PharmaSUG China 2017 - Paper 63

Covering Proc Compare Limits

Mary Zeychelle Hundana, PPD, Bonifacio Global City, Taguig, Philippines

ABSTRACT

Dataset validation is one of the key factors to ensure that the data produced is of good quality. In SAS, PROC COMPARE includes several options and statements that offer great convenience in identifying differences of two datasets. However, this SAS procedure is still incapable of providing value comparisons among different types and relating the two datasets regardless of the sorting.

There could be instances that base dataset has a character date variable displayed in DATE9 format but the compare dataset date variable is in numeric date9 format. Although the user can see manually that they have the same values being displayed, PROC COMPARE will not make value comparison possible. Another common issue is that the result of the comparison might show difference in records even though two datasets contain the exact information, but with different record arrangement. The result should have been a 100-percent data match with sorting as the only difference but PROC COMPARE displays this as a complete mismatch. This is because of PROC COMPARE relates observations row by row. The result generated might show difference in records as to when the data are compared by the system as a whole.

%CompareDS is a macro code designed not to replace proc compare, but rather cover the limits of proc compare. This macro aims to focus on non-matching records by removing observations that have exactly the same values across all variables, whether it is in the same row or not. Also, it aims to promote comparability among variables that tend to have the same and/or conflicting data types.

INTRODUCTION

This paper intends to discuss the uses of the macro, %CompareDS. The said macro aims to allow the user to be able to proceed with data comparison, wherein some variables have different data types. Though data type checking is part of the validation process, it doesn't automatically mean that comparison should stop. How about allowing the user to compare the values he literally sees then just provide a dataset containing variables with different attribute? The said macro also aims to find matching records, assuming that the 2 datasets have different sorting. For example, the 2 datasets, A and B, being compared have only 3 observations. If dataset A contains values 1, 2, and 4 respectively; and dataset B contains values 2, 1, and 5 respectively. Would it be easier for the user if we eliminate the matching records then displaying only 4 and 5 as the mismatch? If we use PROC COMPARE, it would display 3 observations as mismatch. But if we use %CompareDS, it would display only 1 observation as a mismatch. Let's say for example that the difference has been fixed. %CompareDS would display this as matching datasets but alerting the user of the difference in sorting.

This macro does not do all the things that PROC COMPARE is capable of doing but will be a great help when the user is just starting his validation. Most of the time, the user tends to fix the non-matching records first then later on fix the difference in labels, length, formats, and other attributes.

