

A Proc Compare Enhancement macro

+

•

○

Chris Weiner

Bristol-Myers-Squibb Corporation

Purpose

Make the diagnosis of any
proc compare differences
easier and spot other
potential data issues in
the process

Facilitate first validation
draft on new data

Aid implementation of a
new ADaM spec or other
data set spec

Assist in early stages of
initial validation, and
when new differences
emerge in new data
snapshots

Add some more data
checks and to improve
readability of the reports

Advantages of qcompare

Wrapping of columns
and display of full
text strings
highlighting
differences

Reporting of all
duplicates and
analysis of
distinguishing
variables

Comparison of
univariate statistics

Clear and complete
reporting of
duplicates by id
variables

Clearer listing of
unmatched id
variables

Identification of
special characters
(outside ascii 32-
137) in values and
labels.

Reduction of time
with further
investigation of
differences

Finding of general
data issues that may
not be immediately
apparent in proc
compare output

Example Macro Call

```
%qcompare(  
  base = basesub,  
  compare = comparesub,  
  idvars = usubjid paramcd adt,  
  abs_tolerance = .0000001,  
  pct_tolerance = ,  
  idgroups = usubjid | paramcd | usubjid paramcd,  
  dup_base_output_ds = base_dups,  
  dup_compare_output_ds = compare_dups,  
  diff_output_ds = diffs_output,  
  match_output_ds = match,  
  max_obs_to_print = 500  
);
```

Input and
Output

- Report is in HTML format
- The only required macro arguments and input are base = and compare =.
- All others are optional

Parameter Definitions

Parameter Name	Description	Required	Example
Base	Base dataset name	Yes	basesub
Compare	Compare dataset name	Yes	comparesub
Idvars	ID variables	No. Defaults to observation number	Usubjid paramcd adt
Abs_tolerance	Tolerance level to consider two numeric values equal	No ,default 0	.00000001
Pct_tolerance	Percent tolerance level to consider two numeric values equal	No, default 0	could use 1, but not used in our test
Idgroups	Check the agreement of base and compare datasets of these groupings. Any number of groupings can be specified, delimited by .	No	Usubjid paramcd usubjid paramcd
Dup_base_output_ds	Base duplicate output dataset name	No	Base_dups
Dup_compare_output_ds	Compare output dataset name	No	Compare_dups
Diff_output_ds	Differences output dataset name	No	Diff_output
Match_output_ds	Matched output dataset name	No	match
Max_obs_to_print	Maximum observations to print in any one report	No, default 100	500
Outfile	Output Filename	No, default to program name	none used in test, could use compare_outfile

Base Dataset

usubjid	paramcd	adtcmt	ablfl	ittfl	aval	pkfl
STUDY-001-1234567890123	ALB	6/12/2019	Y	Y	4	
STUDY-001-1234567890124	ALB	6/13/2019	Y	Y	5	
STUDY-001-1234567890124	ALB	6/13/2019	Y	Y	5	
STUDY-001-1234567890125	ALB	6/13/2019	Y	Y	5	
STUDY-001-1234567890127	HCT	6/15/2019	Y	Y	5	
STUDY-001-1234567890128	HGB	6/16/2019	Y	Y	7	
STUDY-001-1234567890129	ALP	6/22/2019	Y	Y	7	

Exactly Duplicated Records

Long comment

This is a very long comment intended to illustrate how the qcompare macro handles very long strings and special characters

Nonprintable Character

Character not in Compare

Compare Dataset

usubjid	paramcd	adtcmt	comment	saffl	pcfl	base	aval	ablfl	atoxgrn	atoxtrn	anl01fl	aphasen
STUDY-001-1234567890123	ALB	6/12/2019		Y			5	Y	1			
STUDY-001-1234567890123	ALB	6/12/2019		Y			4	Y	1	2		
STUDY-001-1234567890124	ALB	6/14/2019		Y			4	Y	1	2		
STUDY-001-1234567890125	ALB	6/13/2019		Y			4	Y	1	2		
STUDY-001-1234567890126	ALB	6/13/2019		Y			5	Y	1	2	Y	1
STUDY-001-1234567890126	ALB	6/13/2019		Y			5	Y	1	2	N	2
STUDY-001-1234567890127	HCT	6/15/2019	This is a very long comment intended to illustrate how the qcompare macro handles some very long strings and special characters	Y			8	Y	1	2	Y	2
STUDY-001-1234567890128	HGB	6/16/2019		Y			88888888	Y	1	2	Y	2

