Proc Compare: Wonderful Procedure!

Anusuiya Ghanghas, inVentiv International Pharma Services Pvt Ltd, Pune, India
Rajinder Kumar, inVentiv International Pharma Services Pvt Ltd, Pune, India

ABSTRACT

In Pharmaceutical industry, we work with huge volume of data. Sometime we want to cross check the data, if it is as per specification or sometimes after making some modifications we want to cross check if only desired changes has been made. Here comes in picture one of the wonderful procedures, PROC COMPARE provided by SAS. It helps us in comparing two datasets for their equality or difference with respect to number of observations, number of variables, dataset label, variable attributes and finally values of those variables. If used properly, this procedure can be of great help. It checks for all these, along with some other things in two different datasets or sometime in same dataset. But if not used properly, it can be quite problematic also.

Every SAS programmer will be happy to see "`NOTE: No unequal values were found. All values compared are exactly equal.`" as PROC COMPARE output. But it is not only getting this message, many other things also need to be checked before considering this as fully compared. This paper helps in understanding all other things in PROC COMPARE and if there are mismatches then how to handle them and get above message for full comparison.

INTRODUCTION

While comparing two datasets for their match or mismatch in values or observations, below things needs to be checked cautiously. Because "`NOTE: No unequal values were found. All values compared are exactly equal.`" Message alone doesn't guarantee full comparison of both the datasets. Here are few things listed which should be checked in PROC COMPARE result:

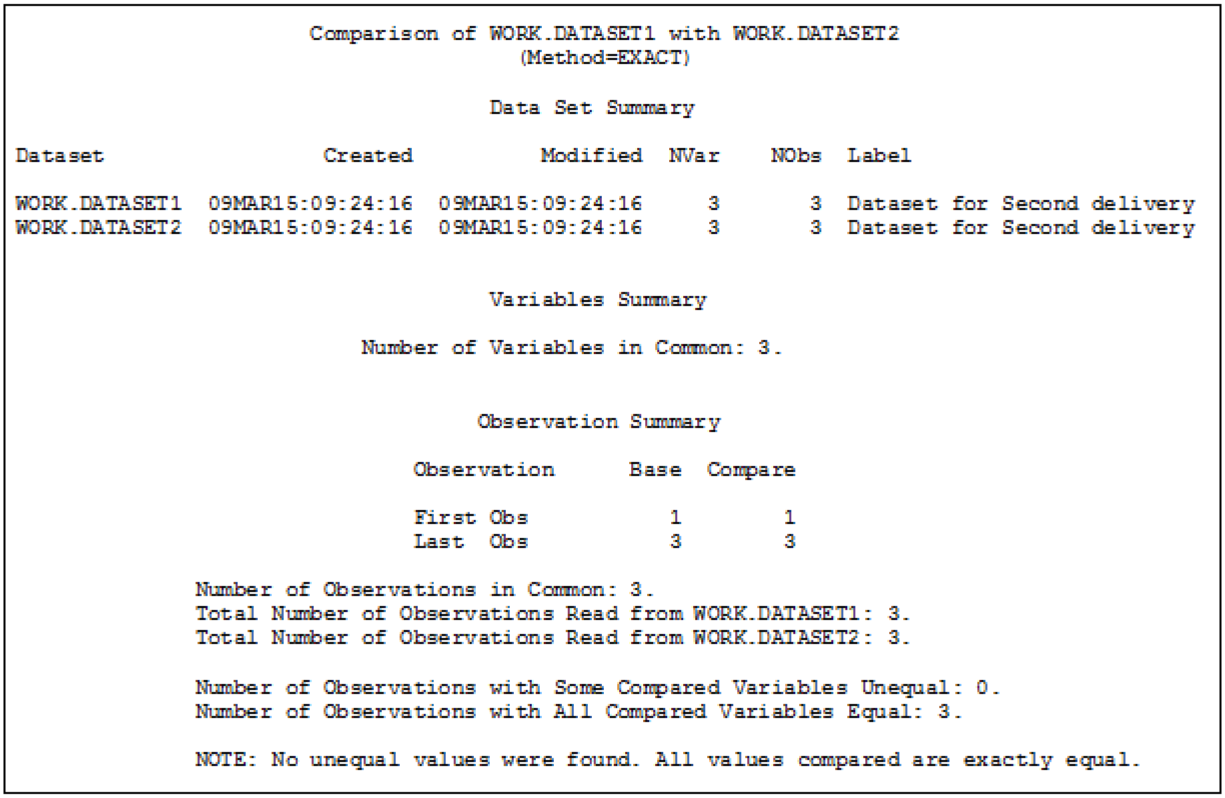
Table 1

Item	**Check**
Label of dataset	Are you comparing right datasets?
No. of observations	Do you have right no. of observations as per your expectations and it is same in both the datasets?
No. of variables	Are number of variables same as expected and it is same in both the datasets?
No. of common observations	Is there any difference in observations?
No. of common variables	Are all the variables same in both the datasets?
Values of the variables	Is value of all the variables in all the observations are exactly same in both the datasets?

