

A way out of the PROC COMPARE labyrinth

Judith Kanz

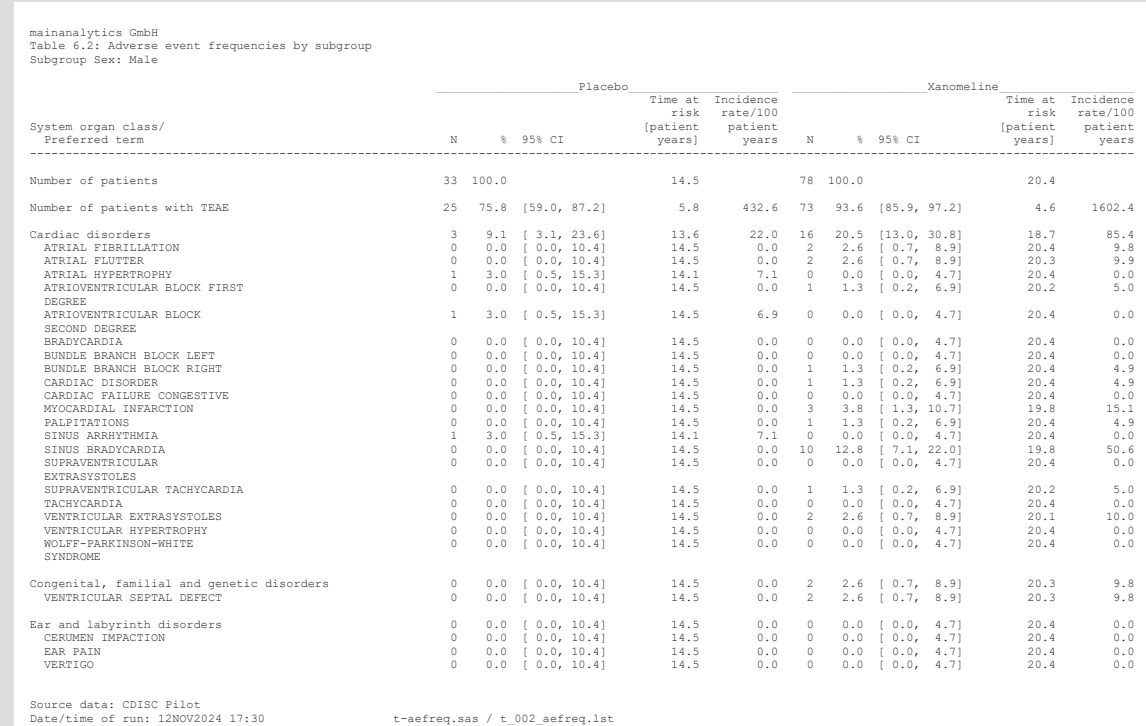
Principal Statistical Programmer
mainanalytics GmbH
judith.kanz@mainanalytics.de

Validation per double programming is the usual method with comparison done via PROC COMPARE.

The output might be overwhelming. Here is guidance on how to untangle the output. A macro can help to tidy up.

Validation steps

- Log file check
- Output review
- Source code review of reporting procedure
- Double programming of input dataset



```
proc report data=dt1f.ouid nowd headline headskip missing split="*" spacing=2;
  column subgroupn subgroup order aesoc aesocd text treat, (freq perc ci trisk rate) dummy;
  define subgroupn / group id order=data noprint;
  define subgroup / group id order=data noprint;
  define order / group id order=data noprint;
  define aesoc / group id order=data noprint;
  define aesocd / group id order=data noprint;
  define text / group id order=data width=55 flow spacing=0;
  define treat / " " order=data across;
  define freq / display;
  define perc / display;
  define trisk / display;
  define rate / display;
  define dummy / noprint;
  break after aesoc / skip;
  compute before _page_;
  line @1 subgroup $50.;
endcomp;
run;
```

