PharmaSUG 2020 - Paper AD-142

Data Library Comparison Macro %COMPARE_ALL

Jeffrey Meyers, Mayo Clinic

ABSTRACT

Reproducible research and sharing of data with repositories are becoming more standard, and so the freezing of data for specific analyses is more crucial than ever before. Maintaining multiple data freezes requires knowing what changed within the data from one version to another. In SAS there is the COMPARE procedure that allows the user to compare two data sets to see potential new variables, lost variables, and changes in values. Relying on the COMPARE procedure can be tedious and cumbersome when maintaining a database containing several data sets. The COMPARE_ALL macro was written to ease this burden by generating a Microsoft Excel report of a comparison of two data libraries instead of just two data sets. The report indicates any new or lost data sets, variables or observations and checks for changed data values within all variables. Multiple ID variables can be specified and the macro will determine which variables are relevant with each data set for comparison. The COMPARE_ALL macro is a fantastic tool for managing multiple versions of the same SAS database.

INTRODUCTION

The research field of clinical oncology has many reasons for keeping copies, or freezes, of study data including: standardized reporting such as to the Data Safety Monitoring Board (DSMB), sharing data to a repository such as Project Data Sphere, or when publishing the analysis in an abstract or manuscript. Studies typically have multiple data freezes over several years, and keeping a log of the changes is difficult. The COMPARE Procedure is useful for comparing one data set at a time, but studies often contain many data sets and using the COMPARE Procedure on each one is tedious. The COMPARE_ALL macro was created to make an easy to read report in an Excel that compares entire SAS libraries to check for various changes such as new or lost data sets, new or lost variables, change to variable attributes, and change to variable values. The generated report is a powerful tool for summarizing data changes and ensuring that the data changes are expected.

REPORT FEATURES

The COMPARE_ALL macro produces an Excel report with four different types of tables. The first is an overall summary of the data sets within the base and comparison libraries, the second is a summary of the variables within a specific data set as well as several variable attributes, the third is a summary of the types of data changes within each data set, and the last is a variable level change summary.

TABLE TYPE 1: LIBRARY SUMMARY

This table is only listed once in the report and is always listed first. The table lists each data set in either the base or comparison library in alphabetical order. If a data set exists within both libraries several comparisons are highlighted:

- Date data sets were last updated
- Number of variables
- Number of observations
- Number of variable attributes that changed (e.g. length, type, label)
- Number of lost observations (when an ID variable is specified)
- Number of new observations (when an ID variable is specified)
- Number of data changes (variable value changes for same observation/ID combination)

The table also indicates which ID variables in the ID list are available to be used in comparisons for the given data set. The ID variables used can be different within each data set.

Figure 1 displays the Type 1 Table.

Base Library (FREEZE): /frozen_data/											
Compare Library (LIVE): /live_data/											
Summary of Datasets Compared											
	Base			Compare			Any Differences				
Dataset Name	Last Updated	Number of	Number of Observation	Last Updated	Number of	Number of Observation	Variable Attribute	Lost Observation	New Observation	Data Change	ID Variables Used
AE_MAX_GRADE	02/12/2019	20	16316	11/03/2019	21	12163	0	6369	2216	0	protnum, dcntr_id
BASELINE	02/12/2019	12	40928	11/03/2019	13	38368	0	2560	0	0	protnum, dcntr_id
BIOMARKERS	02/12/2019	5	22494	01/07/2020	7	21206	3	1726	438	9812	protnum, dcntr_id
DZCHAR_PRIORTRT	02/12/2019	19	40928	11/03/2019	20	38368	0	2560	0	11	protnum, dcntr_id
DZ_ASSESS	02/12/2019	27	133619	11/03/2019	28	133471	1	148	0	0	protnum, dcntr_id, merged_day
LABS_BASELINE	02/12/2019	11	35758	11/03/2019	12	35164	0	594	0	0	protnum, dcntr_id
LABS	02/12/2019	23	235656	11/03/2019	24	234760	0	896	0	0	protnum, dcntr_id, visit, study_day
MET_SITES	02/12/2019	27	32029	11/03/2019	27	31266	0	763	0	0	protnum, dcntr_id
OUTCOMES	02/12/2019	37	40081	11/03/2019	20	37766	2	2315	0	5827	protnum, dcntr_id
PRIMARY_SITE	02/12/2019	7	29336	11/03/2019	8	27657	0	1679	0	0	protnum, dcntr_id
PRIOR_CHEMO	02/12/2019	6	28841	11/03/2019	7	18379	0	10462	0	0	protnum, dcntr_id, start_dt, end_dt
PROT_TRT	02/12/2019	11	40928	11/03/2019	13	28395	1	12533	0	24141	protnum, dcntr_id
PROT_TRT_MAINT				11/03/2019	16	4660	0				
PROT_TRT_SEQUENCE				11/03/2019	22	9746	0				
SUBSEQUENT_CHEMO	02/12/2019	16	16357	11/03/2019	17	16327	0	30	0	0	protnum, dcntr_id, start_dt
VITALS	02/12/2019	11	164559	11/03/2019	12	163301	0	1258	0	0	protnum, dcntr_id, visit

