the values in the dataset do not match the datatype noted for a parameter in value-level metadata (VLM) in the define.xml. For example, the value is 'float' type in the dataset but the define.xml has a datatype of 'integer' for a given –TESTCD/--TEST. It has also been triggered when there are null values in the data. In this case, SD1230 fires as a false positive message. This rule has been adjusted to not fire in this case and the fix was released in the P21 2010.1 engine. Due to this fix, SD1230 should no longer be explained as a false positive in the cSDRG. The issue should be fixed in the define.xml so that rule does not surface in the validation report.

Rule ID	Rule Message	Typical Explanation
SD1231	Variable value is longer than defined max length %Variable.@Clause.Length% when value-level condition occurs	False positive due to bug in P21E. Length is correct in define.xml. Bug has been reported and according Pinnacle 21 will be corrected in a future release.

SD1231 checks that the length of the values in the dataset match the length assigned in the define.xml for a given parameter in the VLM. It has been reported as a false positive for 'float' datatype values because the validation engine also counts the decimal point in the length. For example, for a 'float' datatype value of '4.67', the length should be '3' but if the define.xml has a length of '3', SD1231 will be triggered because the engine does not recognize the 'float' datatype. When this rule fires, it indicates that the max length in the define.xml is not long enough. Currently, the 'float' datatype is the only instance where SD1231 fires as a false positive. All other cases should be rectified in the define.xml and not explained as a false positive. P21 has since updated the engine to recognize 'float' values and the fix will be available in a future release. In the interim, users can increase the length in the define.xml for the affected parameters and the rule will no longer fire.

‘False Positives’ for SD1082 and SD1212

Rule ID	Rule Message	Typical Explanation
SD1082	Variable length is too long for actual data	According to FDA technical conformance guide section 3.3.3: The allotted length for each column containing character (text) data should be set to the maximum length of the variable used across all datasets in the study except for suppqual datasets. Pinnacle 21 provides false positive information since it only checks the length of the variable within the data set.

SD1082 fires as a false positive message when the lengths of the variables are set to be consistent across all datasets. This follows the guidance provided in the FDA sdTCG. When the data is trimmed in this manner, the explanation provided above seems reasonable. The fix for this rule will require an engine update where variable length will be checked across all datasets rather than within the same dataset. This change should be available in a future release.

Rule ID	Rule Message	Typical Explanation
SD1212	--STRESN does not equal --STRESC	The values in RESTRESN and RESTRESC are identical. These two records are false positives generated because the validation software rounded RESTRESN before comparing it to the character value in RESTRESC.

SD1212 checks that the value in –STRESN is the same as the value in –STRESC. Most of the time, the rule works as expected and the data should be updated so that the rule will no longer fire.

Sometimes, the rule will fire as a false positive:

1. When the value contains more then 7 decimal places, the validation engine will round the value in –STRESN before it compares to the value in –STRESC
2. When the value in –STRESC contains leading or trailing zeroes that are not carried over to –STRESN

If one or both of these two cases is present in the data, then the explanation provided above is valid. Otherwise, the data should be updated and this issue should not explained in the cSDRG. Refining this particular rule requires an update to the engine that will be in a future release.

CT2002 Issue Explanation Analysis

Rule ID	Rule Message
CT2002	Variable value not found in extensible codelist

CT2002 accounts for a little over 1 in 5 explanations in the cSDRG. There are 1,015 unique datasets that have an explanation for this rule. As the most explained rule, covering 22.24% of all explanations, there are about 15,415 unique explanations (~50%). While this does imply that more explanations are unique to the specific data for which CT2002 is triggered, it is also an indication that the explanation for this rule is still very generic and lacks granularity. The below word cloud highlights the most commonly occurring words used for CT2002. The size of each word is an indication of frequency.

