

I work for Ephicity as Technical lead - Programming, having around 10 years of experience in statistical programming in clinical domain. My expertise revolves around submission activities in SDTM/ADaM programming, TLF programming having worked in several therapeutic areas such as Oncology, Immunology and Neurosciences. SDTM/ADaM programming activities includes specifications creation, annotation and creating define.xml. My interests include creation of automation macros helping make programming easier for fellow colleagues. Previously, I've served Accenture Services, Quintiles(Now IQVIA).

Macro Trio in SAS: Dataset Comparison and Reporting



PHUSE US Connect 2020, Orlando, 11th March 2020,
Nagaraju Mancha,
Technical Lead Programmer (Biostatistics & Programming)

Agenda

- Introduction
- Why should we use macros to compare datasets?
- Macro1: Comparing datasets of two libraries and creating report
- Macro2: Compress/zip the datasets
- Macro3: Creating folder and back-up of the datasets
- Conclusion



Introduction

- As Quality is utmost priority in clinical data submission, it's a programmer duty to deliver quality data or reports which are going to submit regulatory agencies. In order to make sure the datasets or reports have no issues; we need to have an efficient quality checks to trace the errors. SAS is a great tool having macros functionality to automate the repeated tasks. Programmers expert in macros can develop various quality check macros to track the issues whereby quality datasets can be delivered to our clients within short span of time.
- **Requirement:** As clinical trial analysis is ongoing process so when latest data available, Programmers will have to execute their programs to reflect the latest data into SDTM/ADaM. In order to know what changes were happened we need to compare datasets of earlier with latest. Comparing individual dataset might be an easy task with proc compare procedure.
- Can we compare all datasets of two libraries at once? Yes, it is possible with SAS macros. This paper illustrates a group of macros which will do:
- **Macro1: %domain_compare** will compare all the datasets of libraries at once and create the report which contains mismatches or updates of datasets with clear explanations. The output report of this program will have variable, structural, label and data differences thereby we can improve quality of data and tracing the updates.

Why should we use macro to compare datasets?

- Can we compare datasets with proc compare procedure without using of macros?
Certainly, we can. What if more no of datasets to be compared? It is time consuming process sometimes there is a chance to miss out the datasets.
- Proc compare alone can't compare all datasets at once but macro will compare all datasets of two libraries and generate output with clear details.
- Output generated by macro will have clear details such as variable level changes, data changes and changes will be highlighted with colors and it gives excel outputs to know the record level changes. But proc compare out generate pdf report by default so programmer must check each section to exactly know the changes.
- This macro will send reports to programmers through mail by enabling required parameter. But proc compare will not send reports on mail.
- Compare procedure issues error message when data not available in datasets to be compared but macro creates a report for blank or null datasets.
- macros representing in this paper are standalone so these can be used by any team who has SAS access.

Proc compare Syntax and options

- `proc compare base=sdtm.lb compare=vsdtm.v_lb out=lb_diff outbase outcomp outnoequal outdiff;`
- `run;`
- - **Base** = <dataset> and **compare** = <dataset> are datasets to be compared.
- - **Out** = <dataset> will be used when we want to see the differences in a dataset.
- - **outbase** option writes records from Base dataset
- - **outcomp** option writes records from Compare dataset
- - **Outnoequal** option Suppress the observations when they are equal in base and compare.
- - **outdiff** option writes the observations to output dataset
- Proc compare procedure has lot of other options, those can be used as per programmer requirement and please refer SAS documentation (<http://support.sas.com/documentation/>) for complete list of options and its uses.

Macro to compare datasets of two libraries and creating report

- Comparison Macro **%domain_compare** has following parameters. Programmer must provide appropriate values to each parameter for the program's successful execution. Look at the below macro call and its parameters.

