



Enhancing the quality of data harmonisation – a customisable tool for validation of bespoke SDTM implementation at the Infectious Diseases Data Observatory (IDDO)

Dick Hung^{1,2}, Sadie Kelly^{1,2}, Kalynn Kennon^{1,2}

¹Infectious Diseases Data Observatory (IDDO), Oxford, UK ²Centre for Tropical Medicine and Global Health, Nuffield Department of Clinical Medicine, University of Oxford, Oxford, UK

1. Background

- To validate IDDO-specific-customised rules not covered by Pinnacle 21 (P21), a customised-rule-based error checking tool has been proposed. Customised rules are used to check individual SDTM domains and values between different domains, issuing error alerts for review.
- The aim of the tool implementation is to improve the accuracy and the efficiency of the standardisation process at IDDO by reducing the amount of manual checking required and reducing the number of errors found by quality checks performed post-curation prior to data release.

2. Input

- Depending on the specific domain and rule checking, files containing weight unit/date conversion formulae, external rules logic are required in addition to the curated SDTM domain file and rules file. A parameter file drives the dynamic customisation behaviour for the system. Users can perform checks on different SDTM domains with different rule files without changing the R script.

3. Process

- **Multiple conditional checks** entail the capacity to conduct checks on a particular subset of records that conform to one or more prescribed conditions. For example, records may be selected that are solely associated with the male gender, are within a designated age bracket, and are undergoing a specific drug regimen. This approach expedites the checking process and enables the user to concentrate their attention on those records that are of particular interest. As sample data, illustrated in Tables 1&2, the USUBJID A_00020 where SEX in DM = M and the subject has data in the RP domain.

USUBJID	SEX
A_00010	F
A_00011	M
A_00020	M

Table 1

USUBJID	RPTEST	RPORRES
A_00010	Pregnant Indicator	Y
A_00020	Pregnant Indicator	Y

Table 2

VSTEST	VSORRES	VSORRESU	VSSTRESN	VSSTRESU
Weight	75	Kg	75	kg
Weight	75	Kg	7.5	kg
Weight	4200	G	4200	g
Weight	140	LB	280	kg

Table 3

- **Calculated conversion checks** to highlight any errors in the standardised--STRES variables for measurements such as weight or temperature.Example shown in Table 3, all error records highlighted. Only the top row has corrected conversions, others are incorrect – row 2 incorrect value for standardised result VSSTRESN, row 3 not standardised (original value and units), row 4 incorrect calculation in VSSTRESN.
- **IDDO-customized controlled terminology checks** confirm the correct use of malaria species terminology that are not currently in SDTM controlled terminology. Details as shown as Table 4.

MBTESTCD	MBTEST
PFALCIPA	Plasmodium falciparum, Asexual
PMALARS	Plasmodium malariae, Sexual
PLSMDMA	Plasmodium, Asexual
PFALCIPA	Plasmodium falciparum
PLMSDM	Plasmodium, Asexual

Table 4

4. Output

- A P21-like format

No	A	B	C	D	E	F
1	Details					
2	1 ODDI P21 Validation Report					
3	2 Rules file: D:/R/SK/Cross_rules_2_SK_v2.csv					
4	3 Data file: D:/R/SK/DM_test.csv					
5	4 Data file: D:/R/SK/DS_test.csv					
6	5 Data file: D:/R/SK/SA_test.csv					
7	6 Generate Date: 07-06-2023					
8	7 Software Version: 7					
9	8 Total data size for DM Domain: 4					
10	9 Total size in batch DM Domain: 1					
11	10 Total data size for DS Domain: 6					
12	11 Total size in batch DS Domain: 1					
13	12 Total data size for SA Domain: 7					
14	13 Total size in batch SA Domain: 1					
15	14					
16	15					
17	16 chunk_size larger than record size, chunk_size has been updated to: 1					

Validation Summary

A	B	C	D	E
1 DOMAIN	RuleNo	Group.3	Rejected Records	
2 DM	1		2	
3 DS	2		2	
4 DS	3		1	
5 SA	4		2	
6 SA	5		1	
7				
8				
9 total:			8	

Dataset Summary

Processing batch: 10 record #: 9001 to 10000 at: 07-06-2023 19:09:50
Processing batch: 11 record #: 10001 to 11000 at: 07-06-2023 19:09:57
Processing batch: 12 record #: 11001 to 12000 at: 07-06-2023 19:09:58
Processing batch: 13 record #: 12001 to 13000 at: 07-06-2023 19:09:59
Processing batch: 14 record #: 13001 to 14000 at: 07-06-2023 19:10:01
Processing batch: 15 record #: 14001 to 15000 at: 07-06-2023 19:10:02
Processing batch: 16 record #: 15001 to 16000 at: 07-06-2023 19:10:04
Processing batch: 17 record #: 16001 to 17000 at: 07-06-2023 19:10:05
Processing batch: 18 record #: 17001 to 18000 at: 07-06-2023 19:10:07
Processing batch: 19 record #: 18001 to 19000 at: 07-06-2023 19:10:08

Log file with timestamp

No	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Domain	TESTCD	Target	RuleNo	DataLine	RefNo	Desc	STD_Value	STD_Unit	RuleRef	DataRef	
2	1	DM		DTHDTC	1	2		DTHDTC (DM)			1, DM,, DTHDABCDE, DM, ABCDE_2, 06/06		
3	2	DM		DTHDTC	1	3		DTHDTC (DM)			1, DM,, DTHDABCDE, DM, ABCDE_3, 06/06		
4	3	DS		DSSTDTC	2	4		DSSTDTC (DS)			2, DS,, DSSTD ABCDE, DS, ABCDE_2, 2, Died		
5	4	DS		DSSTDTC	2	5		DSSTDTC (DS)			2, DS,, DSSTD ABCDE, DS, ABCDE_3, 1, Dece		
6	5	DS		DSSTDTC	3	3		DSSTDTC (DS)			3, DS,, DSSTD ABCDE, DS, ABCDE_2, 1, Com		
7	6	SA		SASTDTC	4	4	C_1	SASTDTC (SA)			4, SA,, SASTD ABCDE, SA, ABCDE_2, Fever,		
8	7	SA		SASTDTC	4	5	C_1	SASTDTC (SA)			4, SA,, SASTD ABCDE, SA, ABCDE_2, Head		
	8	SA		SASTDTC	5	7	C_2	SASTDTC (SA)			5, SA,, SASTD ABCDE, SA, ABCDE_3, Head		

Checking results

5. Conclusion

- The customised tool augments the SDTM-compliance checking capabilities of P21 whilst providing a flexible platform for further checks against IDDO’s implementation of SDTM.
- A batch processing mode with a scheduler that can be executed automatically at a specified period can be developed. Rules operators can be extended to allow for production of summary statistics.



 dick.hung@iddo.org
 info@iddo.org
 [@IDDOnews](https://twitter.com/IDDOnews)
 iddo.org