To check all the above items let's start with basic things of PROC COMPARE procedure. First and foremost thing for comparing two datasets is sorting order, both the datasets should be sorted on same variables. Below is basic code for comparing two datasets:

```
PROC COMPARE BASE = dataset1 COMP = dataset2;
RUN;
```

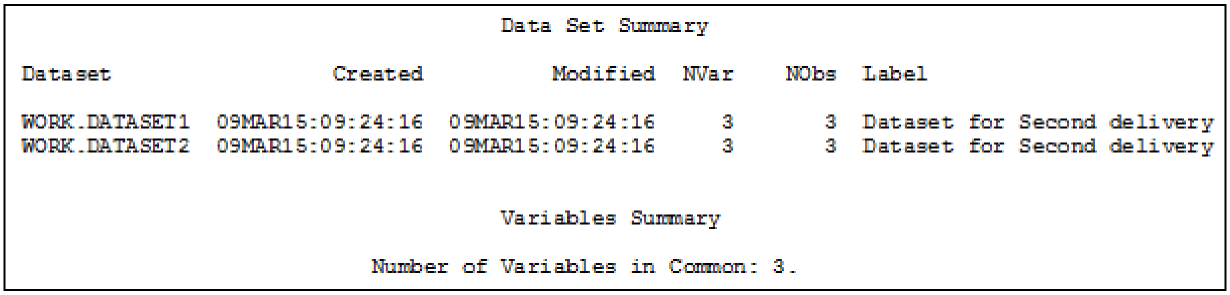
When all items mentioned in above table are checked and they are matching, in that case below output (Display 1) is generated form PROC COMPARE.



```
                        Comparison of WORK.DATASET1 with WORK.DATASET2
                                       (Method=EXACT)

                                     Data Set Summary

Dataset                 Created          Modified  NVar    NObs  Label

WORK.DATASET1  09MAR15:09:24:16  09MAR15:09:24:16     3       3  Dataset for Second delivery
WORK.DATASET2  09MAR15:09:24:16  09MAR15:09:24:16     3       3  Dataset for Second delivery


                                     Variables Summary

                            Number of Variables in Common: 3.


                                    Observation Summary

                               Observation      Base  Compare

                               First Obs           1        1
                               Last  Obs           3        3

              Number of Observations in Common: 3.
              Total Number of Observations Read from WORK.DATASET1: 3.
              Total Number of Observations Read from WORK.DATASET2: 3.

              Number of Observations with Some Compared Variables Unequal: 0.
              Number of Observations with All Compared Variables Equal: 3.

              NOTE: No unequal values were found. All values compared are exactly equal.
```

Display 1. Basic output for PROC COMPARE

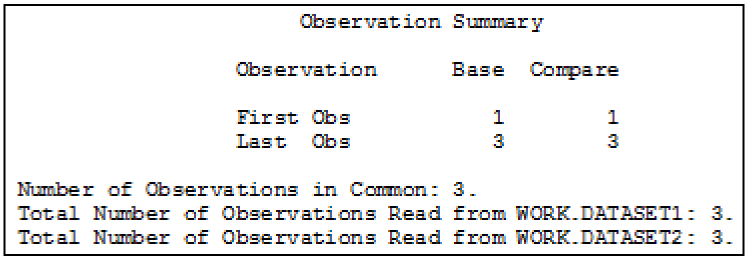
In Display 2 one can see that label of dataset is same which indicates that datasets used for comparison are correct. Now it can be seen that number of variables in both the datasets are same and they are exactly same as number of variables in common.



```
                                     Data Set Summary

Dataset                 Created          Modified  NVar    NObs  Label

WORK.DATASET1  09MAR15:09:24:16  09MAR15:09:24:16     3       3  Dataset for Second delivery
WORK.DATASET2  09MAR15:09:24:16  09MAR15:09:24:16     3       3  Dataset for Second delivery


                                     Variables Summary

                            Number of Variables in Common: 3.
```

Display 2. Data set and Variables summary

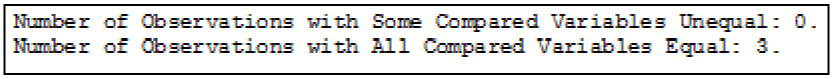
Further in Display 3 it can be seen, numbers of observations in each dataset are same and it is matching with number of observations in common and read from each dataset.



```
                      Observation Summary

                 Observation      Base  Compare

                 First Obs          1        1
                 Last  Obs          3        3

Number of Observations in Common: 3.
Total Number of Observations Read from WORK.DATASET1: 3.
Total Number of Observations Read from WORK.DATASET2: 3.
```

Display 3. Observation summary

Once all above checks are fine, then values of variables should be checked. Below output (Display 4) suggests that there is no observation which has some compared variables unequal.



```
Number of Observations with Some Compared Variables Unequal: 0.
Number of Observations with All Compared Variables Equal: 3.
```

Display 4. Values comparison for compared variables

Now final line "`NOTE: No unequal values were found. All values compared are exactly equal.`" gives immense pleasure to any programmer. But it is not that simple as it looks from above example. Because sometimes there can be mismatches in datasets with above final line appearing in output also. In such scenario one needs to be cautious, now let's check these things in compare output, where it is not matching.

1. MISMATCH IN DATASET LABEL

Sometimes all other checks are fine but label of the datasets are not matching or label is missing in one dataset. It looks a small thing, but sometimes it can be due to wrong dataset picked for comparison. For example there are two datasets in one library, one dataset was created on the time of last delivery 2 months back with name dataset1 and another one recently created for current delivery with name dataset11. For comparison purpose this dataset11 should be used, but by mistake due to old code available using dataset1 as compare dataset, it shows fully matched. In such cases, label of the dataset along with time each dataset was created and last modified should be checked for ensuring right datasets are getting compared.

Output 1 in appendix 1 shows compare output for such example. Where apart from label rest of the things looks fine. But when checking the label, it can be found that it is comparing with last delivery's data and it is not a small mistake.

2. NUMBER OF OBSERVATIONS

After ensuring right datasets are picked for comparing, next thing is ensuring numbers of observation are correct and matching in both the datasets. For example in dataset1, programmer knows, there should be 3 observations as per raw data and programming logic applied. Hence it should be checked that both the datasets have 3 observations only. Appendix 1, output 2 shows such example where dataset2 has one extra observation. If numbers of observations are ignored from PROC COMPARE result, then rest of things looks fine. And sometime in haste, it gives impression that it is fully matching. But that's not the truth. Simple PROC COMPARE gives only information that number of observations are not matching. To get more details for further investigation, LISTALL option along with ID statement are helpful. If VAR1 is used as ID variable then it gives below information for debugging:

Observation 4 in WORK.DATASET2 not found in WORK.DATASET11: var1 = 3;

Now reason behind having an extra observation with VAR1 = 3 value in DATASET2 can be checked. It can be due to selection of all the patients in place of only ITT or SAFETY population. This can be verified and accordingly corrected in corresponding dataset.