Supporting Macro

- Focus PROC COMPARE output to what is really needed
- Create temporary datasets for detailed investigations
 - First dataset: ID variables only plus flag variables for origin and discrepancies
 - Second dataset: As first dataset, but with all other variables with prefixes PROD_ and QC_ side by side
- Create log file messages

Macro QC_COMPARE		
Parameter	Default	Description
OUTID		Output ID for TLF comparison
PRODQS		ADS name for ADS comparison
QCQS	out	Name of QC dataset
KEY	_none_	Variable names for PROC COMPARE ID (also possible: _all_)
KEEP	_all_	Variable names taken into account for PROC COMPARE (also possible: _all_)
DROP	_none_	Variable names removed for PROC COMPARE (also possible: _none_)
INCDENT	_none_	Variable names for comparison of indentations (also possible: _all / _none_)
COMPRESS	_none_	Variable names to be compressed before PROC COMPARE (also possible: _all / _none_)
CRITERION	0.00001	Value of CRITERION option used for PROC COMPARE
LABELS	_all_	Variable names for comparison of labels (also possible: _all / _none_)
FORMATS	_all_	Variable names for comparison of formats (also possible: _all / _none_)
MODIFY		Dataset statements for PROCQS modifications

```
ERROR: output /data/projects/sandbox/sandbox_1/05-Proc-Env/kanz/ouid1/t_002_aesreq.list does not exist
ERROR: QC of DTSLF_T_002_AESREQ - Dataset DTSLF_T_002_AESREQ does not exist

WARNING: QC of DTSLF_T_002_AESREQ - variable DUMMY in PROD dataset only
WARNING: QC of DTSLF_T_002_AESREQ - variable ORDER in PROD dataset only
WARNING: QC of DTSLF_T_002_AESREQ - variable PERC with conflicting types
WARNING: QC of DTSLF_T_002_AESREQ - variable CI with different labels
WARNING: QC of DTSLF_T_002_AESREQ - variable PERC with different labels
WARNING: QC of DTSLF_T_002_AESREQ - variable RATE with different labels
WARNING: QC of DTSLF_T_002_AESREQ - variable TRISK with different labels
WARNING: QC of DTSLF_T_002_AESREQ - values for key variables differ
WARNING: QC of DTSLF_T_002_AESREQ - values differ
```

Untangle PROC COMPARE outputs

1

Listing of Variables in WORK.MCCOMP_PRODQS but not in WORK.MCCOMP_QCQS

Variable	Type	Length
ORDER	Num	8
DUMMY	Num	8

Check in output procedure which variables are used and restrict to them with macro parameters KEEP or DROP

2

Listing of Common Variables with Conflicting Types

Variable	Dataset	Type	Length	Format	Label
PERC	WORK.MCCOMP_PRODQS	Num	8	5.1	%
	WORK.MCCOMP_QCQS	Char	5		

Values cannot be compared, QC dataset needs to be updated

3

Listing of Common Variables with Differing Attributes

Variable	Dataset	Type	Length	Format	Label
SUBGROUP	WORK.MCCOMP_PRODQS	Char	45		
TREAT	WORK.MCCOMP_QCQS	Char	50		Treatment
TRISK	WORK.MCCOMP_PRODQS	Char	100		
CI	WORK.MCCOMP_PRODQS	Num	8	8.1	Time at risk [patient years]
FREQ	WORK.MCCOMP_QCQS	Char	15		the sum, RISKYEARS
RATE	WORK.MCCOMP_PRODQS	Num	8	3.	95% CI
TEXT	WORK.MCCOMP_QCQS	Num	8	9.1	Incidence rate/100 patient years
	WORK.MCCOMP_QCQS	Num	8		
	WORK.MCCOMP_PRODQS	Char	200		System organ class/5 Preferred term
	WORK.MCCOMP_QCQS	Char	250		System organ class/5 Preferred term

Check in output procedure which labels and formats are used and restrict to them with macro parameters FORMATS and LABELS, use larger lengths than in PROD dataset where truncation may be possible

