PhUSE 2015

Paper PP09

Errors beyond OpenCDISC Validator

Usha Kumar, inVentiv Health Clinical, Mumbai, India
Neha Mohan, inVentiv Health Clinical, Mumbai, India
Gayatri Karkera, inVentiv Health Clinical, Mumbai, India

ABSTRACT

In clinical trials there is a continuing effort towards streamlining the processes to facilitate review by the regulatory bodies. One of the efforts in these lines was the development of the CDISC standard for datasets. The CDISC mission is "to develop and support global, platform-independent data standards that enable information system inter-operatability to improve medical research and related areas of healthcare". For regulatory submissions we usually run our SDTM and ADaM datasets through the OpenCDISC validator in order to ensure compliance with CDISC standards. CDISC validator captures lots of errors like inconsistency of visit values across domains, missing values in key variables etc. The datasets may individually adhere to the standards but would they guarantee traceability from ADaM to SDTM? We will see in this paper some examples of such errors that may not necessarily be thrown up through the validator.

INTRODUCTION

OpenCDISC Validator is a simple, yet powerful open source tool developed in an effort to make the compliance of data standards easily manageable within the pharmaceutical industry. The following list of some key checks illustrates the scope of the tool:

- Matching of variable's data against pre-defined set of values
- Uniqueness of variable's value across all records
- Conditional check between variable's values
- Check on specific value for a variable when pre-defined conditions are met
- Cross-data checks across multiple domains
- Meta-data checks

The checks built within the tool provide guidance in ensuring conformance to the data standards in terms of the structure and format; they are not be looked at as a means for ensuring complete data integrity. We will now see some examples that demonstrate the kind of errors we need to be cautious about beyond the OpenCDISC validation report.

CASE 1: CONSISTENCY FOR TRACEABILITY

DATA FORMAT

Consider a case where we are pooling data into an integrated ADaM from individual SDTM of contributing studies for integrated safety/efficacy analysis. Individually each of the SDTM dataset conforms to the data standards applicable to that specific domain. However, this need not guarantee conformance when we pool data from SDTM into integrated ADaM. The key feature of ADaM being traceability back to SDTM, it is required that we ensure data format consistency between SDTM and pooled ADaM. We will see below a scenario further explaining this case where SDTM DM from Study 1 and Study 2 are used to create integrated ADSL analysis dataset

Snapshot of SDTM DM from Study 1

STUDYID	DOMAIN	USUBJID	SUBJID	RFSTDTC
001	DM	001-99-123	123	2015-10-21
001	DM	001-99-124	124	2015-10-21

Snapshot of SDTM DM from Study 2

STUDYID	DOMAIN	USUBJID	SUBJID	RFSTDTC
02	DM	02-88-200	200	2015-12-25
02	DM	02-88-201	201	2015-12-25

Snapshot of pooled data in ADSL

STUDYID	DOMAIN	USUBJID	SUBJID	RFSTDTC
001	DM	001-99-123	123	2015-10-21
001	DM	001-99-124	124	2015-10-21
002	DM	002-88-200	200	2015-12-25
002	DM	002-88-201	201	2015-12-25

In the above snapshots, we see that while trying to have a consistent format of variables STUDYID and USUBJID in the pooled ADSL dataset, we have missed to take care of the traceability from ADaM to individual SDTM source. The records highlighted in red hence need correction in ADSL to ensure that the data of the variables USUBJID and STUDYID when carried over from SDTM to ADaM do not undergo any transformation. In other words we need to ensure the variables being carried forward from SDTM to ADaM should remain unchanged.

Corrected ADSL

STUDYID	DOMAIN	USUBJID	SUBJID	RFSTDTC
001	DM	001-99-123	123	2015-10-21
001	DM	001-99-124	124	2015-10-21
02	DM	02-88-200	200	2015-12-25
02	DM	02-88-201	201	2015-12-25

RE-DERIVATION

As part of the standard guidelines, complete traceability must exist between SDTM and ADaM. The variable being carried over from SDTM to ADaM should not hence be re-derived, though not a collected field.
Let us understand this case using the below example from CM domain.

Snapshot of SDTM CM

STUDYID	DOMAIN	USUBJID	CMSEQ	CMTRT	CMDECOD	CMDOSE	CMDOSU	CMSTDTC
1234	CM	1234	1	Med B	Med B	50	Mg	2004-01-17
1234	CM	1234	2	Med C	Med C	200	Mg	2002-01-15

Snapshot of ADaM ADCM

STUDYID	DOMAIN	USUBJID	CMSEQ	CMTRT	CMDECOD	CMDOSE	CMDOSU	CMSTDTC
1234	ADCM	1234	1	Med B	Med B	200	Mg	2002-01-15
1234	ADCM	1234	2	Med C	Med C	50	Mg	2004-01-17

In the above example, the SDTM follows SEQ creation by sort order of medication term CMTRT whereas the same has been re-derived in ADaM by sorting the data by the medication start date CMSTDTC. In order to ensure full traceability between SDTM and ADaM, the variable being carried over from SDTM to ADaM should not hence be re-derived.

CASE 2

IMPUTATION OF VALUES

SDTM must present the data as collected; no imputation is allowed in the collected fields.
In some countries, only birth year is collected. In such case, though imputation is done to derive age, we should not be imputing the birth date for completeness in SDTM dataset. Say we have only birth year captured as year 2015. Age is being derived in analysis dataset by imputing '01' for day and month components of the birth date. However, SDTM BRTHDTC variable should have the value as '2015' and not as '01-01-2015'.