MACRO BASICS

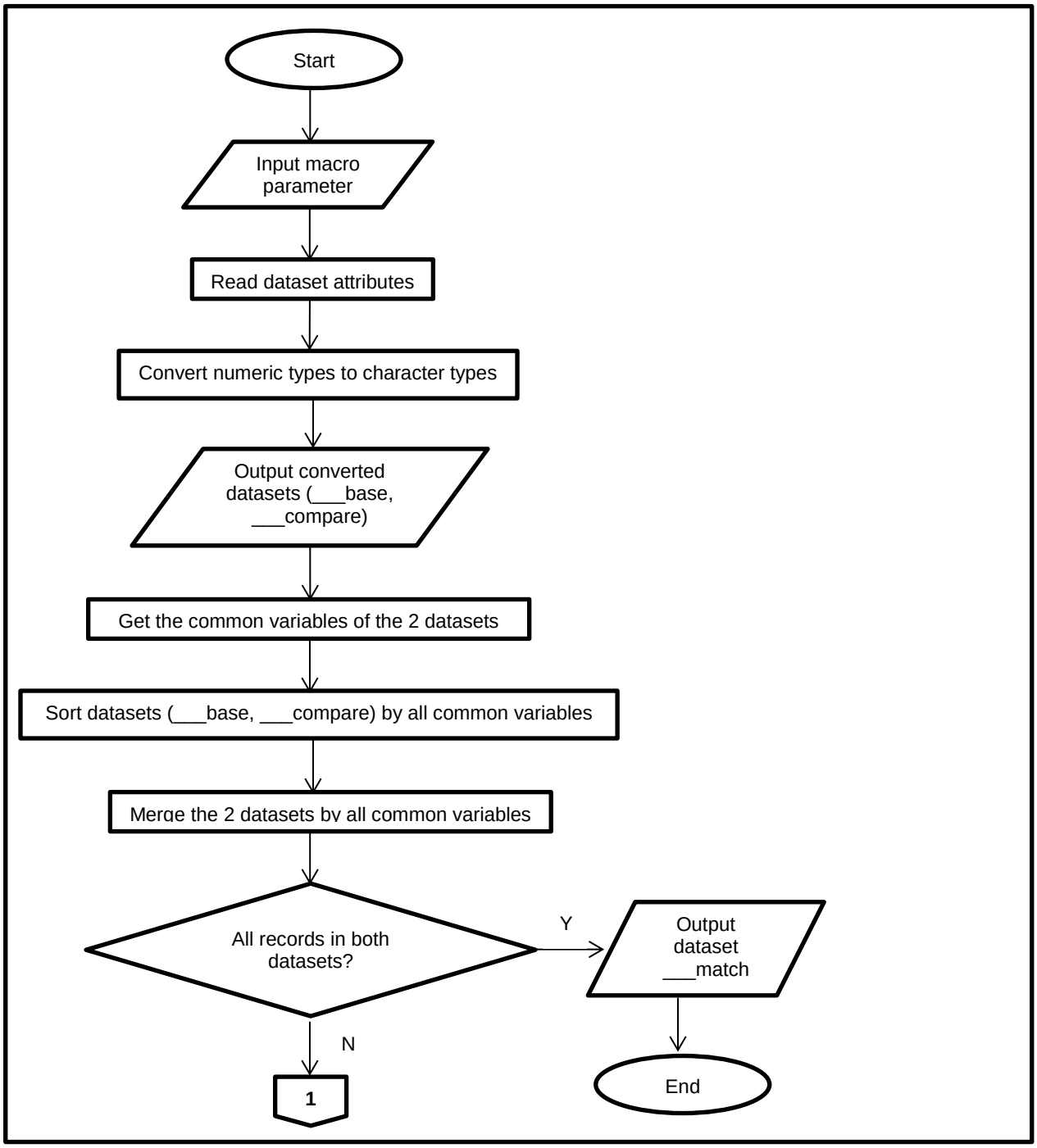
Below are the details regarding the datasets produced, the macro parameters, and the flow of the program.

Parameter	Description
___base	Base dataset used in comparison, wherein numeric variables have been converted to character type
___baserecs	Contains non-matching records of base dataset
___compare	Compare dataset used in comparison, wherein numeric variables have been converted to character type
___comparerecs	Contains non-matching records of compare dataset
___match	Contains matching records of base & compare datasets
___check	Contains mismatching records of base & compare datasets
___checkattrib	Contains mismatching attributes of base & compare datasets

Table 1. Output datasets

Parameter	Description	Example
base	Base dataset	folder.baseds, baseds, work.baseds
compare	Compare dataset	folder.compareds, compareds, work.compareds
sortby	Sorting to be used after the matching records have been eliminated. Should be separated by space	site subject foldername aeterm

