Errors beyond OpenCDISC Validator

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

For FDA submissions we usually run our SDTM and ADaM datasets through the OpenCDISC validator in order to ensure compliance with CDISC standards. The datasets may individually adhere to the standards but would they guarantee traceability from ADaM to SDTM? In this poster we provide some examples 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 for SDTM and CRF or traceability

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. It is however just a supportive tool in ensuring conformance to the data standards in terms of the structure and format; it is not to be looked at for ensuring complete data integrity. The following list of some key checks illustrates the scope of the tool

CDISC Compliance Checks	
Matching of variable's data against pre-defined set of values	✔
Consistency for traceability (Data format)	X
Imputations of values	X
Uniqueness of variable's value across all records	✔
Replication of information	X
Conditional check between variable's values	✔
Check on specific value for a variable when pre-defined conditions are met	✔
Consistency for traceability (Re-derivations)	X
Cross-data checks across multiple domains	✔
Existence of dummy data	X
Correct mapping to CODELIST	X
Consistency of values	X
Complete CRF data is captured in SDTM	X
Redundant NULL variables	X
Meta-data checks	✔

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'

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

CONSISTENCY OF DATA FORMAT FOR TRACEABILITY

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

CONSISTENCY FOR TRACEABILITY (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

From the above example, we see that the CMSEQ value in the SDTM and ADaM have got interchanged. SDTM follows SEQ creation by sort order of medication term 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

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

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

CONSISTENCY 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

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

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

Conclusions and Recommendations

The examples covered in this paper are not an exhaustive list of the issue that we need to be careful about which are not going to be captured by the OpenCDISC validator. The basis behind this paper is to just highlight that we need to be aware of what OpenCDISC validator is designed to capture and hence use it as just a supporting tool for validating SDTM/ADaM

Contact information

Usha Kumar: usha_cool@hotmail.com ; Neha Mohan: neha.nm@gmail.com ;
Gayatri Karkera: gayatri.karkera@gmail.com