sponsor defined
unit
extended terminology
sponsor extension
treatment cycle
sdtm ig
provided
extensible codelist
additional term
study protocol
ct
codelist epoch
treatment added
terminology codelist
unknown
added
already existing
codelist extended
epoch value
test code
collected
quality
extensible
study need
treatment period
issue
found
new
extended
new term
mapped
cdisc codelist
race
test name
cdisc ct
added term
false positive
allowed
include
value
existing term
codelist
extensible
period treatment
meet study
checked
extensible epoch
variable
protocol needs
multiple
open label
crf
controlled terminology
study specific
ng ml
extensible code
include value
included
used
study design
misspelling
codelist race
duplicate value
values already
code list
epoch codelist
cdisc controlled
unit extensible
follow
epoch extensible

The table below shows the top 10 most frequent explanations for CT2002 from the text analysis sample.

#	Explanation	Count
1	The codelist is extensible and additional terms were added to meet study needs. These added terms were checked for duplicate values, already existing terms or misspelling against the CT.	231
2	Codelist is extensible	199
3	Acceptable, as codelist is extensible.	178
4	Sponsor has defined controlled terminology (CT) that has not yet been defined by CDISC. These CT values are not synonymous to any of the existing terms in the CDISC codelist.	154
5	This CDISC controlled terminology codelist is extensible. New terms are allowed to be added.	138
6	The CDISC controlled terminology codelist is extensible. New term is permitted to be added.	132
7	This is a false positive. The codelist is extensible and additional terms were added to meet study needs. These added terms were checked for duplicate values, already existing terms or misspelling against the CT.	132
8	This CDISC controlled terminology codelist is extensible. New terms are allowed to be added.-√°	130
9	Extensible Codelist, no issue	118
10	Codelist is extensible.	117

Please note that #7 on this list mentions ‘false positive’ for CT2002. CT2002 is checking that values in the data for a variable that is subject to CDISC Controlled Terminology match a term in that particular extensible codelist. Here, ‘extensible’ means that values can be added to extend it, if needed. This rule only checks against CDISC CT and does not take into account implementation strategy provided in the SDTMIG, e.g. ‘OTHER’ or ‘MULTIPLE’. When the rule is triggered, it is NOT firing a false positive message. The rule is working as expected!

In the future, this rule could be further refined to consider implementation strategy in the SDTMIG so that common terms such as ‘MULTIPLE’ and ‘OTHER’ do not trigger CT2002. Hopefully, this will reduce the number of times the rule fires for a study as well as how often it needs to be explained in the cSDRG. Also, since this rule has to be explained so often when this doesn’t necessarily mean there is an issue in the data, perhaps there can be a threshold for when CT2002 needs to be explained in the cSDRG.

SO REALLY, HOW IS IT GOING?

Since this topic was last presented in 2018, how are things going? It is this author’s opinion that overall, explanations provided in the cSDRG have improved. There seem to be far less instances of generic explanations that are simply not useful to a reviewer. Examples include: ‘As per data’, ‘Data as collected’, or ‘Mapped as-is’. For the most part, explanations reflect that the issue has been investigated and a complete, correct response is provided.

That being said, there is still work to be done in determining whether a rule is truly a ‘false positive’ or not. Hopefully, over time, users will become even more experienced in making the distinction regarding whether a rule is working properly or not.

As always, the P21 tool is always being improved in order to ensure that industry receives accurate validation results. Though the cSDRG is one small piece of a submission, it is an integral part of facilitating review in an effort to get drugs (and vaccines) 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

[6] Kelly, Kristin. 2018. ‘Best Practice for Explaining Validation Results in the Study Data Reviewer’s Guide’. PharmaSUG 2018, Phuse US Connect 2018, and Phuse EU 2018. https://www.pharmasug.org/proceedings/2018/SS/PharmaSUG-2018-SS13.pdf

ACKNOWLEDGEMENTS

Thank you to Kirsten Myers, Data Scientist, that provided the metrics and analytics for this topic. All I had to do was review them, write up a paper and present it.

CONTACT INFORMATION

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

Kristin Kelly
Pinnacle 21, LLC
kkelly@pinnacle21.com