Duplicated
usubjid-
paramcd-
adt

Duplicated
usubjid-
paramcd-
adt

Distinct
atoxgrn

Distinct
anl01fl

Distinct
anl01fl

Long
comment

Nonprintable
Character

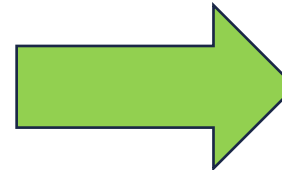
Word not in Base

Dataset and variable level comparison

These tables of metadata differences generated by qcompare are similar to proc compare

Data Set Summary

Dataset	NObs	Created	Modified	NVar
WORK.BASESUB	7	16JAN26:15:53:51	16JAN26:15:53:51	8
WORK.COMPARESUB	8	16JAN26:15:53:51	16JAN26:15:53:51	13



Compare has one more obs than base
Base has 5 more variables

Variables Summary

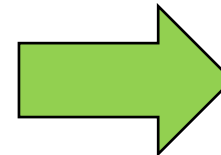
Number of Variables in Common: 6
Number of Variables in WORK.BASESUB but not in WORK.COMPARESUB: 2
Number of Variables in WORK.COMPARESUB but not in WORK.BASESUB: 7
Number of ID Variables: 3



Base and compare share 6 variable names
2 variable names in base not compare
7 variable names in compare not in base
3 ID variables were specified by user

Listing of Common Variables with Differing Attributes

Variable	Attribute	WORK.BASESUB	WORK.COMPARESUB
ABLFL	LABEL	Baseline~Record Flag	ablfl
PARAMCD	LABEL	Parameter Code	Param(13)Code
PARAMCD	LENGTH	100	10000
USUBJID	LABEL	Subjid Identifier	
USUBJID	LENGTH	100	10000



Base label different from compare for
ABLFL, PARAMCD, USUBJID
Base length different from compare for
PARAMCD and USUBJID

Variables list and proc compare special chars

Listing of Variables in WORK.COMPARESUB but not in WORK.BASESUB

Variable Name	Type	Length
ANL01FL	Char	1
APHASEN	Num	8
ATOXGRN	Num	8
ATOXTRN	Num	8
BASE	Num	8
PCFL	Char	1
SAFFL	Char	1

Listing of Variables in WORK.BASESUB but not in WORK.COMPARESUB

Variable Name	Type	Length
ITTFL	Char	1
PKFL	Char	1

Variables not common to base and compare datasets

The COMPARE Procedure
Comparison of WORK.BASESUB with WORK.COMPARESUB
(Method=EXACT)

Listing of Common Variables with Differing Attributes

Variable	Dataset	Type	Length	Label
usubjid	WORK.BASESUB	Char	100	Subjid Identifier
	WORK.COMPARESUB	Char	10000	
paramcd	WORK.BASESUB	Char	100	Parameter Code
	WORK.COMPARESUB	Char	10000	Param
				Code
ablfl	WORK.BASESUB	Char	1	Baseline~Record Flag
	WORK.COMPARESUB	Char	1	ablfl

Nonprintable
carriage return in
paramcd label
in compare

Non-common ID records in proc compare

➤ Proc compare provides a descriptive listing of nonmatching IDs where qcompare provides a table

A duplicate is noted for STUDY-001-1234567890123 ALB 12JUN2019 in compare however only the first of all duplicated IDs in compare dataset will be reported

One-by-one record match within ID causes one of the STUDY-001-1234567890123 ALB 12JUN2019 base duplicate to match compare and the other to be not found in compare

A duplicate is noted for STUDY-001-1234567890124 ALB 13JUN2019 in compare however only the first of all duplicated IDs in compare dataset will be reported