Sometimes numbers of observations are same in both the datasets but they are not common. Output 3 in appendix 1 shows such case. This needs to be corrected before moving to next item.

3. NUMBER OF VARIABLES

Number of variables in both the datasets should be same. They should be same variables. Output 4 in appendix 1 shows that dataset2 has one extra variable, although rest of the things is matching. But in real, it can be that important variable, which was recently added and due to its unavailability in dataset2, one can miss to compare its correct values. In some cases, even numbers of variables are same but number of common variables can be different. Output 5 in appendix 1 provides example for such case only. Here one of the reasons can be different name for the variable in both the datasets. If that is the issue then it can be corrected by renaming it in compare dataset or with the help of VAR and WITH statement, as mentioned in below piece of code.

```
PROC COMAPRE BASE = dataset11 COMP = DATASET2;

VAR var1 var2 var3;

WITH var1 var2 var13;

RUN;
```

If VAR and WITH statements are used then PROC COMPARE considers that extra variable for comparison, otherwise simply it ignores.

Other important thing related to variables, is their type. If type of variable is not matching even name of the variable is same, then SAS doesn’t compare them and shows result provided in output 6 in appendix 1. This should be handled by appropriately correcting the type of variable in corresponding dataset.

4. VALUES OF VARIABLES

Once number of observations, number of common observations, number of variables and number of common variables are matched, and then values of those variables needs to be checked for their match. Before jumping directly to comparison of values, there are some easy tips which can be time saving and helpful in faster comparison.

Below are few STATMENTS and OPTIONS listed which can be helpful in debugging the mismatches in values of variables.

Table 2

STATEMENT/TASK	**Task**
ID Statement	It helps in comparing values on these ID variables. ID variables are set of variable/s which defines values in the datasets uniquely.
BY Statement	It provides comparison by the BY variables. Sometimes it provides lengthy output and it becomes difficult to debug. Hence it is not used that much frequently.
VAR Statement	If one wants to compare only few variables, or for debugging purpose want to go for smaller chunk of variables first, then this statement can be really helpful.
WITH Statement	If names of the variables are not same in both the dataset, then this statement is helpful in comparing values of two variables with different names.
LISTALL Option	This option enables to provide all the mismatches in detail. This is helpful in debugging reason behind mismatches.
CRITERION Option	Sometimes in compare result it shows values (formatted) are equal but still it shows in comparison result that there are mismatches. It can be due to mismatches at precision level in values. This option allows user to ignore mismatches after defined precision in CRITERION option.
MAXPRINT Option	This option enables SAS to print more observations of comparison result, in place of standard number of observations provided by default SAS session.

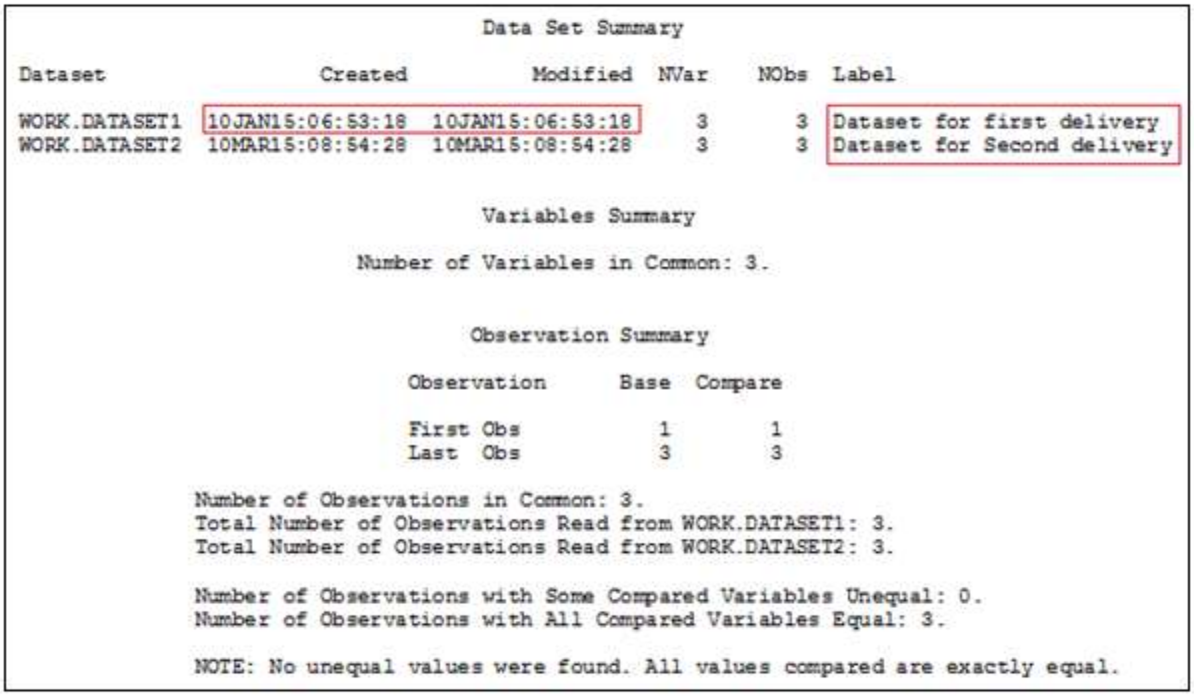
If there are high numbers of variables with mismatches, then it would be good idea to use VAR statement for selecting a chunk of variables for comparing initially. Then next set can be picked. One should remember that in the last entire dataset without VAR statement or with VAR statement having all the variables should be compared. Because sometime after making some changes in few variables, it can cause differences in other variables which were earlier matching.