Figure 1. Compares the same library of data sets at different data freezes

The report automatically uses colors to highlight data changes and to reference the same items across tables. Attributes for the base library will always be highlighted blue, the compare library will be grey, ID variables will be green, and data differences will be red. In table 1 any changes between base and compare will be highlighted orange to draw the user's attention immediately. The header displays the file paths of both the base and comparison libraries so that the user can ensure the correct libraries are being compared.

TABLE TYPE 2: DATA SET LEVEL SUMMARY

The macro report will include a data set level summary for each data set that exists within both the base and compare libraries. Variables are listed in the same order as they appear in the data set by default, but can be listed alphabetically as well. Each variable indicates if it is an ID variable, has the TYPE, LABEL, FORMAT, and LENGTH attributes listed as well as whether the variable is lost (in base but not in compare), new (in compare but not in base), and for how many observations did the variable's values change for a given ID combination.

Figure 2 displays one of the Type 2 Tables.

Dataset Name: PROT_TRT													
ID Variables: protnum dcntr_id													
Return to Top Summary													
		Base				Compare				Any Differences			
Variable Name	ID Variables?	Type	Label	Format	Length	Type	Label	Format	Length	Lost Variable	New Variable	No. Data Changes	
PROTNUM	Yes	Numeric	Study Name	PROT	8	Numeric	Study Name	PROT	8	No	No		
DCNTR_ID	Yes	Character	Patient ID	$	25	Character	Patient ID	$	25	No	No		
NEW_ID	No	Character				Character	ARCAD ID		25	No	Yes		
LINE_TRIAL	No	Numeric	1st, 2nd, or >=3 Line Study	LINE_TRIA	8	Numeric	1st, 2nd, or >=3 Line Study	LINE_TRIA	8	No	No	0	
STUDY_START_DT	No	Numeric	Study Start Date	MMDDYY	8	Numeric	Study Start Date	MMDDYY	8	No	No	0	
ARM_STRAT_INDEX	No	Numeric	Arm Stratification		8	Numeric	Arm Stratification		8	No	No	1230	
ARM_STRAT_TEXT	No	Character	Arm Stratification (Decode)		40	Character	Arm Stratification (Decode)		100	No	No	22911	
TARGET	No	Numeric	Regimen Includes Any Target Agents?	TARGET	8	Numeric	Regimen Includes Any Target Agents?	TARGET	8	No	No	0	
CHEMO_BACKBONE	No	Character				Character	Chemotherapy Backbone		40	No	Yes		
ANG	No	Numeric	Regimen Includes Any Angiogenic	ANG	8	Numeric	Regimen Includes Any Angiogenic	ANG	8	No	No	0	
EGFR	No	Numeric	Regimen Includes Any Anti-EGFR	EGFR	8	Numeric	Regimen Includes Any Anti-EGFR	EGFR	8	No	No	0	
TRT_DAYS	No	Numeric	Duration (Days) of Protocol Treatment (Excluding Strategy Trials)		8	Numeric	Duration (Days) of Protocol Treatment (Excluding Strategy Trials)		8	No	No	0	
STUDY_END_DT	No	Numeric	Study End Date	MMDDYY	8	Numeric	Study End Date	MMDDYY	8	No	No	0	