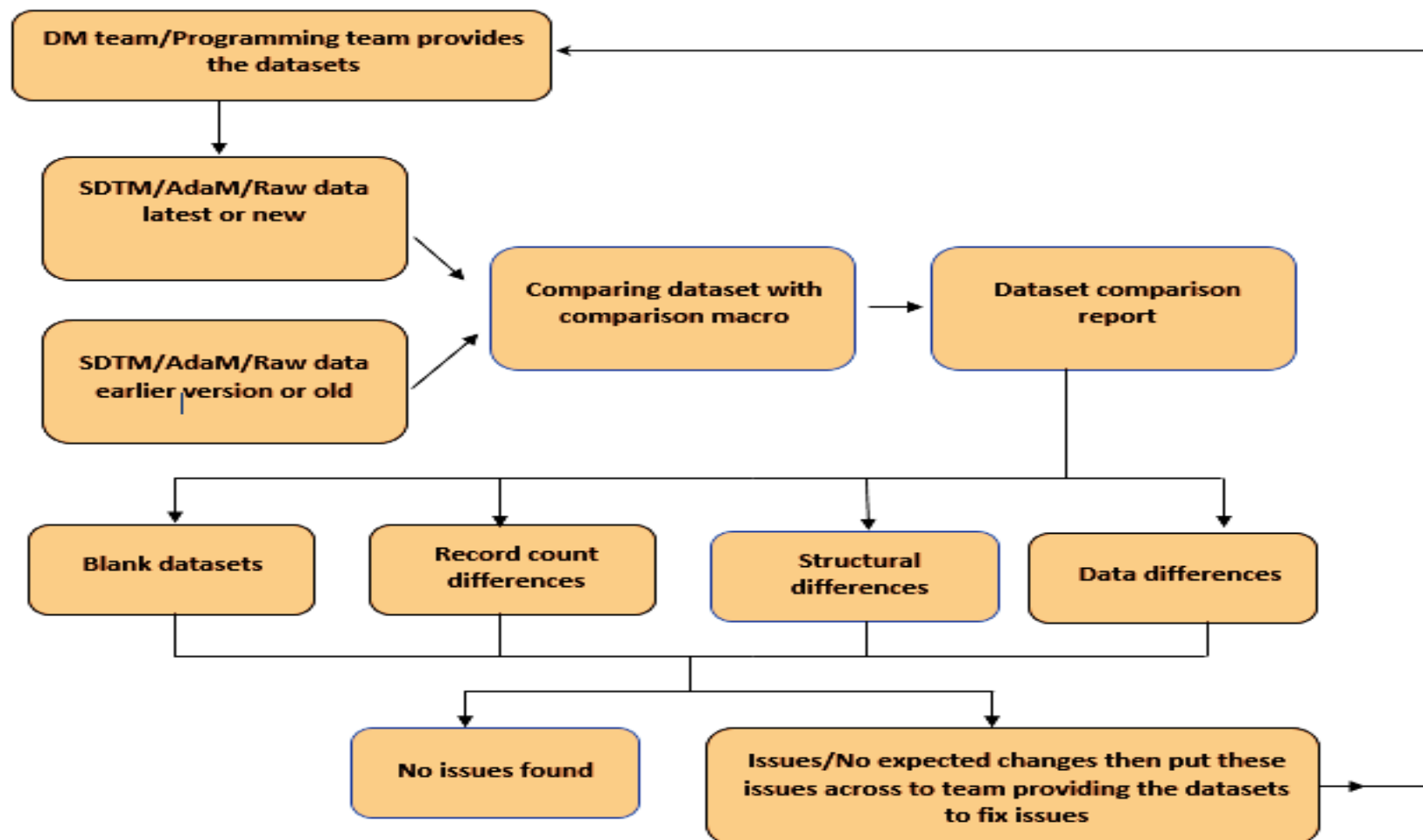
```
%domain_compare (prot=TEST-001,  
    old_path=c:/sdtm/datasets_15JAN19,  
    new_path=c:/sdtm/datasets_23JAN19,  
    output=c:/sdtm/compare,uneqc=Y,  
    ftype=SDTM,mail_this_report=Y);
```

- **Prot=** Protocol number or STUDYID as it is defined in protocol. It will appear in proc compare output.
- **Old_path=** Provide the path for earlier version/old datasets available.
- **New_path=** Provide the path where latest datasets available.
- **Output=** Specify the path to save the output comparison report. Comparison report will be saved with date and time at specified location. PDF and excel outputs will be saved in same location specified in the parameter.

Macro to compare two libraries and creating report

- **Uneqc=** if two datasets to be compared have unequal records, proc compare output gives lot of discrepancies; so, if we don't want compare datasets having unequal records then specify as N else specify as Y.
- **Ftype=** specify the category of file such as SDTM or ADaM or RAW/EDC datasets
- **mail_this_report=** if programmer wishes to get this report on mail, specify a value Y to this parameter. Specify N if we don't want this report on mail.(Note: Mail setup feature should be available in SAS installed in your system).
- In order to execute the program without having issues, User has to provide valid values to each parameter. Invalid values of any parameter may lead to issue error message in log window and it will stop executing the program.

