

Confusing Data Validation Rules Explained

PhUSE US Connect 2018

Paper #RG03

Michael Beers, Pinnacle 21

Sources of Confusion

- › Changing data standards and regulatory requirements
- › Contradictory guidance between SDOs and regulatory agencies
- › Unclear or complicated validation message
- › Misconception of a certain concept
- › Misunderstanding of the purpose of the validation rule

Data Quality Rules

These validation rules identify issues with the collection of the data, deficiencies in the data, or issues that may otherwise affect reviewability.

Duplicate Records (SD1117)

- › This validation rule looks to identify potential duplicate records, based on a set of meaningful, common industry-wide keys.
- › Variables without standard definitions are not used (sponsor-defined variables and suppquals)
- › This rule does not check for duplicate records against the sponsor defined keys in the define.xml.

Duplicate Records (SD1117)

- › Common scenarios:

- › Actual duplicate information, except for –SEQ

USUBJID	LBSEQ	LBTEST	LBORRES	LBORRESU	LBDTC
001-001	1	Hemoglobin	15	g/dL	2012-03-06T10:10
001-001	2	Hemoglobin	15	g/dL	2012-03-06T10:10

- › Same timing information but results are different

USUBJID	LBSEQ	LBTEST	LBORRES	LBORRESU	LBDTC
001-001	1	Hemoglobin	15	g/dL	2012-03-06T10:10
001-001	2	Hemoglobin	18	g/dL	2012-03-06T10:10

Duplicate Records (SD1117)

- › Common scenarios, cont.:
 - › Same timing information but one record is NOT DONE.

USUBJID	LBSEQ	LBTESTCD	LBORRES	LBORRESU	LBSTAT	LBNAM	VISIT	LBDTC
001-001	1	ALB	4.7	g/dL		ABC	UNSCHEDULED	2008-12-15
001-001	2	ALB			NOT DONE	ABC	UNSCHEDULED	2008-12-15
001-001	3	ALP	97	IU/L		ABC	UNSCHEDULED	2008-12-15
001-001	4	ALP			NOT DONE	ABC	UNSCHEDULED	2008-12-15

- › Only differentiated by a sponsor-defined variable.

USUBJID	LBSEQ	LBSPID	LBTEST	LBORRES	LBORRESU	LBDTC
001-001	1	123456	Hemoglobin	15	g/dL	2012-03-06T10:10
001-001	2	654321	Hemoglobin	15	g/dL	2012-03-06T10:10

Duplicate Records (SD1117)

- Typical explanation from sponsors:
 - "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.
 - Many times sponsors explain that --SPID, --GRPID, etc. need to be used, but the define.xml doesn't describe what these variables contain.

Controlled Terminology Rules

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

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

- › There are some paired variables (--TESTCD/--TEST, TSPARMCD/TSPARM, etc.) in CDISC CT. The code variable value and decode variable value share the same NCI code.

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

> Example:

> Values in CDISC Controlled Terminology

Code	Codelist Code	Codelist Extensible (Yes/No)	Codelist Name	CDISC Submission Value	CDISC Synonym(s)
C74683	C65047		Laboratory Test Code	CYTYRO	Tyrosine Crystals
C74683	C67154		Laboratory Test Name	Tyrosine Crystals	Tyrosine Crystals
C74756	C65047		Laboratory Test Code	CYTRPHOS	Triple Phosphate Crystals
C74756	C67154		Laboratory Test Name	Triple Phosphate Crystals	Triple Phosphate Crystals

> Values in the dataset:

USUBJID	LBTESTCD	LBTEST	LBORRES	LBORRESU	VISIT	LBDTC
001-001	CYTYRO	Triple Phosphate Crystals	1+		Visit 3	2016-05-13T10:27
001-001	CYTYRO	Triple Phosphate Crystals	2+		Visit 4	2016-06-13T11:30

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

- › Typical explanation from sponsors:
 - › “Controlled Terminology was applied where applicable. LBTESTCD/LBTEST codelist is extensible.”

Metadata Rules

These validation rules identify issues with the define.xml, such as 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 looks to identify values in the dataset that are not listed in the associated codelist in the define.xml.

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

- › Common scenarios:
 - › Casing difference or misspelling between the define.xml codelist and the data value.
 - › 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.

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

› Example:

- › LBTOXGR has a codelist assigned in the define.xml:

Variable	Label	Key	Type	Length	Controlled Terms or Format	Origin
LBTOXGR	Standard Toxicity Grade		text	1	Common Terminology Criteria for Adverse Events	Assigned

- › The codelist in the define.xml has these values:

Common Terminology Criteria for Adverse Events [CL.TOXGR]	
Permitted Value (Code)	Display Value (Decode)
1	Grade 1; 1
2	Grade 2; 2
3	Grade 3; 3
4	Grade 4; 4
5	Grade 5; 5

- › The dataset has these values:

LBSEQ	LBTESTCD	LBTEST	LBCAT	LBNRIND	LBTOXGR	VISITNUM
1	ALB	Albumin	CHEMISTRY	LOW	1	-1
2	ALB	Albumin	CHEMISTRY	LOW	1	1
3	ALB	Albumin	CHEMISTRY	NORMAL	0	104

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

- › Typical explanation from sponsors:
 - › “The codelist is extensible.”
- › Metadata issues should be corrected.

Regulatory Conformance Rules

These validation rules look for violations of regulatory agency (FDA and PMDA) guidance in their Technical Conformance Guides, such as missing requested data, and implementations inconsistent with the regulatory guidance.

FDA Expected variable not found (SD1077)

- › This validation rule looks to make sure that the EPOCH variable is provided in the appropriate domains.
- › Confusion is due to
 - › CDISC guidance listing this variable as permissible
 - › Technical Conformance Guide stating that the EPOCH variable should be provided

FDA Expected variable not found (SD1077)

- › Common scenarios:
 - › EPOCH is not provided for domains that the sponsor deems unnecessary, such as IE, etc.
 - › EPOCH is missing due to no collected timing information (and therefore not possible to derive).
 - › EPOCH is not provided for domains that the sponsor thinks conflicts with CDISC guidance, such as SE and SV.
 - › EPOCH is just missing when it should have been provided.

FDA Expected variable not found (SD1077)

- › Typical explanation from sponsors:
 - › "EPOCH is a permissible variable as per SDTM IG 3.1.3, hence not included."
- › It does not seem appropriate to disregard regulatory guidance based on CDISC variable core status.
- › In most cases, EPOCH should be provided and populated.

SDTM Compliance Rules

These validation rules identify issues with how study data was mapped to SDTM, such as implementation inconsistencies, discrepancies, or deficiencies per CDISC guidance. .

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

- › This validation rule 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.

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

- › Common scenarios:
 - › Incorrect implementation of the SV domain (missing scheduled visits in the SV domain).
 - › Unscheduled visits are missing in the SV domain.
 - › VISIT values that a sponsor feels are not actual visits, and have not been mapped to the SV domain (log pages for example).
 - › Dates are missing for a visit for a subject, and therefore the records are not mapped to the SV domain.

Summary

- › 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 validation rules leads to:
 - › Important issues not being corrected
 - › Regulatory agencies not receiving the information they need
 - › Data issues not being explained sufficiently

Questions

Michael Beers

mbeers@pinnacle21.net