Other major thing is, printing unmatched variables and their values. By default SAS has set MAXPRINT= (50,500) values, if there are more variables and lengthy values then these options can be used for enhancing the limit of printing mismatch values.

There are some more options witch are not use that much frequently but some of them are really very useful. Like OUTDIF, BRIEFSUMMARY, NOPRINT and TRANSPOSE. These options with some more options and their task can be found in the appendix 2.

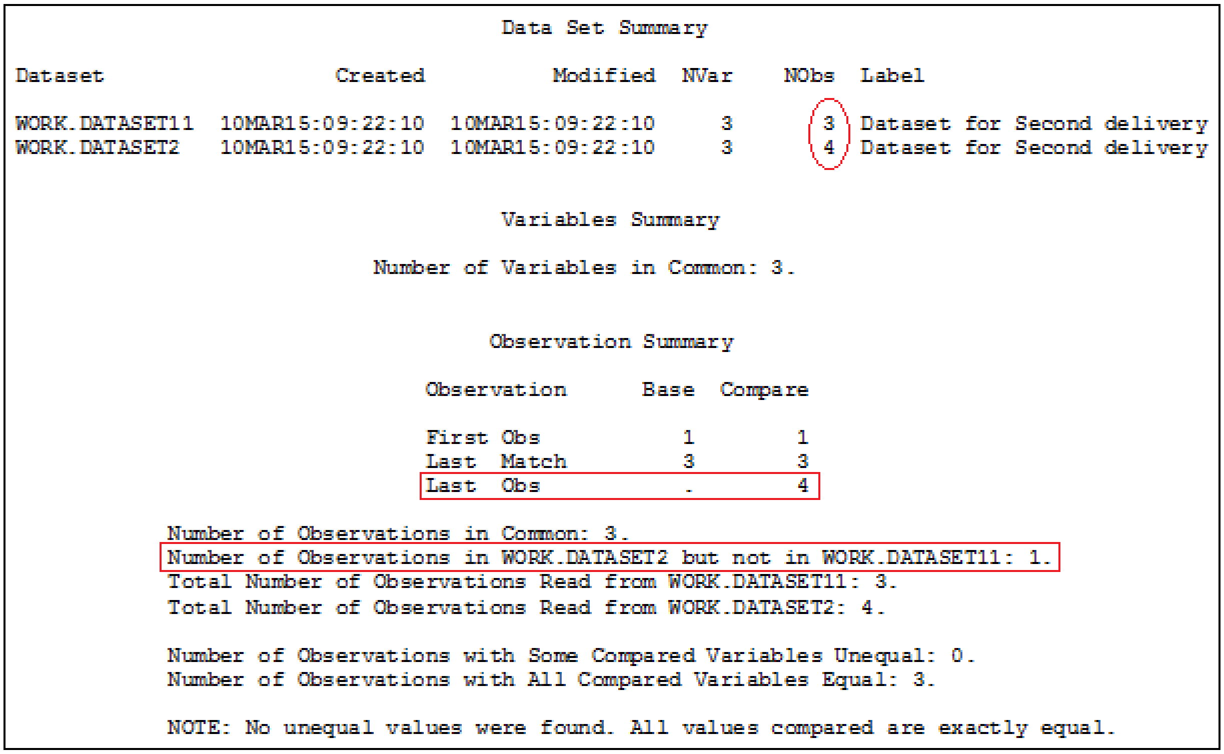
APPENDIX 1

OUTPUT 1



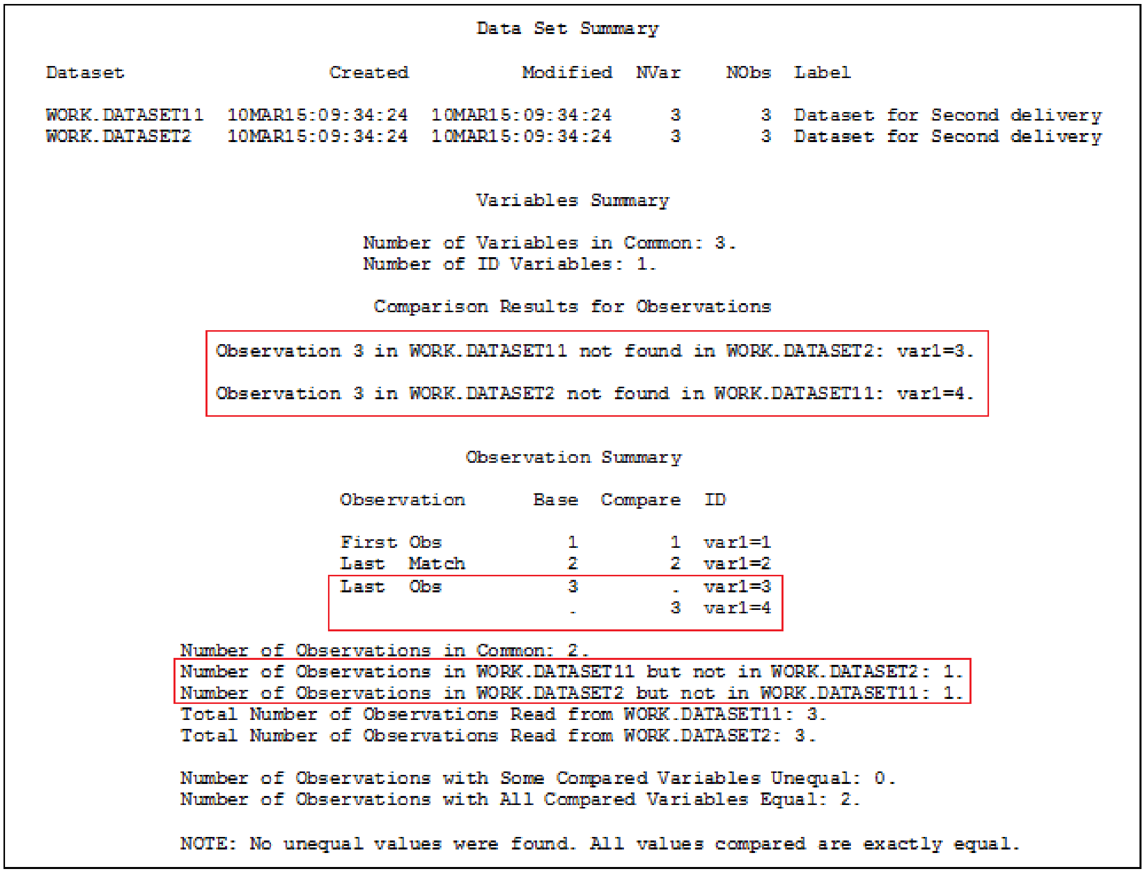
```
                                Data Set Summary

Dataset                 Created          Modified  NVar    NObs  Label

WORK.DATASET1  10JAN15:06:53:18  10JAN15:06:53:18     3       3  Dataset for first delivery
WORK.DATASET2  10MAR15:08:54:28  10MAR15:08:54:28     3       3  Dataset for Second delivery


                                Variables Summary

                         Number of Variables in Common: 3.


                               Observation Summary

                           Observation      Base  Compare

                           First Obs           1        1
                           Last  Obs           3        3

               Number of Observations in Common: 3.
               Total Number of Observations Read from WORK.DATASET1: 3.
               Total Number of Observations Read from WORK.DATASET2: 3.

               Number of Observations with Some Compared Variables Unequal: 0.