STUDY-001-1234567890124 ALB 14JUN2019 and
STUDY-001-1234567890126 ALB 13JUN2019
Are in compare and not base datasets

```
WARNING: The data set WORK.COMPARESUB contains a duplicate observation at observation number 2.
NOTE: At observation 2 the current and previous ID values are:
      usubjid=STUDY-001-1234567890123 paramcd=ALB adt=12JUN2019.
NOTE: Further warnings for duplicate observations in this data set will not be printed.
```

Comparison Results for Observations

```
Observation 2 in WORK.COMPARESUB not found in WORK.BASESUB: usubjid=STUDY-001-1234567890123
paramcd=ALB adt=12JUN2019.
```

```
Observation 2 in WORK.BASESUB not found in WORK.COMPARESUB: usubjid=STUDY-001-1234567890124
paramcd=ALB adt=13JUN2019.
```

```
WARNING: The data set WORK.BASESUB contains a duplicate observation at observation number 3.
NOTE: At observation 3 the current and previous ID values are:
      usubjid=STUDY-001-1234567890124 paramcd=ALB adt=13JUN2019.
NOTE: Further warnings for duplicate observations in this data set will not be printed.
```

```
Observation 3 in WORK.BASESUB not found in WORK.COMPARESUB: usubjid=STUDY-001-1234567890124
paramcd=ALB adt=13JUN2019.
```

```
Observation 3 in WORK.COMPARESUB not found in WORK.BASESUB: usubjid=STUDY-001-1234567890124
paramcd=ALB adt=14JUN2019.
```

```
Observation 5 in WORK.COMPARESUB not found in WORK.BASESUB: usubjid=STUDY-001-1234567890126
paramcd=ALB adt=13JUN2019.
```

```
Observation 6 in WORK.COMPARESUB not found in WORK.BASESUB: usubjid=STUDY-001-1234567890126
paramcd=ALB adt=13JUN2019.
```

```
Observation 7 in WORK.BASESUB not found in WORK.COMPARESUB: usubjid=STUDY-001-1234567890129
paramcd=ALP adt=22JUN2019.
```

STUDY-001-1234567890124 ALB 13JUN2019 and
STUDY-001-1234567890129 ALP 22JUN2019
Are in base and not compare datasets

Non-common ID records in qcompare

usubjid-paramcd-adt records in WORK.BASESUB and not found in WORK.COMPARESUB

usubjid	paramcd	adt	Observation Count	Observation Numbers	Duplicate ID Value?
STUDY-001-1234567890124	ALB	13JUN2019	2	2,3	Yes, WARNING
STUDY-001-1234567890129	ALP	22JUN2019	1	7	

STUDY-001-1234567890124 ALB 13JUN2019 and
STUDY-001-1234567890129 ALP 22JUN2019
Are in base and not compare datasets

STUDY-001-1234567890124 ALB 13JUN2019 has 2 duplicates

Duplicates located at obs 2 and 3

usubjid-paramcd-adt records in WORK.COMPARESUB and not found in WORK.BASESUB

usubjid	paramcd	adt	Observation Count	Observation Numbers	Duplicate ID Value?
STUDY-001-1234567890124	ALB	14JUN2019	1	3	
STUDY-001-1234567890126	ALB	13JUN2019	2	5,6	Yes, WARNING

STUDY-001-1234567890124 ALB 14JUN2019 and
STUDY-001-1234567890126 ALB 13JUN2019
Are in compare and not base datasets

STUDY-001-1234567890126 ALB 13JUN2019 has 2 duplicates

Duplicates located at obs 5 and 6

ID matching in presence of duplicates

- Proc compare reports STUDY-001-1234567890123 ALB 12JUN2019 as both a duplicate ID and a mismatching ID between base and compare
- Qcompare only reports this as a duplicate and not an ID mismatch
- Since base and compare datasets both have these IDs, for our QCs we do not need to report this as an ID mismatch
- The QC issue in our context is only to resolve the compare dataset duplicates
- Proc compare is matching base and compare in a one-to-one fashion based on their sorted order within the duplicated matched ID
- Qcompare uses sql join to find all matches within the duplicated matched ID group

```
WARNING: The data set WORK.COMPARESUB contains a duplicate observation at observation number 2.  
NOTE: At observation 2 the current and previous ID values are:  
       usubjid=STUDY-001-1234567890123 paramcd=ALB adt=12JUN2019.  
NOTE: Further warnings for duplicate observations in this data set will not be printed.  
  
Comparison Results for Observations  
  
Observation 2 in WORK.COMPARESUB not found in WORK.BASESUB: usubjid=STUDY-001-1234567890123  
       paramcd=ALB adt=12JUN2019.
```