Dataset comparison process flow



Comparison report details

- This is a first page of report here we can see the study details and author who did execute and when it was created.
- List of checks has been used in program.
- Path for current and earlier version of datasets.
- Study name, username and program version used to create the report.
- Programmer has to check the details displayed in this section to make sure everything is correct.

COMPARE REPORT FOR THE STUDY:TEST1001

02:25 Mon

STUDY DETAILS

STUDY INFORMATION	
STUDY:	TEST1001
USERNAME:	MANCHAN
AUTHOR:	EPHICACY ANALYSIS TEAM
PROC COMPARE VESION:	1.0
LIST OF CHECKS USED IN THIS PROC COMPARE:	
1. RECORD COUNTS BETWEEN OLD AND NEW DATASETS.	
2. EXISTENCE OF DATASETS BETWEEN TWO LIBRARIES.	
3. FINDING THE VARIABLE DIFFERENCES IN LENGTH,TYPE,FORMAT, LABEL AND ITS VALUES.	
4. COMPARING OLD DATASETS WITH NEW DATASETS AND EXPORTING THE DIFFERENCES	
PATH FOR THE OLD SDTM DATASETS: C:/SDTM/OLD	
PATH FOR THE NEW SDTM DATASETS: C:/SDTM/NEW	

Record count differences

- In this sections we can see the no of records in each dataset and the record count differences between current and earlier datasets.
- User can also find when the datasets were created.
- Blue color indicates that records have been included and red color indicates records have been deleted.
- No comments will be provided if there is no record count differences between current and prior.
- Dates when the datasets were created will also be displayed here.
- if no data was collected then record count would be zero. Programmer will have to check the reason for not having data

RECORD COUNT COMPARE OF OLD DATASET VS NEW DATASET

Dataset Library Name	Dataset Name	Records in Old Library	Old Library dataset creation date	Records in New Library	New Library dataset creation date	Comments
NEW	AE	240	15JAN19:19:49:05	242	23JAN19:12:55:50	2 Records have been Included
NEW	CM	445	15JAN19:19:49:19	445	23JAN19:12:56:10	
NEW	DA	786	15JAN19:19:49:29	786	23JAN19:12:56:21	
NEW	DM	31	15JAN19:19:50:45	31	23JAN19:12:59:26	
NEW	DS	52	15JAN19:19:49:33	52	23JAN19:12:56:26	
NEW	DV	56	11DEC18:22:58:46	56	15JAN19:19:54:17	
NEW	EG	137	15JAN19:19:49:38	138	23JAN19:12:56:33	1 Records have been Included
NEW	EX	42	15JAN19:19:49:41	42	23JAN19:12:56:38	
NEW	IE	0	15JAN19:19:49:12	0	23JAN19:12:55:55	
NEW	LB	6130	15JAN19:19:50:37	6161	23JAN19:12:58:10	31 Records have been Included
NEW	LB_2709	5798	27SEP18:10:29:09	5798	27SEP18:10:29:09	
NEW	MH	401	15JAN19:19:49:47	401	23JAN19:12:56:45	
NEW	MI	65	15JAN19:19:50:16	65	23JAN19:12:57:10	
NEW	MQ	203	15JAN19:19:49:24	204	23JAN19:12:56:15	1 Records have been Included
NEW	PR	137	15JAN19:19:49:51	136	23JAN19:12:56:52	1 Records have been Deleted

Blank/Null datasets and variable level changes

BLANK DATASETS WITH NO OBSERVATIONS

Comments	library	Dataset
These domains don,t have data please check and confirm...	NEW	IE
These domains don,t have data please check and confirm...	NEW	SUPPIE

VARIABLE LEVEL REPORT OF OLD DATASET VS NEW DATASET

Obs	domain	name	old	new	comment
1	DV				The domain DV has been added please check
2	LB_2709				The domain LB_2709 has been added please check
3	SUPPDV				The domain SUPPDV has been added please check
4	SUPPLB				The domain SUPPLB has been added please check
5	LB	LBSCAT			The variable LBSCAT has been added in LB..Please check..
6	MI	MIREASND			The variable MIREASND has been deleted in MI..Please check..