4

Comparison Results for Observations

Observation 25 in WORK.MCCOMP_QCQS not found in WORK.MCCOMP_PRODQS:
SUBGROUP=11 TEXT= APPLICATION SITE DESQUAMATION TREAT= Placebo_

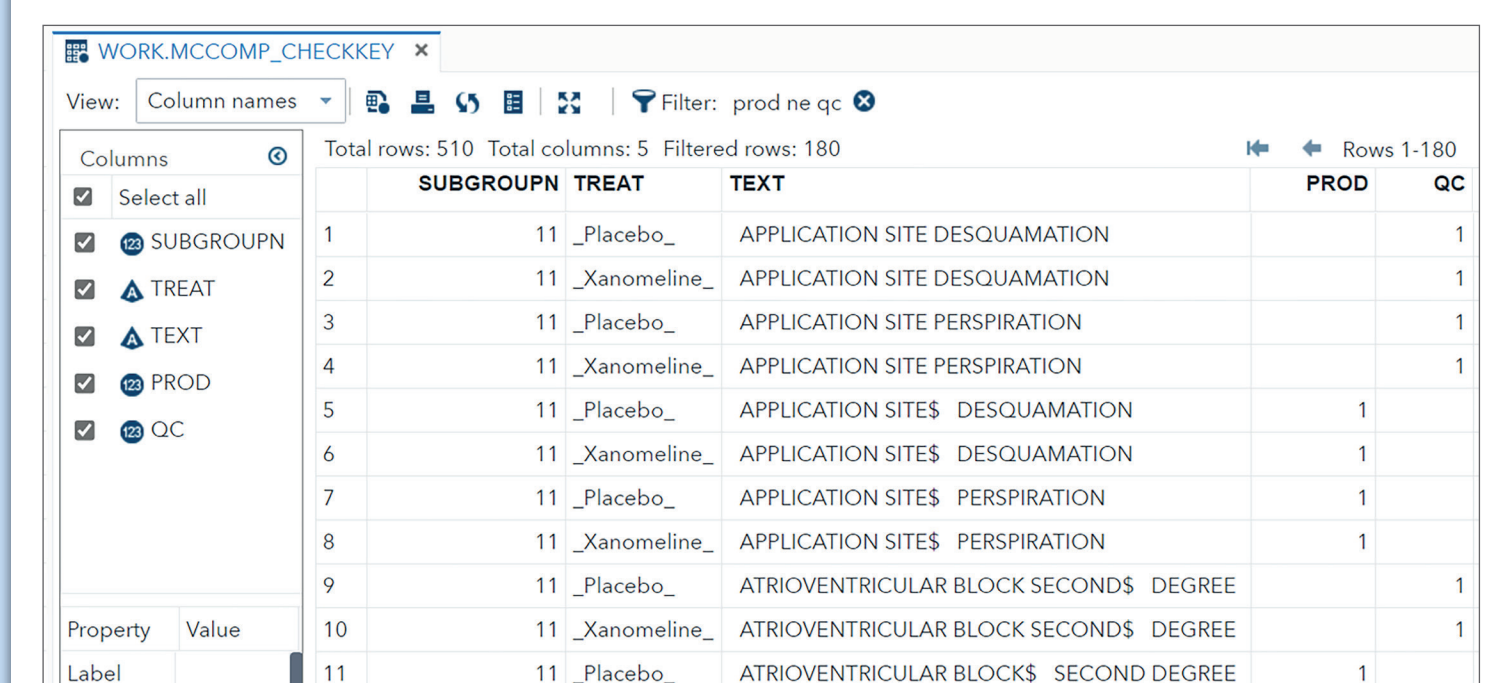
Observation 26 in WORK.MCCOMP_QCQS not found in WORK.MCCOMP_PRODQS:
SUBGROUP=11 TEXT= APPLICATION SITE DESQUAMATION TREAT= Xanomeline_

