



Two macros for dataset validation

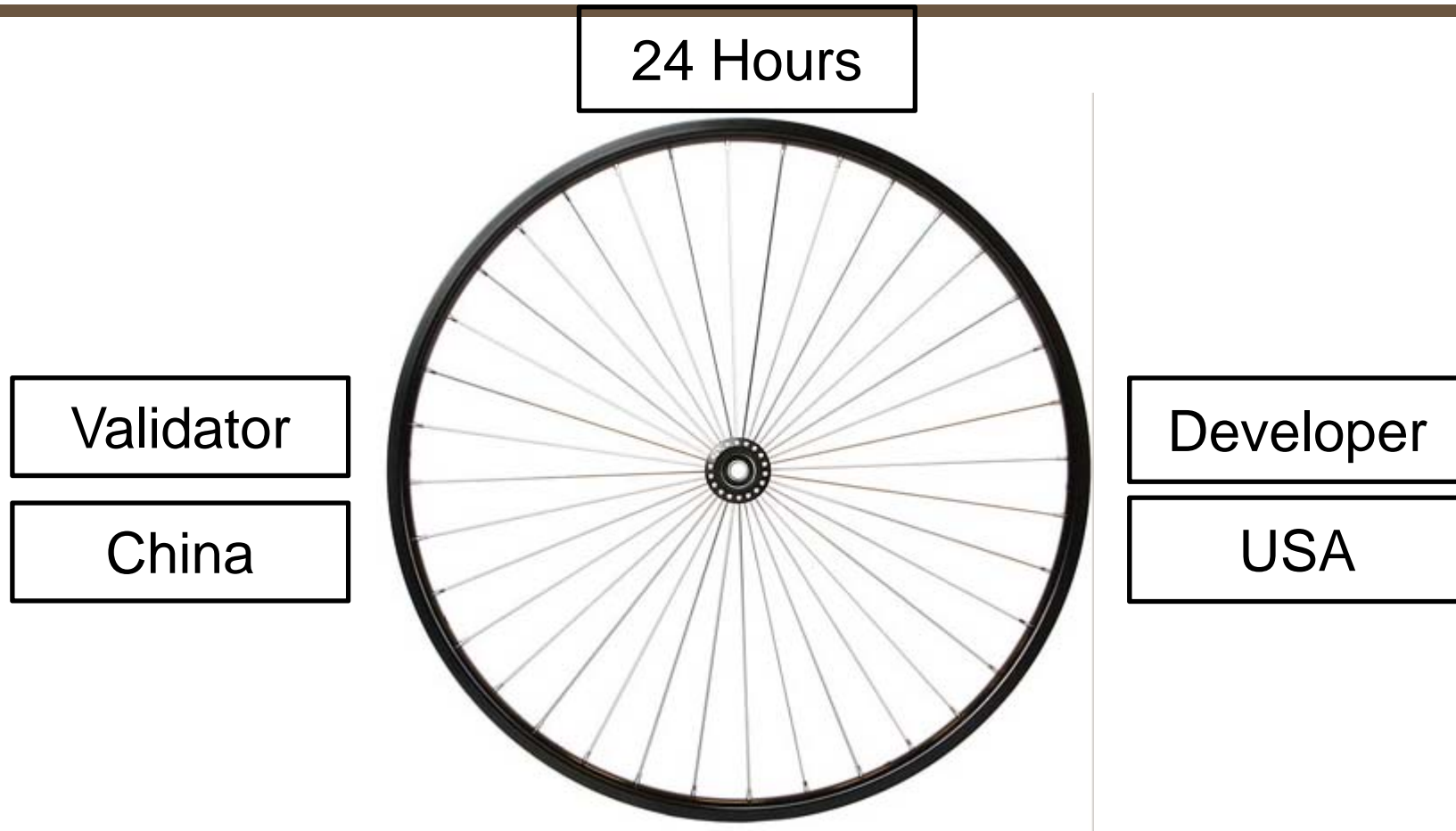
Chunpeng Zhao

2015-08-17

International Team



Scenario



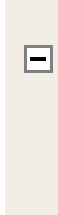
- What to do for Validator to improve validation process?