Compare

usubjid	paramcd	adt	comment	saffl	pcfl	base	aval	ablfl	atoxgrn	atoxtrn	anl01fl	aphasen
STUDY-001-1234567890123	ALB	6/12/2019		Y			5	Y	1			
STUDY-001-1234567890123	ALB	6/12/2019		Y			4	Y	1	2		

Base

usubjid	paramcd	adt	comment	ablfl	ittfl	aval	pkfl
STUDY-001-1234567890123	ALB	6/12/2019		Y	Y	4	

Observation Summary in qcompare

Observation Summary

No entire records are exactly same between base and compare

Four records are common between base and compare matching by usubjid-paramcd-adt

Total Number of Observations Read from WORK.BASESUB: 7
Total Number of Observations Read from WORK.COMPARESUB: 8
Number of Observations with Some Compared Variables Unequal: 4
Number of Observations with All Compared Variables Equal: 0
Number of Observations in Common: 4
Number of Observations in WORK.BASESUB but not in WORK.COMPARESUB: 3

For three records in base, no match could be found in source by usubjid-paramcd-adt

The observation summary is similar to proc compare

Value Comparison in %qcompare

Variables with Unequal Values

Variable	Number of Value Differences
AVAL	4
COMMENT	1

There were 4 differences in AVAL and 1 in COMMENT

Value Comparison Results for Variables

AVAL						
usubjid	paramcd	adt	AVAL WORK.BASESUB	AVAL WORK.COMPARESUB	diff	pct_diff
STUDY-001-1234567890123	ALB	12JUN2019	4	5	1	25
STUDY-001-1234567890125	ALB	13JUN2019	5	4	-1	-20
STUDY-001-1234567890127	HCT	15JUN2019	5	8	3	60
STUDY-001-1234567890128	HGB	16JUN2019	7	88888888	88888881	1269841157.14285

First mismatching character beginning from start of base and compare strings

COMMENT

usubjid	paramcd	adt	COMMENT WORK.BASESUB	COMMENT WORK.COMPARESUB
STUDY-001-1234567890127	HCT	15JUN2019	This is a very long comment intended to illustrate how (16) the qcompare macro handles very long strings and special characters	This is a very long comment intended to illustrate how (18) the qcompare macro handles some very long strings and special characters

First mismatching character beginning from end of base and compare strings

Value Comparison in Proc Compare

The COMPARE Procedure
Comparison of WORK.BASESUB with WORK.COMPARESUB
(Method=EXACT)

Value Comparison Results for Variables

usubjid	paramcd	adt	Base Value comment	Compare Value comment
STUDY-001-1234567890	HCT	15JUN2019	This is a very long	This is a very long

usubjid	paramcd	adt	Base aval	Compare aval	Diff.	% Diff
STUDY-001-1234567890	ALB	12JUN2019	4.0000	5.0000	1.0000	25.0000
STUDY-001-1234567890	ALB	13JUN2019	5.0000	4.0000	-1.0000	-20.0000
STUDY-001-1234567890	HCT	15JUN2019	5.0000	8.0000	3.0000	60.0000
STUDY-001-1234567890	HGB	16JUN2019	7.0000	88888888	88888881	1.26984E9