Table 2. Parameters of Macro %CompareDS



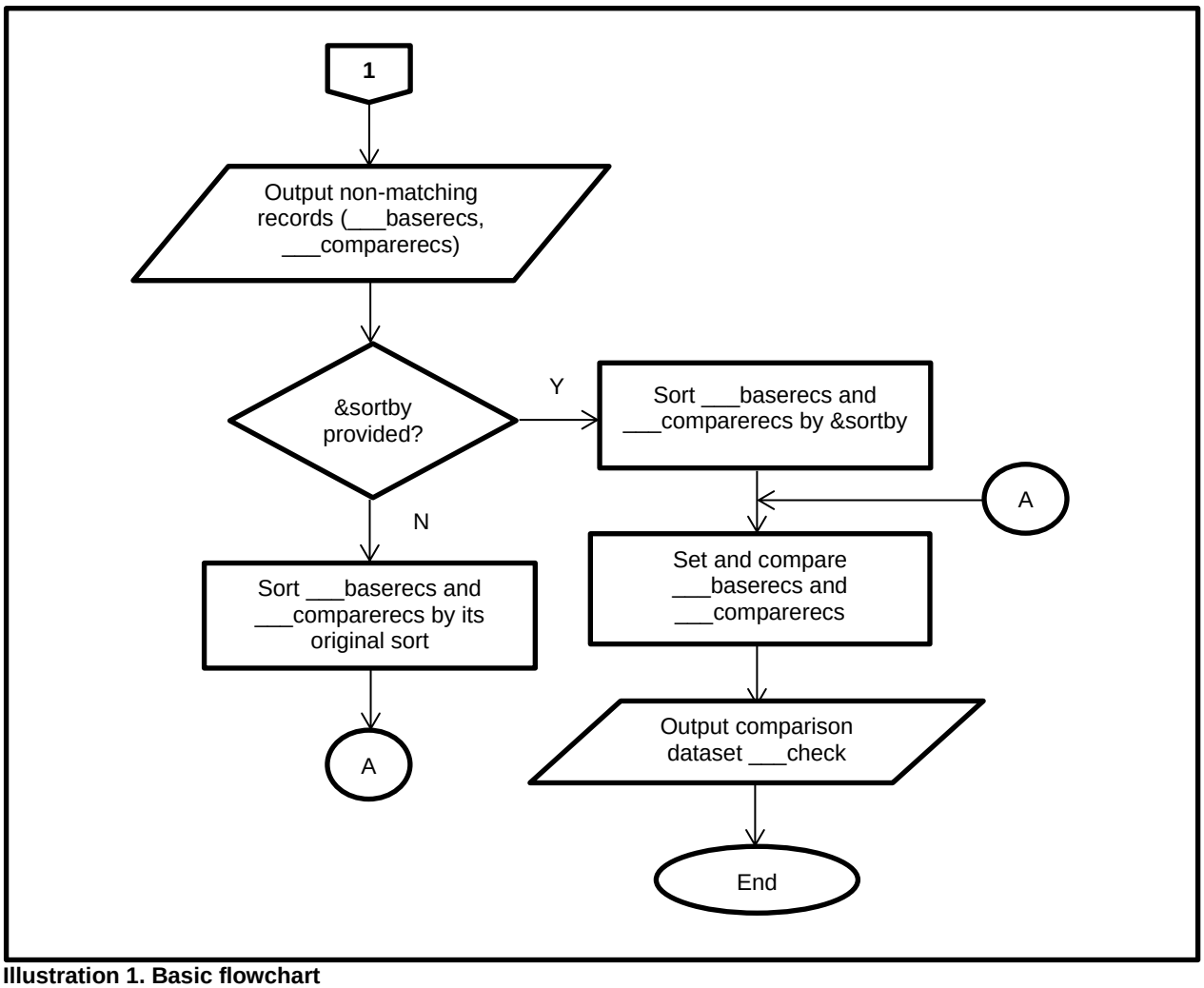


Illustration 1. Basic flowchart

DATASET COMPARISON

In SAS, comparison of different data types is not allowed; however, it would help the user if he can already start looking for the differences in values itself. Let's assume for example that the dataset is printed, would the user know if a value is a character or numeric type? Of course not. This is because, most of the time, people intend to compare what they actually see.

Below is an example of a base and compare datasets, and its result:

- **Column Properties**

Column Name	Type	Length
site	Text	765
subject	Text	150
aeterm	Text	600
aerel	Text	39
aestdt	Text	25
aeendt	Text	25

Display 1. Base dataset column properties

Column Name	Type	Length
SUBJECT	Text	150
SITE	Text	765
AETERM	Text	600
AESTDT	Num...	8
AEENDT	Num...	8
AEREL	Text	39

Display 2. Compare dataset column properties

Looking at Display 1 and Display 2, we can see that base dataset have all variables in character type; while compare dataset have numeric format for the two variables(aestdt, aeendt) and character format for the remaining variables. In SAS PROC COMPARE, it would only compare the values for the variables site, subject, aeterm, and aerel then informing the user that aestdt, and aeendt have different data types.

- **Sorting**

Attribute	Value
Compressed	No
Row Length	1604
Deleted Rows	0
Reuse	No
Point to Observation	Yes
Sorted by	site subject aerel

Display 3. Base dataset sorting

Attribute	Value
Compressed	No
Row Length	1576
Deleted Rows	0
Reuse	No
Point to Observation	Yes
Sorted by	SITE SUBJECT AETERM

Display 4. Compare dataset sorting

Looking at Display 3 and Display 4, we can see that the two datasets have difference in sorting. More likely, this would result to more mismatch being displayed. The user would tend to fix the sorting issue first before he can finally start with the 'true' comparison of the records.