               Number of Observations with All Compared Variables Equal: 3.

               NOTE: No unequal values were found. All values compared are exactly equal.
```

OUTPUT 2



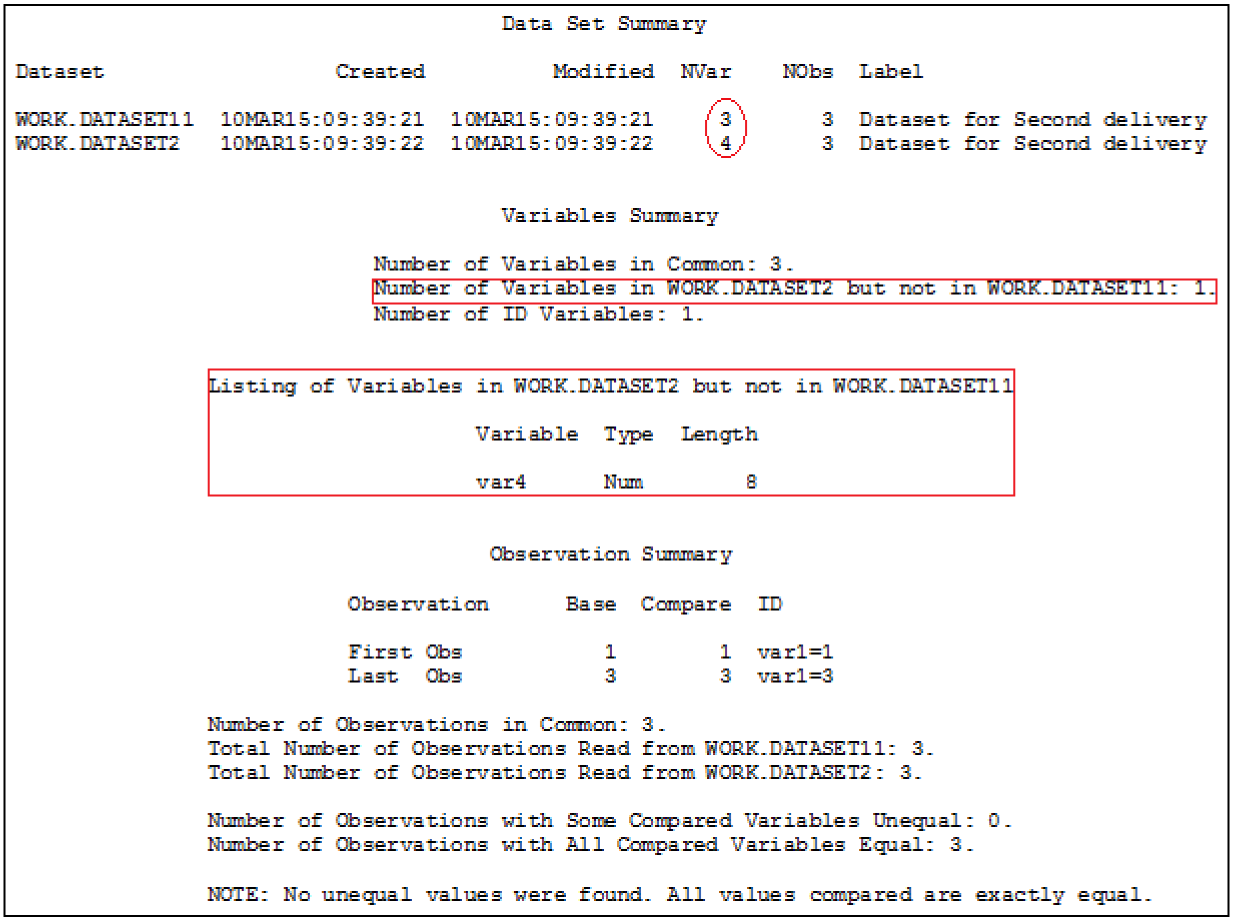
```
                                     Data Set Summary

Dataset                   Created          Modified  NVar    NObs  Label

WORK.DATASET11  10MAR15:09:22:10  10MAR15:09:22:10     3       3  Dataset for Second delivery
WORK.DATASET2   10MAR15:09:22:10  10MAR15:09:22:10     3       4  Dataset for Second delivery


                                     Variables Summary

                            Number of Variables in Common: 3.


                                    Observation Summary

                                Observation      Base  Compare

                                First Obs           1        1
                                Last  Match         3        3
                                Last  Obs           .        4

            Number of Observations in Common: 3.
            Number of Observations in WORK.DATASET2 but not in WORK.DATASET11: 1.
            Total Number of Observations Read from WORK.DATASET11: 3.
            Total Number of Observations Read from WORK.DATASET2: 4.

            Number of Observations with Some Compared Variables Unequal: 0.
            Number of Observations with All Compared Variables Equal: 3.

            NOTE: No unequal values were found. All values compared are exactly equal.
```