+ denotes truncation of the long
comment strings

Usubjid is truncated
without any indication
within the report

Value Comparison in proc compare

Obs	_TYPE_	_OBS_	comment	usubjid	paramcd	adt	ablfl	aval
1	BASE	1		STUDY-001-1234567890123	ALB	12JUN2019	Y	4
2	COMPARE	1		STUDY-001-1234567890123	ALB	12JUN2019	Y	5
3	DIF	1		STUDY-001-1234567890123	ALB	12JUN2019	.	1
4	COMPARE	2		STUDY-001-1234567890123	ALB	12JUN2019	Y	4
5	BASE	2		STUDY-001-1234567890124	ALB	13JUN2019	Y	5
6	BASE	3		STUDY-001-1234567890124	ALB	13JUN2019	Y	5
7	COMPARE	3		STUDY-001-1234567890124	ALB	14JUN2019	Y	4
8	BASE	4		STUDY-001-1234567890125	ALB	13JUN2019	Y	5
9	COMPARE	4		STUDY-001-1234567890125	ALB	13JUN2019	Y	4
10	DIF	2		STUDY-001-1234567890125	ALB	13JUN2019	.	-1
11	COMPARE	5		STUDY-001-1234567890126	ALB	13JUN2019	Y	5
12	COMPARE	6		STUDY-001-1234567890126	ALB	13JUN2019	Y	5
13	BASE	5	This is a very long comment intended to illustrate how □ the qcompare macro handles very long strings and special characters	STUDY-001-1234567890127	HCT	15JUN2019	Y	5
14	COMPARE	7	This is a very long comment intended to ilustrate how □ the qcompare macro handles some very long strings and special characters	STUDY-001-1234567890127	HCT	15JUN2019	Y	8
15	DIF	3	XXX XXXXXXXXXXXXXXXXXXXXX.....	STUDY-001-1234567890127	HCT	15JUN2019	.	3
16	BASE	6		STUDY-001-1234567890128	HGB	16JUN2019	Y	7
17	COMPARE	8		STUDY-001-1234567890128	HGB	16JUN2019	Y	88888888
18	DIF	4		STUDY-001-1234567890128	HGB	16JUN2019	.	88888881
19	BASE	7		STUDY-001-1234567890129	ALP	22JUN2019	Y	7

Proc compare output dataset has full strings

Would need to reformat font to align X with strings

Univariate Statistics Comparison

Additional comparative report provided by qcompare

Univariate statistics comparison between WORK.BASESUB and WORK.COMPARESUB
Sorted by descending pct missing in base dataset

Variable	WORK.BASESUB								WORK.COMPARESUB							
	Records nonmissing	Records missing	Total records	% Missing	Min	Max	Median	Mean	Records nonmissing	Records missing	Total records	% Missing	Min	Max	Median	Mean
adt	7	0	7	0.00%	21712	21722	21713	21714.857	8	0	8	0.00%	21712	21716	21713	21713.5
aval	7	0	7	0.00%	4	7	5	5.4285714	8	0	8	0.00%	4	88888888	5	11111115
aphasen	-	-	-	-	-	-	-	-	4	4	8	50.00%	1	2	2	1.75
atoxgrn	-	-	-	-	-	-	-	-	8	0	8	0.00%	1	1	1	1
atoxtrn	-	-	-	-	-	-	-	-	7	1	8	12.50%	2	2	2	2
base	-	-	-	-	-	-	-	-	0	8	8	100.0%	-	-	-	-

There is an unusually large aval to investigate

All values of base are missing

Alternative ID groups comparison

Comparing for example two ADLB datasets, qcompare can tell us

- Which entire subjects are not common to base and compare
- Which entire paramcds are not common to base and compare
- Whether a given subject has an entire paramcd not common to base and compare

Subject STUDY-001-1234567890126 is only in base

Subject STUDY-001-1234567890129 is only in compare

usubjid not common to WORK.BASESUB and WORK.COMPARESUB

usubjid	Value in WORK.BASESUB Only	Value in WORK.COMPARESUB Only
STUDY-001-1234567890126	Y	
STUDY-001-1234567890129		Y

The ALP paramcd is only in base

paramcd not common to WORK.BASESUB and WORK.COMPARESUB

paramcd	Value in WORK.BASESUB Only	Value in WORK.COMPARESUB Only
ALP		Y

For subject STUDY001-1234567890126
the ALB paramcd is only in base

usubjid-paramcd not common to WORK.BASESUB and WORK.COMPARESUB

usubjid	paramcd	Value in WORK.BASESUB Only	Value in WORK.COMPARESUB Only
STUDY-001-1234567890126	ALB	Y	
STUDY-001-1234567890129	ALP		Y