Observation 37 in WORK.MCCOMP_QCQS not found in WORK.MCCOMP_PRODQS:
SUBGROUP=11 TEXT= APPLICATION SITE PERSPIRATION TREAT= Placebo_

Observation 38 in WORK.MCCOMP_QCQS not found in WORK.MCCOMP_PRODQS:
SUBGROUP=11 TEXT= APPLICATION SITE PERSPIRATION TREAT= Xanomeline_

Observation 47 in WORK.MCCOMP_PRODQS not found in WORK.MCCOMP_QCQS:
SUBGROUP=11 TEXT= APPLICATION SITE DESQUAMATION TREAT= Placebo_

Use first temporary dataset to get a better overview about non-matching ID variables



Consider to use macro parameter MODIFY to remove split characters and check appearance per output review

7

All Variables Compared Have Unequal Values

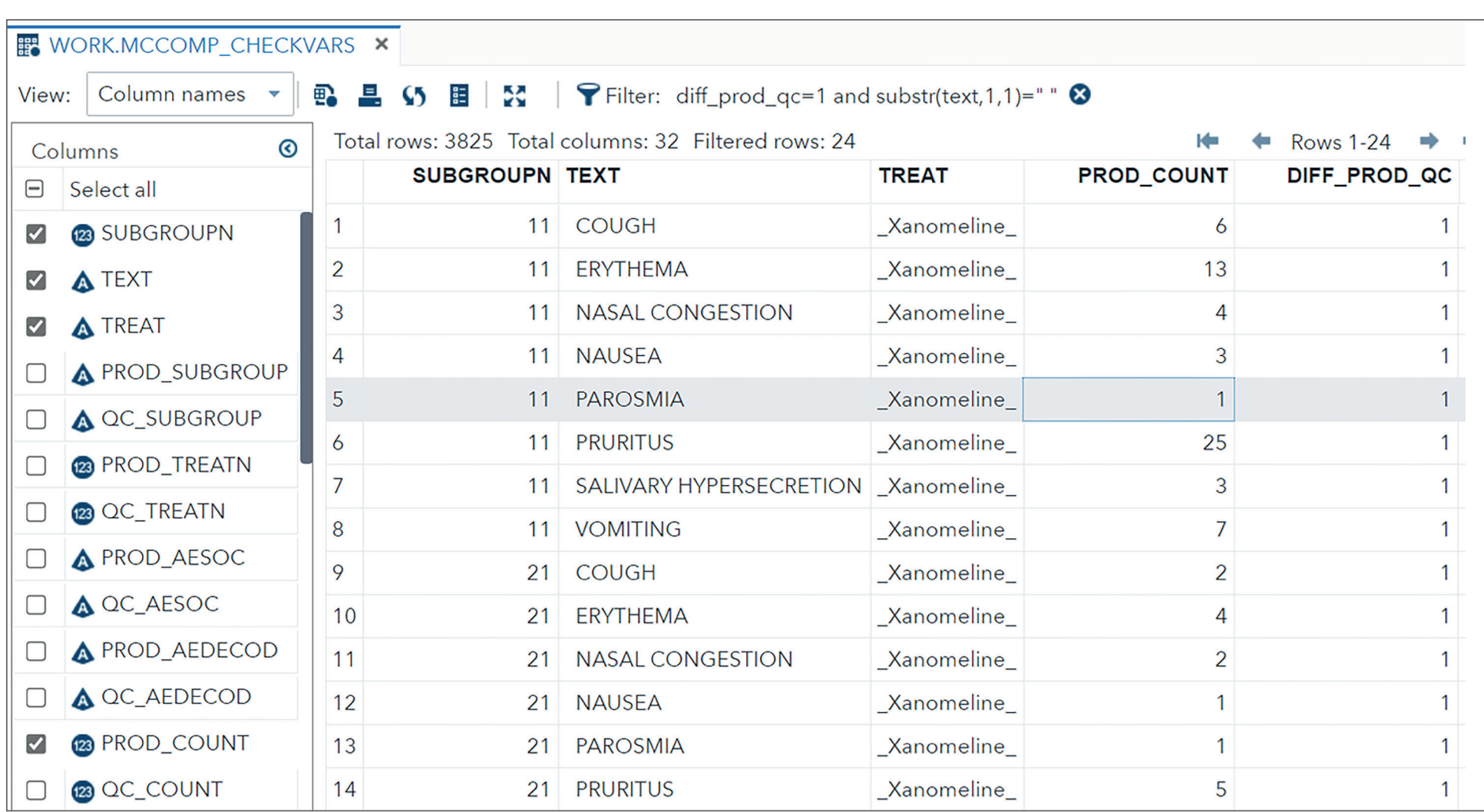
Variable	Type	Len	Label	Compare Label	Ndif	MaxDif
TRISK	NUM	8	Time at risk [patient years]	the sum, RISKYEARS	39	0.0082
RATE	NUM	8	Incidence rate/100 patient years		39	1565