- **Dummy Datasets**

	site	subject	aeterm	aerel	aestdt	aeendt
1	005-H Lee Moffitt Cancer Center and Research Institute	0051001	ALANINE AMINOTRANSFERASE (ALT) INCREASED	Not suspected	23FEB2016:00:00:00.000	
2	005-H Lee Moffitt Cancer Center and Research Institute	0051001	ASPARTATE AMINOTRANSFERASE(AST) INCREASED	Not suspected	09FEB2016:00:00:00.000	15MAR2016:00:00:00.000
3	005-H Lee Moffitt Cancer Center and Research Institute	0051001	HYPERKALEMIA	Not suspected	26APR2016:00:00:00.000	26APR2016:00:00:00.000
4	005-H Lee Moffitt Cancer Center and Research Institute	0051001		Suspected	22MAR2016:00:00:00.000	26APR2016:00:00:00.000
5	005-H Lee Moffitt Cancer Center and Research Institute	0051001	NAUSEA	Suspected	03MAY2016:00:00:00.000	08MAY2016:00:00:00.000
6	005-H Lee Moffitt Cancer Center and Research Institute	0051002		Not suspected	20MAR2016:00:00:00.000	24MAR2016:00:00:00.000
7	005-H Lee Moffitt Cancer Center and Research Institute	0051002	DIZZINESS	Not suspected	22MAR2016:00:00:00.000	14APR2016:00:00:00.000
8	005-H Lee Moffitt Cancer Center and Research Institute	0051002	VOMITING (EMESIS)	Not suspected	22MAR2016:00:00:00.000	22MAR2016:00:00:00.000
9	005-H Lee Moffitt Cancer Center and Research Institute	0051002		Not suspected	22MAR2016:00:00:00.000	24MAR2016:00:00:00.000
10	005-H Lee Moffitt Cancer Center and Research Institute	0051003	SHOULDER PAIN	Not suspected	18MAY2016:00:00:00.000	

Display 5. Base dataset

	SITE	SUBJECT	AETERM	AEREL	AESTDT	AEENDT
1	005-H Lee Moffitt Cancer Center and Research Institute	0051001	ALANINE AMINOTRANSFERASE (ALT) INCREASED	Not suspected	23FEB2016:00:00:00.000	.
2	005-H Lee Moffitt Cancer Center and Research Institute	0051001	ASPARTATE AMINOTRANSFERASE(AST) INCREASED	Not suspected	09FEB2016:00:00:00.000	15MAR2016:00:00:00.000
3	005-H Lee Moffitt Cancer Center and Research Institute	0051001	HYPERKALEMIA	Not suspected	26APR2016:00:00:00.000	26APR2016:00:00:00.000
4	005-H Lee Moffitt Cancer Center and Research Institute	0051001	HYPERTENSION	Suspected	22MAR2016:00:00:00.000	26APR2016:00:00:00.000
5	005-H Lee Moffitt Cancer Center and Research Institute	0051001	NAUSEA	Suspected	03MAY2016:00:00:00.000	08MAY2016:00:00:00.000
6	005-H Lee Moffitt Cancer Center and Research Institute	0051002	DIZZINESS	Not suspected	22MAR2016:00:00:00.000	14APR2016:00:00:00.000
7	005-H Lee Moffitt Cancer Center and Research Institute	0051002	GENERALIZED MUSCLE WEAKNESS	Not suspected	20MAR2016:00:00:00.000	24MAR2016:00:00:00.000
8	005-H Lee Moffitt Cancer Center and Research Institute	0051002	NAUSEA	Not suspected	22MAR2016:00:00:00.000	24MAR2016:00:00:00.000
9	005-H Lee Moffitt Cancer Center and Research Institute	0051002	VOMITING (EMESIS)	Not suspected	22MAR2016:00:00:00.000	22MAR2016:00:00:00.000
10	005-H Lee Moffitt Cancer Center and Research Institute	0051003	SHOULDER PAIN	Not suspected	18MAY2016:00:00:00.000	.

Display 6. Compare dataset

Looking at Display 5 and Display 6, we can see that some observations in base dataset matches some observations in compare dataset. To give the readers a better look, here are the matching records of base and compare datasets respectively:

- Observation 1 – Observation 1
- Observation 2 – Observation 2
- Observation 3 – Observation 3
- Observation 5 – Observation 5
- Observation 7 – Observation 6
- Observation 8 – Observation 9
- Observation 10 – Observation 10

The non-matching records for base dataset are in observations 4, 6, and 9; while the non-matching records for compare datasets are in observations 4, 7, and 8.