OUTPUT 3



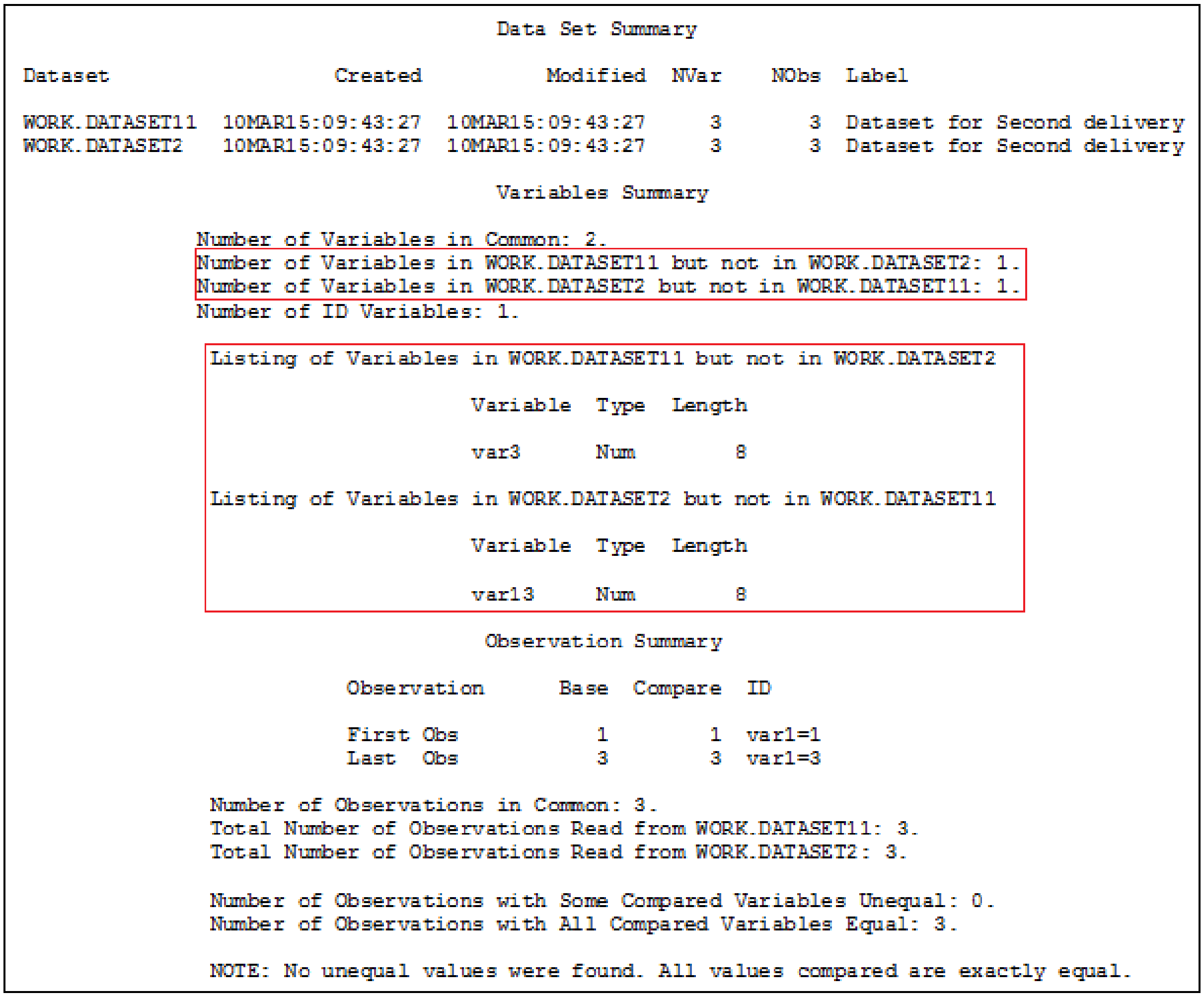
```
                                  Data Set Summary

Dataset                Created           Modified  NVar    NObs  Label

WORK.DATASET11  10MAR15:09:34:24  10MAR15:09:34:24     3       3  Dataset for Second delivery
WORK.DATASET2   10MAR15:09:34:24  10MAR15:09:34:24     3       3  Dataset for Second delivery


                                  Variables Summary

                           Number of Variables in Common: 3.
                           Number of ID Variables: 1.

                            Comparison Results for Observations

              Observation 3 in WORK.DATASET11 not found in WORK.DATASET2: var1=3.

              Observation 3 in WORK.DATASET2 not found in WORK.DATASET11: var1=4.


                                 Observation Summary

                         Observation      Base  Compare  ID

                         First Obs           1        1  var1=1
                         Last  Match         2        2  var1=2
                         Last  Obs           3        .  var1=3
                                             .        3  var1=4

           Number of Observations in Common: 2.
           Number of Observations in WORK.DATASET11 but not in WORK.DATASET2: 1.
           Number of Observations in WORK.DATASET2 but not in WORK.DATASET11: 1.
           Total Number of Observations Read from WORK.DATASET11: 3.
           Total Number of Observations Read from WORK.DATASET2: 3.

           Number of Observations with Some Compared Variables Unequal: 0.
           Number of Observations with All Compared Variables Equal: 2.

           NOTE: No unequal values were found. All values compared are exactly equal.
```