Check for dependencies between variables

Value Comparison Results for Variables

SUBGROUPN	TEXT	TREAT	Base TRISK	Compare TRISK	Diff.	% Diff
11	COUGH	_Xanomeline_	19.2	19.2498	0.008214	0.0427
11	ERYTHEMA	_Xanomeline_	16.1	16.1205	0.008214	0.0510
11	NASAL CONGESTION	_Xanomeline_	19.3	19.3155	0.008214	0.0425
11	NAUSEA	_Xanomeline_	19.9	19.9042	0.008214	0.0413
11	PAROSMIA	_Xanomeline_	20.2	20.1725	0.008214	0.0407
11	PRURITUS	_Xanomeline_	15.5	15.4935	0.008214	0.0530
11	SALIVARY HYPERSECR	_Xanomeline_	19.7	19.6660	0.008214	0.0418
11	VOMITING	_Xanomeline_	18.9	18.9131	0.008214	0.0434
11	Gastrointestinal dis	_Xanomeline_	16.4	16.3888	0.008214	0.0501
11	Nervous system disor	_Xanomeline_	16.1	16.0903	0.008214	0.0511
11	Number of patients w	_Xanomeline_	4.6	4.5640	0.008214	0.1803
11	Respiratory, thoraci	_Xanomeline_	18.1	18.1410	0.008214	0.0453

Use second temporary dataset to find a start for manual check on patient level



6

Value Comparison Results for Variables

SUBGROUPN	TEXT	TREAT	Base Value CI	Compare Value CI
11	ABDOMINAL DISCOMFO	_Placebo_	[0.0, 10.4]	[0.0, 10.4]
11	ABDOMINAL DISCOMFO	_Xanomeline_	[0.2, 6.9]	[0.2, 6.9]
11	ABDOMINAL PAIN	_Placebo_	[0.0, 10.4]	[0.0, 10.4]
11	ABDOMINAL PAIN	_Xanomeline_	[0.2, 6.9]	[0.2, 6.9]
11	ACROCHORDON EXCISI	_Placebo_	[0.0, 10.4]	[0.0, 10.4]
11	ACROCHORDON EXCISI	_Xanomeline_	[0.0, 4.7]	[0.0, 4.7]
11	ACTINIC KERATOSIS	_Placebo_	[0.0, 10.4]	[0.0, 10.4]
11	ACTINIC KERATOSIS	_Xanomeline_	[0.2, 6.9]	[0.2, 6.9]
11	AGITATION	_Placebo_	[0.0, 10.4]	[0.0, 10.4]
11	AGITATION	_Xanomeline_	[0.0, 4.7]	[0.0, 4.7]