Proc compare output details

- Compare report of each dataset will be displayed here. User can check any domain directly through the bookmark in the report.
- Blank or null datasets will not be compared however null datasets will be reported in other section of same report.
- Each dataset compare report will have dataset summary, variable summary, variable attribute summary and observation summary.
- User not required to check the variable and dataset summary as these changes will be reported in other section of same report.

The screenshot displays the 'SDTM COMPARE REPORT FOR THE STUDY:TEST1001' interface. On the left, a 'Bookmarks' pane shows a tree view of reports, with 'VARIABLE LEVEL REPORT OF OLD DATASET VS NEW DATASET' selected. The main pane shows the details for comparing 'EX(21AUG18:15:53:36)' with 'EX(23JAN19:12:56:38)'. The report includes a 'Data Set Summary' table, a 'Variables Summary' section, and a 'Listing of Common Variables with Differing Attributes' table.

SDTM COMPARE REPORT FOR THE STUDY:TEST1001

COMPARING EX(21AUG18:15:53:36) WITH EX(23JAN19:12:56:38)

The COMPARE Procedure
Comparison of OLD.EX with NEW.EX
(Method=EXACT)

Data Set Summary

Dataset	Created	Modified	NVar	NObs	Label
OLD.EX	21AUG18:06:23:36	21AUG18:06:23:36	18	42	Exposure
NEW.EX	23JAN19:02:26:38	23JAN19:02:26:38	18	42	Exposure

Variables Summary

Number of Variables in Common: 18.
Number of Variables with Differing Attributes: 1.

Listing of Common Variables with Differing Attributes

Variable	Dataset	Type	Length	Label
EXADJ	OLD.EX	Char	33	Reason for Dose Adjustment
	NEW.EX	Char	36	Reason for Dose Adjustment

Observation Summary

Data differences in Excel Format

- This is an excel output generated by macro, here record level data changes will be displayed.
- Each dataset mismatches will appear in sheets. if there is no matches no output will not be generate.
- Programmer will have to look at this file closely and if there is an issue with any datasets, To know the reason for the issue, user need to check the raw datasets and programs used to generate the datasets.
- Excel file will be named programmatically with dates of old and new datasets. By means of doing this latest output will not overwrite older outputs.

SDTM_compare_23JAN2019_with_21AUG2018_1415.xlsx - Excel

	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	EXDOSE	EXDOSTXT	EXDOSU	EXDOSFRM	EXDOSFRQ	EXROUTE	EXADJ	EPOCH	EXSTDTC	EXENDTC	EXSTDY	EXENDY	
2	80		mg	CAPSULE	QD	ORAL	RE-START DOSING	TREATMENT	2017-02-02T18:10	2017-11-07	221	499	
3	80		mg	CAPSULE	QD	ORAL	OTHER:PT FORGOT TO DOSE ON 01FEB2017	TREATMENT	2017-02-02T18:10	2017-11-07	221	499	
4		.	..		..		XXXXXXXX..X.XXXX.XX.XXXX.XX.XXXXXXXXXX						
5	160		mg	CAPSULE	QD	ORAL	NOT APPLICABLE, FIRST DOSE	TREATMENT	2016-07-11T19:35		1		
6	160		mg	CAPSULE	QD	ORAL	NOT APPLICABLE, FIRST DOSE	TREATMENT	2016-07-11T19:35	2018-11-27	1	870	
7	.		..		..					XXXXXXXXXX			
8	160		mg	CAPSULE	QD	ORAL	NOT APPLICABLE, FIRST DOSE	TREATMENT	2016-08-15T07:15		1		
9	160		mg	CAPSULE	QD	ORAL	NOT APPLICABLE, FIRST DOSE	TREATMENT	2016-08-15T07:15	2018-12-31	1	869	
10	.		..		..					XXXXXXXXXX			
11	120		mg	CAPSULE	QD	ORAL	ADVERSE EVENT (SPECIFY AE NUMBER)	TREATMENT	2017-07-24		260		
12	120		mg	CAPSULE	QD	ORAL	ADVERSE EVENT (SPECIFY AE NUMBER)	TREATMENT	2017-07-24	2018-12-03	260	757	
13	.		..		..					XXXXXXXXXX			
14	160		mg	CAPSULE	QD	ORAL	RE-START DOSING	TREATMENT	2017-05-11	2018-02-04	149	418	
15	160		mg	CAPSULE	QD	ORAL	RE-START DOSING	TREATMENT	2017-05-11	2017-12-21	149	373	