Requirements to a good validator

- Quickly locate programming issues
- Clearly explain issues to developer
- Raise as many issues as possible at a time
- Never just display discrepancies without any explanations

Challenges in Validation to datasets

- Difficult to locate unmatched records
 - ID variable value is too long
 - ID variables too many



```
proc compare base=derived.dae compare=dae_26;
  id USUBJID EPOCH AETERM AESTDTC;
run;
```

WARNING: The LINESIZE=120 option value is too small to print all of the ID variables in the Value Comparison reports. Only the first 3 ID variables will be printed. The following ID variables will not be printed: AESTDTC.

Value Comparison Results for Variables

USUBJ ID	EPOCH	AETERM	Outcome of Adverse Event Base Value AEOUT	Compare Value AEOUT
1002008	RANDOMIZED TREATMENT	Right leg cramps	RECOVERED/RESOLVED	DURING
1711002	RANDOMIZED TREATMENT	Pneumonia	RECOVERED/RESOLVED	DURING
1711002	RANDOMIZED TREATMENT	pleura effusion	RECOVERED/RESOLVED W	DURING
1711002	RANDOMIZED TREATMENT	pneumo thorax	RECOVERED/RESOLVED W	DURING
2363002	SCREENING	Lung cancer	FATAL	DURING
3050020	SCREENING	Morbus Basedow	RECOVERED/RESOLVED	DURING
5139002	RANDOMIZED TREATMENT	Nose Bleed	RECOVERED/RESOLVED	DURING

Challenges in Validation to datasets

- Not easy to review unmatched values
 - Unmatched values are too long to completely display

USUBJ ID	Date/Time of Birth Base BRTHDTN	Compare BRTHDTN	Diff.	% Diff
.1002014	*****	*****	189302400	-32.5654
.1253023	*****	*****	-4.1394E8	121.1378
.1848003	*****	*****		
.2050025	*****	*****		
.2363001	*****	*****		
.2363002	*****	*****		
.2609003	*****	*****	-43113600	7.3189
.2615032	*****	*****	157766400	321.4789

USUBJID	Ethnicity Base Value ETHNIC	Compare Value ETHNIC
.1253022	OTHER	UNKNOWN
.2224004	OTHER	HISPANIC OR LATINO
.3552012	NOT HISPANIC OR LATI	OTHER

What do we want for compare results?

- Only focus on unmatched records and unmatched variables
 - Completely matched records should be removed
 - Completely matched variables should be removed
- Can review unmatched records and unmatched variables in **context of clinical data**
- Easy to describe QC issues to developer

ID Variables

Unmatched Variables

USUBJID	EPOCH	AETERM	AESTDTC	_DSNAME	AEENRF	AEOUT
_1002008	RANDOMIZED TREATMENT	Right leg cramps	2014-10-06T17:00	BASE	DURING	RECOVERED/RESOLVED
_1002008	RANDOMIZED TREATMENT	Right leg cramps	2014-10-06T17:00	COMPARE		DURING
_1711002	RANDOMIZED TREATMENT	Pneumonia	2014-03-31	BASE	DURING	RECOVERED/RESOLVED
_1711002	RANDOMIZED TREATMENT	Pneumonia	2014-03-31	COMPARE		DURING
_1711002	RANDOMIZED TREATMENT	pleura effusion	2014-03-31	BASE	DURING	RECOVERED/RESOLVED WITH SEQUELAE
_1711002	RANDOMIZED TREATMENT	pleura effusion	2014-03-31	COMPARE		DURING
_1711002	RANDOMIZED TREATMENT	pneumo thorax	2014-04-03	BASE	DURING	RECOVERED/RESOLVED WITH SEQUELAE
_1711002	RANDOMIZED TREATMENT	pneumo thorax	2014-04-03	COMPARE		DURING
_2363002	SCREENING	Lung cancer	2015-04-09	BASE	DURING	FATAL
_2363002	SCREENING	Lung cancer	2015-04-09	COMPARE		DURING

