main analytics

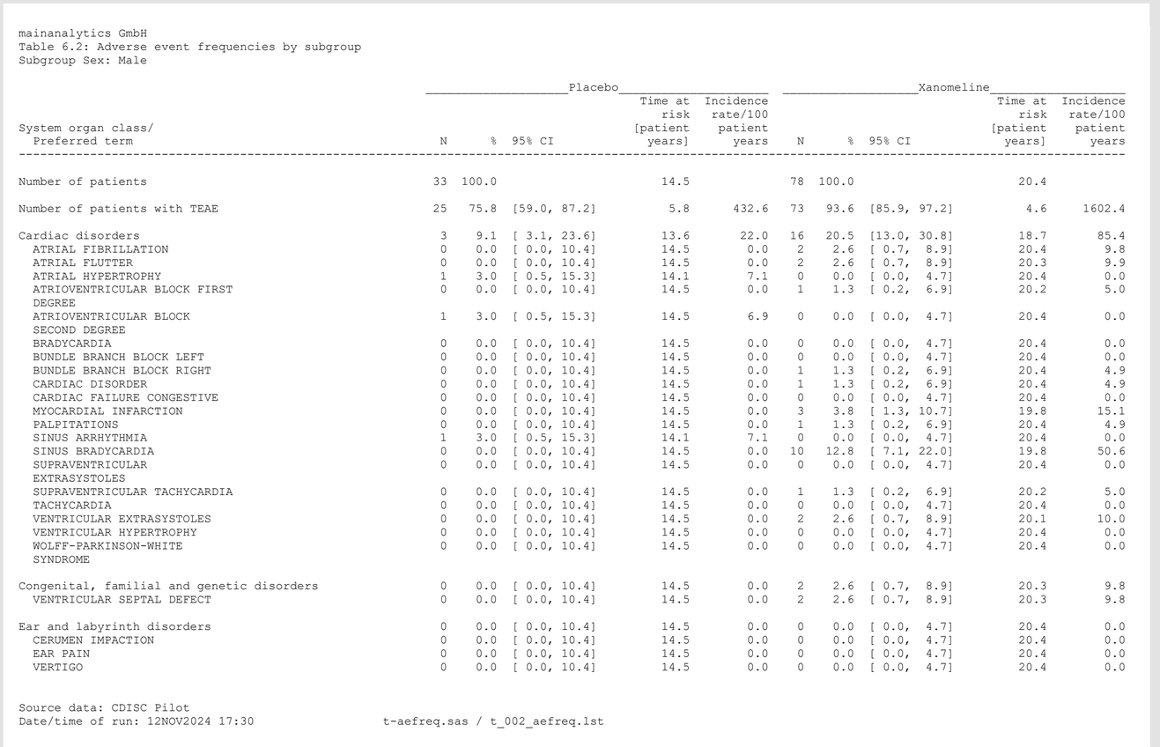
A way out of the PROC COMPARE labyrinth

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=dstlf.&outid nowd headline headskip missing split="$" spacing=2;
  column subgroupn subgroup order aesoc aedecod 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 aedecod / 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;
```

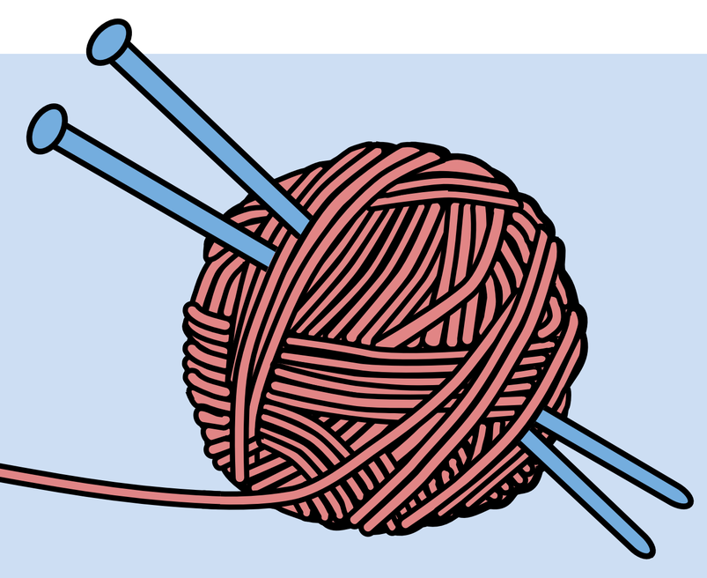
Supporting Macro

- 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
PRODDS		ADS name for ADS comparison
QCDS	out	Name of QC dataset
KEY	_none_	Variable names for PROC COMPARE ID (also possible: _none_)
KEEP	_all_	Variable names taken into account for PROC COMPARE (also possible: _all_)
DROP	_none_	Variable names removed for PROC COMPARE (also possible: _none_)
INDENT	_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	0.000001	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 PRODDS modifications