report compare_DM compare_EX compare_TA compare_TE compare_TS

2. MACRO TO COMPRESS/ZIP THE DATASETS

- Sometimes we need to compress or zip the files effectively and post them on any document repository or place them in standard web tools such as pinnacle 21 to validate the datasets. Files can get corrupted and might consume time when they are transferred over the mail. So, file compression is necessary to save space and to reduce time for transmission over mail. It is easier to compress multiple files to one file and attach to mail or upload them on web tool rather than uploading individual file.
- This macro has two parameters, Programmer must provide appropriate values to successful execution
 - `%create_zip(ftype=xpt, path=/sdtm/xpt);`
- Ftype= file type that we need to select to compress. Folder can have multiple types of files such as xpt, xls, xlsx, doc, etc. if you want compress selectively specific type of file then we must give file type to this parameter.
- Path= A valid path where files available to be compressed. Wrong path can issue error message in the log window.

	datasets_24APR19.zip	4/24/2019 1:02 PM
	datasets_24OCT18.zip	10/24/2018 2:38 PM
	datasets_25MAR19.zip	3/25/2019 8:59 PM
	datasets_26FEB19.zip	2/27/2019 12:15 AM
	datasets_27FEB19.zip	2/27/2019 11:50 AM
	datasets_28MAY19.zip	5/28/2019 11:21 PM
	datasets_29MAY19.zip	5/29/2019 8:55 PM
	ae.xpt	6/21/2019 5:47 PM
☒ 	cm.xpt	6/21/2019 5:48 PM
	da.xpt	6/21/2019 5:49 PM
	dm.xpt	6/21/2019 6:01 PM
	ds.xpt	7/9/2019 2:31 PM
	dv.xpt	6/21/2019 6:00 PM

3. MACRO TO CREATE FOLDER AND BACK UP THE DATASETS

- Manually creating folder and placing the data into folder is time consuming activity and this process may end up with error. We can write a macro to create folder with a name date and placing datasets into folder programmatically. The following macro will help us to do this task. Programmer must provide appropriate parameter values in order to execute.
- `%files_backup(copyin=,copyout=,ftype=rtf);`
- Copyin=** Provide the path where the datasets or files are available to be copied.
- Copyout=** Provide a path where copied datasets must be placed. If you want to create a new folder, folder name can be given in the same path. Macro will create the folder and place files into it.
- Ftype=** provide the file extension such as XPT,XLS and XLSX

datasets_24APR19	4/24/2019 1:02 PM
datasets_24OCT18	10/24/2018 2:38 PM
datasets_25MAR19	3/25/2019 8:59 PM
datasets_26FEB19	2/26/2019 11:16 AM
datasets_27FEB19	2/27/2019 10:41 AM
datasets_28MAY19	5/28/2019 11:21 PM
datasets_29MAY19	5/29/2019 12:44 PM
xpt	6/21/2019 11:44 AM
ae.sas7bdat	6/21/2019 5:47 PM
cm.sas7bdat	6/21/2019 5:48 PM
da.sas7bdat	6/21/2019 5:49 PM
dm.sas7bdat	6/21/2019 6:01 PM
ds.sas7bdat	7/9/2019 2:31 PM
dv.sas7bdat	6/21/2019 6:00 PM

Conclusion

- In this paper, the three macros provide an advantage edge for programmers. Dataset comparison will let us know the difference between datasets. This is a best practice to compare datasets. We can reduce the timelines and improve the quality of datasets leading to delivering quality data to client or submitting the data to regulatory agencies.
- This will help us to optimize the process and reduce manual intervention significantly.
- Clinical programmers, especially working on SDTM, AdaM and TLF programming can use these macros to get quality outputs.
- Clinical programmers will get the EDC data on regular basis and they must run the programs regularly to reflect the fresh data into outputs. Using the macros, it will be easy to know the updates by comparing earlier datasets with current datasets.
- Back up of the datasets also can be taken using the program defined here. The goal is to avoid manual programming, reduce timelines and generate quality outputs.

Any Qs??

Author information:

Name: Nagaraju Mancha

Organization: Efficacy Analytics, Bengaluru, India

Email ID: Nagaraju.Mancha@Efficacy.com

Phone No.: 080 4146 3195 / +91 9739522055