- **Result**

	TYPE	_OBS_	SITE	SUBJECT	AETERM	AEREL	AESTDT	AEENDT
1	BASE	4	005-H Lee Moffitt Cancer Center and Research Institute	0051001		Suspected	22MAR2016:00:00:00.000	26APR2016:00:00:00.000
2	COMPARE	4	005-H Lee Moffitt Cancer Center and Research Institute	0051001	HYPERTENSION	Suspected	22MAR2016:00:00:00.000	26APR2016:00:00:00.000
3	CHECK	.			XXXXXXX			
4	BASE	6	005-H Lee Moffitt Cancer Center and Research Institute	0051002		Not suspected	20MAR2016:00:00:00.000	24MAR2016:00:00:00.000
5	COMPARE	7	005-H Lee Moffitt Cancer Center and Research Institute	0051002	GENERALIZED MUSCLE WEAKNESS	Not suspected	20MAR2016:00:00:00.000	24MAR2016:00:00:00.000
6	CHECK	.			XXXXXXX			
7	BASE	9	005-H Lee Moffitt Cancer Center and Research Institute	0051002		Not suspected	22MAR2016:00:00:00.000	24MAR2016:00:00:00.000
8	COMPARE	8	005-H Lee Moffitt Cancer Center and Research Institute	0051002	NAUSEA	Not suspected	22MAR2016:00:00:00.000	24MAR2016:00:00:00.000
9	CHECK	.			XXXXXXX			

Display 7. %CompareDS Result

Before we arrive at our final output, numeric variables are first converted into character type. This is done by reading the attributes in sashelp.vcolumn then using the value in the format column to change the variable type. If ever the format column is missing, then the numeric variable will be converted to character using best. format. The author of this paper decided to go with all character compare because we are sure that all numeric type variables can be converted to character type variables; while not all character type variables can be converted to numeric type variables.

Once all variables have been converted to character type, then duplicate records will be deleted. If the user wants to avoid deleting of duplicate records, then he must make sure to add a primary key to his dataset. After the deletion of duplicate records, the 2 datasets will then be merged by all its common variables. The records that match successfully will be outputted in dataset ___match; while the non-matching records will be outputted in dataset ___check. The non-matching records will be sorted by using the macro variable &sortby., but if not available will use the original sort of the 2 datasets.

Even though some records from both base and compare datasets are not in the same row, the macro managed to eliminate the sorting issues then focused only on the 'real' non-matching records. The macro was also able to compare the variables (aestdt, aeendt) in different data types. _OBS_ variable in Display 7 indicates where the records is in base/compare. As you can see, it is not necessary for the records to be in the same observation number before it will be compared.

	TYPE	_OBS_	site	subject	aeterm	aerel
1	BASE	4	005-H Lee Moffitt Cancer Center and Research Institute	0051001		Suspected
2	COMPARE	4	005-H Lee Moffitt Cancer Center and Research Institute	0051001	HYPERTENSION	Suspected
3	DIF	4			XXXXXXXXXX	
4	BASE	6	005-H Lee Moffitt Cancer Center and Research Institute	0051002		Not suspected
5	COMPARE	6	005-H Lee Moffitt Cancer Center and Research Institute	0051002	DIZZINESS	Not suspected
6	DIF	6			XXXXXXXX	
7	BASE	7	005-H Lee Moffitt Cancer Center and Research Institute	0051002	DIZZINESS	Not suspected
8	COMPARE	7	005-H Lee Moffitt Cancer Center and Research Institute	0051002	GENERALIZED MUSCLE WEAKNESS	Not suspected
9	DIF	7			XXXXXXXXXX XXXXX XXXXXXX	
10	BASE	8	005-H Lee Moffitt Cancer Center and Research Institute	0051002	VOMITING (EMESIS)	Not suspected
11	COMPARE	8	005-H Lee Moffitt Cancer Center and Research Institute	0051002	NAUSEA	Not suspected
12	DIF	8			XXXXXXX XXXXXXX	
13	BASE	9	005-H Lee Moffitt Cancer Center and Research Institute	0051002		Not suspected
14	COMPARE	9	005-H Lee Moffitt Cancer Center and Research Institute	0051002	VOMITING (EMESIS)	Not suspected
15	DIF	9			XXXXXXX XXXXXXX	

Display 8. PROC COMPARE Result

In PROC COMPARE, more records were shown to have a mismatch. The user deals both with the sorting issue and differences in values itself. In order for the user to resolve this, he has to communicate with the other

programmer first regarding the sorting before he can actually see the differences in values. This would require the availability of the two programmers and might consume more time compared to looking at the 'real' differences then just discussing the sorting and attributes at the later part of the validation process.