```
ERROR: output /data/projects/sandbox/sandbox_1/05-Prog-Env/jkanz/outtlf/t_002_aefreq.lst does not exist
ERROR: QC of DSTLF.T_002_AEFREQ - Dataset DSTLF.T_002_AEFREQ does not exist

WARNING: QC of DSTLF.T_002_AEFREQ - variable DUMMY in PROD dataset only
WARNING: QC of DSTLF.T_002_AEFREQ - variable ORDER in PROD dataset only

WARNING: QC of DSTLF.T_002_AEFREQ - variable PERC with conflicting types

WARNING: QC of DSTLF.T_002_AEFREQ - variable CI with different labels
WARNING: QC of DSTLF.T_002_AEFREQ - variable FREQ with different labels
WARNING: QC of DSTLF.T_002_AEFREQ - variable PERC with different labels
WARNING: QC of DSTLF.T_002_AEFREQ - variable RATE with different labels
WARNING: QC of DSTLF.T_002_AEFREQ - variable TREAT with different labels
WARNING: QC of DSTLF.T_002_AEFREQ - variable TRISK with different labels

WARNING: QC of DSTLF.T_002_AEFREQ - values for key variables differ

WARNING: QC of DSTLF.T_002_AEFREQ - values differ
```



Untangle PROC COMPARE outputs

1

```
Listing of Variables in WORK.MCCOMP_PRODDS but not in WORK.MCCOMP_QCDS

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_PRODDS  Num        8  5.1     %
          WORK.MCCOMP_QCDS    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_PRODDS  Char      45
          WORK.MCCOMP_QCDS    Char      50
TREAT     WORK.MCCOMP_PRODDS  Char     100          Treatment
          WORK.MCCOMP_QCDS    Char     100
TRISK     WORK.MCCOMP_PRODDS  Num        8  8.1     Time at risk [patient years]
          WORK.MCCOMP_QCDS    Num        8          the sum, RISKYEARS
CI        WORK.MCCOMP_PRODDS  Char      15          95% CI
          WORK.MCCOMP_QCDS    Char      15
FREQ      WORK.MCCOMP_PRODDS  Num        8  3.      N
          WORK.MCCOMP_QCDS    Num        8
RATE      WORK.MCCOMP_PRODDS  Num        8  9.1     Incidence rate/100 patient years
          WORK.MCCOMP_QCDS    Num        8
TEXT      WORK.MCCOMP_PRODDS  Char     200          System organ class/$   Preferred term
          WORK.MCCOMP_QCDS    Char     250          System organ class/$   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_QCDS not found in WORK.MCCOMP_PRODDS:
SUBGROUPN=11 TEXT=  APPLICATION SITE DESQUAMATION TREAT=_Placebo_.

Observation 26 in WORK.MCCOMP_QCDS not found in WORK.MCCOMP_PRODDS:
SUBGROUPN=11 TEXT=  APPLICATION SITE DESQUAMATION TREAT=_Xanomeline_.

Observation 37 in WORK.MCCOMP_QCDS not found in WORK.MCCOMP_PRODDS:
SUBGROUPN=11 TEXT=  APPLICATION SITE PERSPIRATION TREAT=_Placebo_.

Observation 38 in WORK.MCCOMP_QCDS not found in WORK.MCCOMP_PRODDS:
SUBGROUPN=11 TEXT=  APPLICATION SITE PERSPIRATION TREAT=_Xanomeline_.