OUTPUT 4



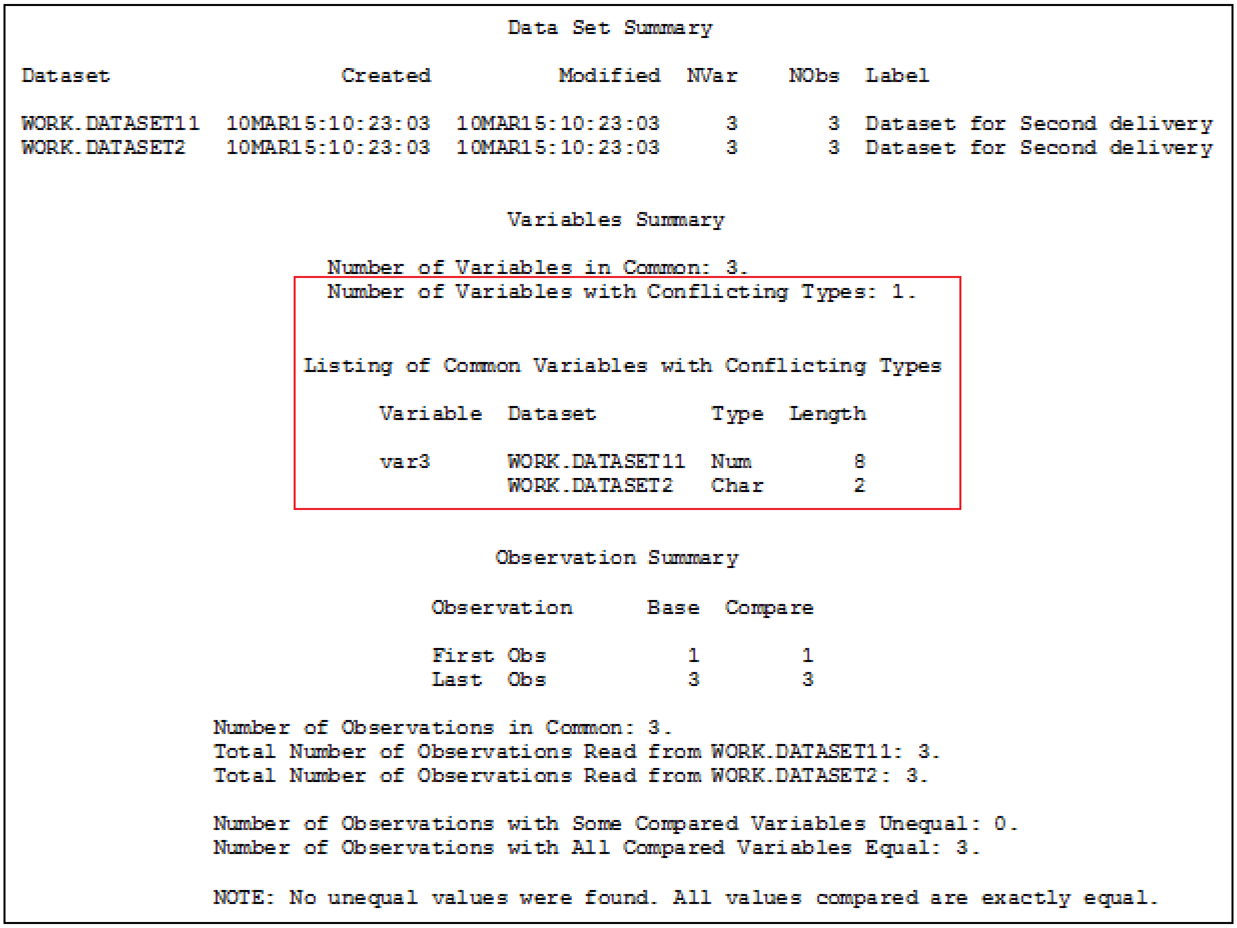
```
                                      Data Set Summary

Dataset                  Created          Modified  NVar    NObs  Label

WORK.DATASET11  10MAR15:09:39:21  10MAR15:09:39:21     3       3  Dataset for Second delivery
WORK.DATASET2   10MAR15:09:39:22  10MAR15:09:39:22     4       3  Dataset for Second delivery


                                      Variables Summary

                            Number of Variables in Common: 3.
                            Number of Variables in WORK.DATASET2 but not in WORK.DATASET11: 1.
                            Number of ID Variables: 1.


               Listing of Variables in WORK.DATASET2 but not in WORK.DATASET11

                                    Variable  Type  Length

                                    var4      Num        8


                                     Observation Summary

                          Observation      Base  Compare  ID

                          First Obs           1        1  var1=1
                          Last  Obs           3        3  var1=3

               Number of Observations in Common: 3.
               Total Number of Observations Read from WORK.DATASET11: 3.
               Total Number of Observations Read from WORK.DATASET2: 3.

               Number of Observations with Some Compared Variables Unequal: 0.
               Number of Observations with All Compared Variables Equal: 3.

               NOTE: No unequal values were found. All values compared are exactly equal.
```