ATTRIBUTE DIFFERENCES

The difference in attributes can be seen in dataset ___checkattrib. The variables being compared here are the common variables of the 2 datasets. This is the only comparison that uses PROC COMPARE as its tool.

	TYPE	_OBS_	name	type	length	varnum	label	format	informat
1	BASE	1	AEENDT	char	25	6			
2	COMPARE	1	AEENDT	num	8	6	End Date	DATETIME22.3	DATETIME22.3
3	DIF	1	.....	XXXX	-17	0	XXX.XXXX.....	XXXXXXXXXXX.....	XXXXXXXXXXX.....
4	BASE	2	AEREL	char	39	4	Relationship to Study Drug	$26.	$39.
5	COMPARE	2	AEREL	char	39	7	Relationship to Study Drug	$26.	$39.
6	DIF	2	.....	....	0	3	.....	.....	.....

Display 9. Example of Output with Difference in Attributes

ALERT MESSAGES

- **No common variables**

```
ALERT_I: No common variables in BASE/COMPARE
```

Display 10. Alert Message if there is no common variables between the 2 datasets

If the message above appears, this means that no comparison took place. The macro %CompareDS needs at least 1 same variable in order to proceed with the comparison.

- **Sorting Issue**

```
Alert_I: Difference in Sorting. Check variables _OBS_BASE, _OBS_COMPARE in ___match dataset.
```

Display 11. Alert Message if there is a sorting issue

The message above will appear if the sorting of the 2 datasets are not the same. Difference in values of variables _obs_base and _obs_compare means that the actual row numbers are different. In order to resolve this issue, the user must make sure that base and compare datasets have the same, enough sorting keys.

	SUBJECT	SITE	AETERM	AESTDT	AEENDT	AEREL	_OBS_BASE	_OBS_COMPARE
1	0051001	005-H Lee Moffitt Cancer Center and Research Institute	ALANINE AMINOTRANSFERASE (ALT) INCREASED	23FEB2016:00:00:00.000		Not suspected	2	1
2	0051001	005-H Lee Moffitt Cancer Center and Research Institute	ASPARTATE AMINOTRANSFERASE(AST) INCREASED	09FEB2016:00:00:00.000	15MAR2016:00:00:00.000	Not suspected	1	2
3	0051001	005-H Lee Moffitt Cancer Center and Research Institute	HYPERKALEMIA	26APR2016:00:00:00.000	26APR2016:00:00:00.000	Not suspected	4	3
4	0051001	005-H Lee Moffitt Cancer Center and Research Institute	HYPERTENSION	22MAR2016:00:00:00.000	26APR2016:00:00:00.000	Suspected	3	4

Display 12. Sample data with difference in sorting

- **Missing dataset**

```
ALERT_I: Either BASE/COMPARE dataset is missing
```

Display 13. Alert Message if the either base/compare dataset is missing

In order to resolve the above issue, the user must make sure that both base and compare datasets are present.

- **Missing Variables**

```
ALERT_I: Variable/s that exist only in compare dataset: FOLDER
```

The message above can be resolved by making sure that the variables in both base and compare dataset are the same.

CONCLUSION

%CompareDS can only be used as an aid when using PROC COMPARE. It is great that we can already start the validation of variables with different data types then just providing the user a dataset which contains the difference in

attributes. It is also of a great convenience that we are able to eliminate the matching records of the 2 datasets with difference in sorting so we can focus on non-matching records. If someone would like to improve this macro, I suggest incorporating the uses of PROC COMPARE so that it can soon be used as a replacement for it.

REFERENCES

SAS 9.2 Online product documentation. Available at support.sas.com/documentation/92/index.html

ACKNOWLEDGMENTS

I would like to express my gratitude to the people who supported me from the beginning up to finishing this paper. Without them this paper would not be possible.

I would like to thank my company, PPD, for allowing me to showcase my knowledge not only within the company, but also outside the organization.

I would also like to thank my colleagues, manager, and all those who offered help in checking my program and paper, those who provided feedbacks and comments, and those who assisted in editing and proofreading.

I would also like to thank my family for providing their utmost support and continuously encouraging me to show people what I am capable of doing.

CONTACT INFORMATION

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

Name: Mary Zeychelle Hundana
Address: 22F Net Park Bldg., 5th Ave, BGC Taguig
City, State ZIP: Taguig, BGC, 1634
E-mail: MaryZeychelle.Hundana@ppdi.com