Vertical View to QC issues

List all records with unmatched specified variables

```
%let keepvar=LABTP SUBJID AVISITN ADTM PARAMCD CRIT1 CRIT2 ;  
  
%cmp_ds(BASE=analysis.adlb(where=(SUBJID="2611020")keep=&keepvar),  
        COMPARE=adlb_qc(where=(SUBJID="2611020") keep=&keepvar),  
        ORDER=%str(LABTP, SUBJID, AVISITN, ADTM, PARAMCD),  
        OUTDSN=qc_rst);
```

ID Variables

Unmatched Variables

SUBJID	AVISITN	ADTM	PARAMCD	_DSNAME	CRIT1	CRIT2
2611020	104	1717687800	CREATUS	BASE	>= 0.3 mg/dL increas	>= 0.5 mg/dL increas
2611020	104	1717687800	CREATUS	COMPARE	>=0.3 mg/dL increase	>=0.5 mg/dL increase
2611020	105	1717774200	CREATUS	BASE	>= 0.3 mg/dL increas	>= 0.5 mg/dL increas
2611020	105	1717774200	CREATUS	COMPARE	>=0.3 mg/dL increase	>=0.5 mg/dL increase
2611020	106	1717860600	CREATUS	BASE	>= 0.3 mg/dL increas	>= 0.5 mg/dL increas
2611020	106	1717860600	CREATUS	COMPARE	>=0.3 mg/dL increase	>=0.5 mg/dL increase
2611020	107	1717947000	CREATUS	BASE	>= 0.3 mg/dL increas	>= 0.5 mg/dL increas
2611020	107	1717947000	CREATUS	COMPARE	>=0.3 mg/dL increase	>=0.5 mg/dL increase
2611020	108	1718033400	CREATUS	BASE	>= 0.3 mg/dL increas	>= 0.5 mg/dL increas
2611020	108	1718033400	CREATUS	COMPARE	>=0.3 mg/dL increase	>=0.5 mg/dL increase
2611020	109	1718787600	CREATUS	BASE	>= 0.3 mg/dL increas	>= 0.5 mg/dL increas
2611020	109	1718787600	CREATUS	COMPARE	>=0.3 mg/dL increase	>=0.5 mg/dL increase
2611020	29001	1718100000	CREATUS	BASE	>= 0.3 mg/dL increas	>= 0.5 mg/dL increas
2611020	29001	1718100000	CREATUS	COMPARE	>=0.3 mg/dL increase	>=0.5 mg/dL increase

Horizontal View to QC issues

List all unmatched variables within minimum records

```
%cmp_ds_v2(BASE=analysis.adlb(where=(SUBJID="2611020")),  
COMPARE=adlb_qc(where=(SUBJID="2611020")),  
ORDER=%str(LABTP, SUBJID, AVISITN, ADTM, PARAMCD, ANRHIN),  
OUTDSN=qc_rst_1);
```

ID Variables

Unmatched Variables

SUBJID	AVISI	ADTM	PARAMCD	ANRHIN	_DSNAME	ADY	ANRHI	ANRLO	CRIT1	CRIT2
2611020	100	1717599600	PLATSI	370	BASE	1	370.0	140.		
2611020	100	1717599600	PLATSI	370	COMPARE	1	370.0	140.		
2611020	109	1718787600	CREATUS	1.16493	BASE	15	1.2	0.7	>= 0.3 mg/dL increas	>= 0.5 mg/dL increas
2611020	109	1718787600	CREATUS	1.16493	COMPARE	14	1.2	0.7	>=0.3 mg/dL increase	>=0.5 mg/dL increase

Summary of two macros for dataset validation

- Vertical view
 - List all records with unmatched specified variables
- Horizontal View
 - List all unmatched variables within minimum records
- Limitation:
 - Not replace SAS compare procedure
 - Not check records with unmatched ID
 - Not check variable attributes
- Totally free: Able to download from PharmaSUG website

SAS macros



cmp_ds.sas



cmp_ds_v2.sas