Observation 47 in WORK.MCCOMP_PRODDS not found in WORK.MCCOMP_QCDS:
SUBGROUPN=11 TEXT=  APPLICATION SITE$   DESQUAMATION TREAT=_Placebo_.
```

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

WORK.MCCOMP_CHECKKEY

View: Column names · Filter: prod ne qc

Total rows: 510 Total columns: 5 Filtered rows: 180 · Rows 1-180

Columns: Select all · SUBGROUPN · TREAT · TEXT · PROD · QC

	SUBGROUPN	TREAT	TEXT	PROD	QC
1	11	_Placebo_	APPLICATION SITE DESQUAMATION		1
2	11	_Xanomeline_	APPLICATION SITE DESQUAMATION		1
3	11	_Placebo_	APPLICATION SITE PERSPIRATION		1
4	11	_Xanomeline_	APPLICATION SITE PERSPIRATION		1
5	11	_Placebo_	APPLICATION SITE$ DESQUAMATION	1	
6	11	_Xanomeline_	APPLICATION SITE$ DESQUAMATION	1	
7	11	_Placebo_	APPLICATION SITE$ PERSPIRATION	1	
8	11	_Xanomeline_	APPLICATION SITE$ PERSPIRATION	1	
9	11	_Placebo_	ATRIOVENTRICULAR BLOCK SECOND$ DEGREE		1
10	11	_Xanomeline_	ATRIOVENTRICULAR BLOCK SECOND$ DEGREE		1
11	11	_Placebo_	ATRIOVENTRICULAR BLOCK$ SECOND DEGREE	1	

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

5

```
Value Comparison Results for Variables

__________________________________________________________________________________
                                               ||  Base Value          Compare Value
SUBGROUPN  TEXT                 TREAT          ||  SUBGROUP            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
```

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

WORK.MCCOMP_CHECKVARS

View: Column names · Filter: prod=1 and qc=1 and prod_subgroup ne qc_subgroup

Total rows: 3915 Total columns: 32 Filtered rows: 249 · Rows 1-

Columns: Select all · SUBGROUPN · TEXT · TREAT · PROD_SUBGROUP · QC_SUBGROUP · PROD_TREATN · QC_TREATN

	PROD_SUBGROUP	QC_SUBGROUP
1	Subgroup Race: American Indian or Alaska Nati	Subgroup Race: American Indian or Alaska Native
2	Subgroup Race: American Indian or Alaska Nati	Subgroup Race: American Indian or Alaska Native
3	Subgroup Race: American Indian or Alaska Nati	Subgroup Race: American Indian or Alaska Native
4	Subgroup Race: American Indian or Alaska Nati	Subgroup Race: American Indian or Alaska Native
5	Subgroup Race: American Indian or Alaska Nati	Subgroup Race: American Indian or Alaska Native
6	Subgroup Race: American Indian or Alaska Nati	Subgroup Race: American Indian or Alaska Native
7	Subgroup Race: American Indian or Alaska Nati	Subgroup Race: American Indian or Alaska Native

6

```
Value Comparison Results for Variables

__________________________________________________________________________
                                               ||  95% CI
                                               ||  Base Value     Compare Value
SUBGROUPN  TEXT                 TREAT          ||  CI              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]
```

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

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

_______________________________________________________________________________________
                                               ||  Time at risk [patient years]
                                               ||  the sum, RISKYEARS
                                               ||      Base    Compare
SUBGROUPN  TEXT                 TREAT          ||     TRISK      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

WORK.MCCOMP_CHECKVARS

View: Column names · Filter: diff_prod_qc=1 and substr(text,1,1)=" "

Total rows: 3825 Total columns: 32 Filtered rows: 24 · Rows 1-24

Columns: Select all · SUBGROUPN · TEXT · TREAT · PROD_SUBGROUP · QC_SUBGROUP · PROD_TREATN · QC_TREATN · PROD_AESOC · QC_AESOC · PROD_AEDECOD · QC_AEDECOD · PROD_COUNT · QC_COUNT

	SUBGROUPN	TEXT	TREAT	PROD_COUNT	DIFF_PROD_QC
1	11	COUGH	_Xanomeline_	6	1
2	11	ERYTHEMA	_Xanomeline_	13	1
3	11	NASAL CONGESTION	_Xanomeline_	4	1
4	11	NAUSEA	_Xanomeline_	3	1
5	11	PAROSMIA	_Xanomeline_	1	1
6	11	PRURITUS	_Xanomeline_	25	1
7	11	SALIVARY HYPERSECRETION	_Xanomeline_	3	1
8	11	VOMITING	_Xanomeline_	7	1
9	21	COUGH	_Xanomeline_	2	1
10	21	ERYTHEMA	_Xanomeline_	4	1
11	21	NASAL CONGESTION	_Xanomeline_	2	1
12	21	NAUSEA	_Xanomeline_	1	1
13	21	PAROSMIA	_Xanomeline_	1	1
14	21	PRURITUS	_Xanomeline_	5	1

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