CASE 3

REPLICATION OF INFORMATION

We need to ensure that the information is not duplicated in multiple domains. There should be just one mapping destination even though the information might seem correct fit in more than one domain

CASE 4

EXISTENCE OF DUMMY DATA

To support analysis, derivations are done and often new records are created in the ADaM datasets. The additional records are not to be added/created if they are not part of the original study design as these will result in addition of dummy records. The inclination to add such dummy records is higher when integrating data as in the example below.

Snapshot of SDTM SE for Study 1 with follow-up visits

STUDYID	DOMAIN	USUBJID	SUBJID	ETCD	ELEMENT	SESTDTC	SEENDTC
001	SE	001-99-123	123	SCRN	Screening	2015-10-01	2015-10-05
001	SE	001-99-123	123	TRTA	Treatment	2015-10-06	2015-10-20
001	SE	001-99-123	123	FUP	Follow-up	2015-10-21	2015-10-30

Snapshot of SDTM SE for Study 2 with no follow-up visits per study design

STUDYID	DOMAIN	USUBJID	SUBJID	ETCD	ELEMENT	SESTDTC	SEENDTC
002	SE	002-99-100	100	SCRN	Screening	2015-10-01	2015-10-05
002	SE	002-99-100	100	TRTA	Treatment	2015-10-06	2015-10-20

In trying to pool the data, there might be an inclination to add dummy record for follow-up visit in Study 2 (highlighted in table below) to match the elements defined in the different studies.

Snapshot of SDTM SE for Study 2 with dummy follow-up visit for pooled analysis

STUDYID	DOMAIN	USUBJID	SUBJID	ETCD	ELEMENT	SESTDTC	SEENDTC
002	SE	002-99-100	100	SCRN	Screening	2015-10-01	2015-10-05
002	SE	002-99-100	100	TRTA	Treatment	2015-10-06	2015-10-20
002	SE	002-99-100	100	FUP	Follow-up	2015-10-20	2015-10-20

The additional dummy record should not to be added as it is not part of the original study design.

CASE 5

CORRECT MAPPING TO CODELIST

There are certain controlled terminologies tied to specific variables as we may see from the SDTM standards document. OpenCDISC validator would only check for the values in the data to be one of the listed values in the controlled terminology applicable for that particular variable. However, we need to be careful that values are mapped to the controlled terminologies correctly.

For instance, variable SEX in the DM domain has the controlled terms ‘M’, ‘F’ and ‘U’ for Male, Female and Unknown respectively. However, for a study, 0/1 could correspond to M/F or F/M based on the way data has been collected. Hence we need to ensure that the collected values are mapped correctly when trying to match with the controlled terms per CDISC standards.

CASE 5

CONSISTENCY OF VALUES

Let us look at an example of pregnancy test results captured as “NEGATIVE” and “POSITIVE”. These original result values are captured in LB domain under variable LBORRES. The corresponding standardized values being the same, variable LBSTRESC will also reflect values as “NEGATIVE” and “POSITIVE” as in LBORRES. The corresponding, numeric equivalent (LBSTRESN) will however be set to missing instead of mapping to any numeric values like 0/1

On the other hand, in cases such as erythema rating score captured as “Clear or almost clear”, “Mild”, “Moderate”, “Severe” with corresponding standardized values as 1, 2, 3, 4 these standardized values will be captured under both STRESC and STRESN variables. We cannot have STRESC populated as “Clear or almost clear”, “Mild”, “Moderate”, “Severe”. These original values will only be captured under ORRES.

CASE 5

COMPLETE CRF DATA IS CAPTURED IN SDTM

We need to ensure that all of the CRF data is captured in one or the other SDTM domains as per the applicable fit guidelines outlined in CDISC document. There is high likelihood of missing out on records when we pull the data for that domain from multiple sources. For example, comments on various domains are captured as free text. As these

are pulled from various sources, we need to ensure that the count of records from different sources match the total records that are created in the CO domain when mapping non CDISC raw data to SDTM.

CASE 6

REDUNDANT NULL VARIABLES

Null variables, if not collected should be removed from the domain if permissible. There are lots of tests that have result captured without any unit. In such a cases, the unit column (STRESU) can be dropped. Similarly, when we have scoring data with coded values captured in FA domain, there are no units. We can hence drop the column FASTRESU from the domain.

CONCLUSION

For FDA submissions we usually run our SDTM and ADaM datasets through the Open CDISC validator in order to ensure compliance with CDISC standards. The datasets may individually adhere to the standards but they might not guarantee traceability from ADaM to SDTM. The examples illustrated in this paper are just a few cases of errors we need to be careful about, which may not necessarily be thrown up through the validator and often come to the fore when we start to relate SDTM and ADaM or SDTM and CRF for traceability.

.

CONTACT INFORMATION

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

Author 1: Usha Kumar
Company: inVentiv Health Clinical
Address: Commerzone, Building No. 4, Floor No. 6, Yerwada Jail Road
City/Postcode: Pune - 411006
Work Phone: +91-02030569112
Email: usha_cool@hotmail.com

Author 2: Neha Mohan
Company: inVentiv Health Clinical
Address: 7th Floor Marwah Centre, Krishanlal Marwah Marg, Andheri (E)
City/Postcode: Mumbai - 400 072
Work Phone: +91-22-40957374
Email: neha.nm@gmail.com

Author 3: Gayatri Karkera
Company: inVentiv Health Clinical
Address: 7th Floor Marwah Centre, Krishanlal Marwah Marg, Andheri (E)
City/Postcode: Mumbai - 400 072
Work Phone: +91-22-40957367
Email: gayatri.karkera@gmail.com