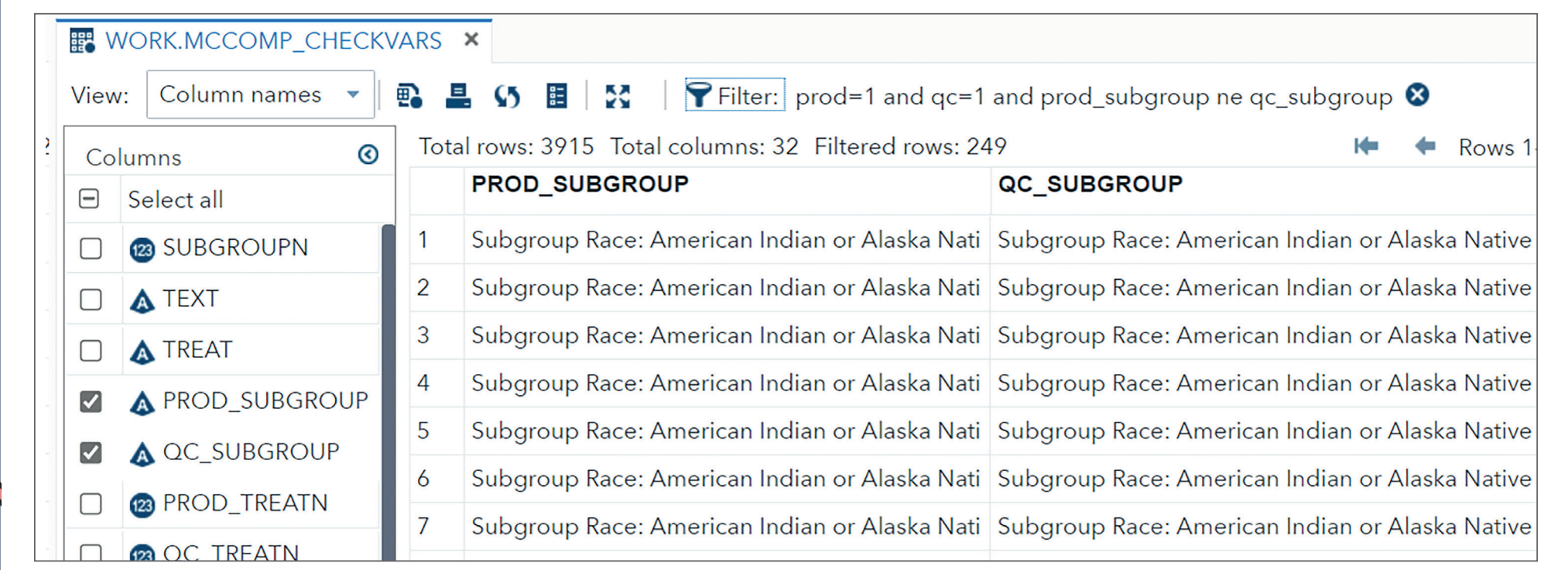
Use macro parameter COMPRESS to compare values only and check appearance per output review, enables use of larger formats than in PROD dataset

5

Value Comparison Results for Variables

SUBGROUPN	TEXT	TREAT	Base Value SUBGROUP	Compare Value SUBGROUP
36	ABDOMINAL DISCOMFO	_Xanomeline_	Subgroup Race: Ameri	Subgroup Race: Ameri
36	ABDOMINAL PAIN	_Xanomeline_	Subgroup Race: Ameri	Subgroup Race: Ameri
36	ACROCHORDON EXCISI	_Xanomeline_	Subgroup Race: Ameri	Subgroup Race: Ameri
36	ACTINIC KERATOSIS	_Xanomeline_	Subgroup Race: Ameri	Subgroup Race: Ameri
36	AGITATION	_Xanomeline_	Subgroup Race: Ameri	Subgroup Race: Ameri
36	AGITATION	_Xanomeline_	Subgroup Race: Ameri	Subgroup Race: Ameri

Use second temporary dataset to check for discrepancies that are not visible in PROC COMPARE output



Conclusion

You can ease your daily work. The macro cannot provide a solution to get rid of all discrepancies, but it can at least help you to get closer to finding a pattern or a hint for the source of mismatches and can speed things up.

Scan QR Code for paper