OUTPUT 5



```
                                      Data Set Summary

Dataset                   Created          Modified  NVar    NObs  Label

WORK.DATASET11  10MAR15:09:43:27  10MAR15:09:43:27     3       3  Dataset for Second delivery
WORK.DATASET2   10MAR15:09:43:27  10MAR15:09:43:27     3       3  Dataset for Second delivery

                                      Variables Summary

              Number of Variables in Common: 2.
              Number of Variables in WORK.DATASET11 but not in WORK.DATASET2: 1.
              Number of Variables in WORK.DATASET2 but not in WORK.DATASET11: 1.
              Number of ID Variables: 1.

               Listing of Variables in WORK.DATASET11 but not in WORK.DATASET2

                                    Variable  Type  Length

                                    var3      Num        8

               Listing of Variables in WORK.DATASET2 but not in WORK.DATASET11

                                    Variable  Type  Length

                                    var13     Num        8

                                     Observation Summary

                          Observation      Base  Compare  ID

                          First Obs           1        1  var1=1
                          Last  Obs           3        3  var1=3

               Number of Observations in Common: 3.
               Total Number of Observations Read from WORK.DATASET11: 3.
               Total Number of Observations Read from WORK.DATASET2: 3.

               Number of Observations with Some Compared Variables Unequal: 0.
               Number of Observations with All Compared Variables Equal: 3.

               NOTE: No unequal values were found. All values compared are exactly equal.
```

OUTPUT 6



```
                                  Data Set Summary

Dataset                   Created          Modified  NVar    NObs  Label

WORK.DATASET11  10MAR15:10:23:03  10MAR15:10:23:03     3       3  Dataset for Second delivery
WORK.DATASET2   10MAR15:10:23:03  10MAR15:10:23:03     3       3  Dataset for Second delivery


                                  Variables Summary

                        Number of Variables in Common: 3.
                        Number of Variables with Conflicting Types: 1.


                      Listing of Common Variables with Conflicting Types

                            Variable  Dataset         Type  Length

                            var3      WORK.DATASET11  Num        8
                                      WORK.DATASET2   Char       2


                                 Observation Summary

                            Observation      Base  Compare

                            First Obs           1        1
                            Last  Obs           3        3

               Number of Observations in Common: 3.
               Total Number of Observations Read from WORK.DATASET11: 3.
               Total Number of Observations Read from WORK.DATASET2: 3.

               Number of Observations with Some Compared Variables Unequal: 0.
               Number of Observations with All Compared Variables Equal: 3.

               NOTE: No unequal values were found. All values compared are exactly equal.
```

APPENDIX 2

Task		**Option**
Control the output data set		
	Create an output data set	OUT=
	Write an observation for each observation in the BASE= and COMPARE= data sets	OUTALL
	Write an observation for each observation in the BASE= data set	OUTBASE
	Write an observation for each observation in the COMPARE= data set	OUTCOMP
	Write an observation that contains the differences for each pair of matching observations	OUTDIF

	Suppress the writing of observations when all values are equal	OUTNOEQUAL
	Write an observation that contains the percent differences for each pair of matching observations	OUTPERCENT
Create an output data set that contains summary statistics		OUTSTATS=
Specify how the values are compared		
	Specify the method for judging the equality of numeric values	METHOD=
	Judge missing values equal to any value	NOMISSBASE and NOMISSCOMP
Control the details in the default report		
	Include the values for all matching observations	ALLOBS
	Print a table of summary statistics for all pairs of matching variables	ALLSTATS and STATS
	Include in the report the values and differences for all matching variables	ALLVARS
	Print only a short comparison summary	BRIEFSUMMARY
	Change the report for numbers between 0 and 1	FUZZ=
	Suppress the print of creation and last-modified dates	NODATE
	Suppress all printed output	NOPRINT
	Suppress the summary reports	NOSUMMARY
	Suppress the value comparison results.	NOVALUES
	Produce a complete listing of values and differences	PRINTALL
	Print the value differences by observation, not by variable	TRANSPOSE
Control the listing of variables and observations		
	List all variables and observations found only in the base data set	LISTBASE
	List all observations found only in the base data set	LISTBASEOBS
	List all variables found only in the base data set	LISTBASEVAR
	List all variables and observations found only in the comparison data set	LISTCOMP
	List all observations found only in the comparison data set	LISTCOMPOBS
	List all variables found only in the comparison data set	LISTCOMPVAR
	List variables whose values are judged equal	LISTEQUALVAR
	List all observations found in only one data set	LISTOBS
	List all variables found in only one data set	LISTVAR

CONCLUSION

No doubt, PROC COMPARE is a wonderful procedure. It just needs some cautious approach and it can be very user friendly. This paper puts light on some of the common problems and their possible solutions, along with some of the important features of PROC COMPARE. While comparing two datasets for their match or expected mismatch, STATEMENTS and OPTIONS mentioned in this paper can be really handy. We hope this paper helps you in saving your time while comparing two datasets.

REFERENCES

SAS Institute Inc. (2009). *SAS onlineDOC®, Version 9.2,* Chapter 11, The COMPARE Procedure; pp 207-258, Cary, NC: SAS Institute Inc.

ACKNOWLEDGMENTS

We would like to thank our mangers Akhil Mishra and Sandeep Sawant respectively and our colleagues for their valuable inputs.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Anusuiya Ghanghas
inVentiv International Pharma Services Pvt Ltd
Commerzone, Yerawada, Pune, INDIA
Phone: +91 20 30569374
E-mail: anusuiya.ghanghas@inventivhealth.com

Rajinder Kumar
inVentiv International Pharma Services Pvt Ltd
Commerzone, Yerawada, Pune, INDIA
Phone: +91 20 30569217
E-mail: rajinder.kumar@inventivhealth.com