For subject STUDY001-1234567890129
the ALP paramcd is only in compare

Duplicates Analysis

Duplicate usubjid-paramcd-adt groups in WORK.BASESUB dataset - Summary - WARNING

usubjid	paramcd	adt	Distinguishing Variables	Observation Numbers	Observation Count
STUDY-001-1234567890124	ALB	13JUN2019	**Exact Copies**	2,3	2

The duplicated records for subject STUDY-001-1234567890124 ALB 13JUN2019 in base were exact copies

The duplicated records for subject STUDY-001-1234567890124 ALB 13JUN2019 in base occurred at obs 2 and 3

Subject STUDY-001-1234567890124 had 2 ALB duplicate records on 13JUN2019 in base

Duplicate usubjid-paramcd-adt groups in WORK.COMPARESUB dataset - Summary - WARNING

usubjid	paramcd	adt	Distinguishing Variables	Observation Numbers	Observation Count
STUDY-001-1234567890123	ALB	12JUN2019	aval	1,2	2
STUDY-001-1234567890126	ALB	13JUN2019	anl01fl, aphasen	5,6	2

Subject STUDY-001-1234567890123 had 2 ALB records on 12JUN2019 in compare

Subject STUDY-001-1234567890126 had 2 ALB records on 13JUN2019 in compare

The 2 STUDY-001-1234567890123 ALB records on 12JUN2019 had different aval values in compare

The 2 STUDY-001-1234567890126 ALB records on 13JUN2019 had different anl01fl and aphasen values in compare

The 2 STUDY-001-1234567890126 ALB records on 13JUN2019 had the duplicate on obs 5 and 6 of compare

The 2 STUDY-001-1234567890126 ALB records on 12JUN2019 had the duplicate on obs 1 and 2 of compare

Duplicate Details

This can help us make sense of many duplicates in a dataset with hundreds of variables

Duplicate usubjid-paramcd-adt groups in WORK.COMPARESUB dataset - Details

usubjid	paramcd	adt	Variable Distinguishing the Duplicate	Record Number within Duplicate	Value Distinguishing the Duplicate
STUDY-001-1234567890123	ALB	12JUN2019	aval	1	5
			aval	2	4
STUDY-001-1234567890126	ALB	13JUN2019	anl01fl	1	Y
			anl01fl	2	N
			aphasen	1	1
			aphasen	2	2

The two STUDY-001-1234567890123 ALB on 12JUN2019 records can only be distinguished by aval values 5 and 4.

The two STUDY-001-1234567890126 ALB on 13JUN2019 records can be distinguished by anl01fl of Y and N

The two STUDY-001-1234567890126 ALB on 13JUN2019 records can be distinguished by aphasen of 1 and 2

Final Results

In the case that mismatch found in qcompare reports, qcompare reports the overall match as N

Final results

Qcompare:WORK.BASESUB WORK.COMPARESUB match?	date	time	Maximum difference tolerance	Maximum percent difference tolerance	Maximum difference observed	Maximum percent difference observed
N	16JAN26	21:53	.0000001	N/A	88888881	12698412

In the case that no mismatch were to be found in qcompare reports, qcompare would report the overall match as Y and then run the corresponding proc compare

The proc compare listing would then be checked both for strings indicating a mismatch within the listing and for the sysinfo value of 0

Final results

Qcompare:WORK.BASESUB WORK.COMPARESUB match?	date	time	Maximum difference tolerance	Maximum percent difference tolerance	Maximum difference observed	Maximum percent difference observed	Proc Compare:WORK.BASESUB WORK.COMPARESUB match?	sysinfo
Y	16JAN26	21:53	.0000001	N/A	0	0	Y	0

Conclusions

- The qcompare macro streamlines the data set comparisons for the specific purposes of validation of datasets in clinical trials setting
- The qcompare macro generates tabular html output in an easy-to-read report.
- It can compare very long character strings and can identify nonprintable characters
- Extra reports are provided for comparison of distinct values of alternative ID groups and for duplicate ID records
- The qcompare macro does not intend to replace proc compare, only to extend it for early stages of the validation in cases where proc compare is reporting a variety of differences that are hard to understand and summarize
- Recommend proc compare be run as a first pass and also as the final step as confirmation once qcompare has found agreement

Contact Information

- chris.weiner@bms.com

Thankyou for your interest!