Figure 2. Compares a specific data set that both libraries share

Variable names or attributes are highlighted in three cases: if they are an ID variable (green), if an attribute such as typing has changed (orange) and if any of the "Any Differences" columns are not equal to "No" (orange) or if the number of data changes are greater than zero (red). This draws the user's attention to the variables and attributes that are of interest for the difference report. The header lists the data set name and gives a list of the ID variables used for that particular data set.

TABLE TYPE 3: DATA CHANGES SUMMARY

The third table type is optional and displays a summary of the data changes either within a data set grouping by either the first N (specified by *IDSUMTABLE* parameter) ID variable(s) available or across all observations. The summary is useful to see a quick description of all the changes, whether all of the changes are coming from the same study or patient, or if the changes are spread out. This also gives a chance to see if related variables are also changing. For example being able to immediately check if the number of patients having death dates has increased the same amount as the number of patients having a survival status changed from alive to dead.

Figure 3 displays one of the Type 3 Tables.

Summary of PROT_TRT Summary of Data Changes									
All Available ID Variables: protnum dcntr_id									
Note: Only first 1 ID variable(s) is used in summary									
Return to Top Summary									
Study Name	Lost Observations	New Observations	N Data Changes	Variable	Base	Compare	N	Mininum	Maximum
Study 1	6	0	171	arm_strat_text	XELOX	CAPOX	171		
Study 2	0	0	471	arm_strat_text	Cap	Capecitabine	156		
					Cap+Bev	Capecitabine + Bevacizumab	157		
					Cap+Bev+ Mitomycin	Capecitabine + Bevacizumab + Mitomycin	158		
Study 3	0	0	0						
Study 4	825	0							
Study 5	0	0	0						
Study 6	0	0	463	arm_strat_text	BSC	Best Supportive Care	232		
					BSC + panitumumab	Best Supportive Care + Panitumumab	231		

Figure 3. Displays a summary of each variable's changes within the first ID variable (study).

Figure 3 summarizes all of the variable changes within the first ID variable as well as displaying the number of lost observations, new observations, and data changes. For a meta-study such as the one in figure 3, this is valuable to see if the changes are coming from the expected studies or if a programming error could have caused the changes. The summary for each variable changes depending on criteria:

- Comparing two character variables: displays a cross tab of the before and after values with frequencies (see arm_strat_text in figure 3)
- Comparing two numeric variables having less than or equal to macro parameter *CROSSTAB_THRESHOLD*'s unique combinations: displays a cross tab of the before and after values with frequencies
- Comparing two numeric variables having greater than the macro parameter *CROSSTAB_THRESHOLD*'s unique combinations: displays the number of times values changed from missing to non-missing, non-missing to missing, and non-missing to non-missing. There will be a frequency and minimum and maximum change values where appropriate.

The header of the table indicates the data set name and ID variable used to be clear to the user.

Figure 4 displays the Type 3 Table with *IDSUMTABLE*=0.

Summary of BIOMARKERS Summary of Data Changes								
All Available ID Variables: protnum dcntr_id								
Note: No ID variables used in summary								
Return to Top Summary								
Lost Observations	New Observations	N Data Changes	Variable	Base	Compare	N	Mininum	Maximum
1726	438	9812	braf		MT	183		
					WT	1250		
				MT		6		
				WT		66		
			kras		MT	1256		
					WT	2111		
			ras		MT	2381		
					WT	2559		

Figure 4. Displays a summary of each variable's changes within all observations

Figure 4 summarizes all of the variable changes, lost observations, new observations, and data changes across all observations of the data sets. This is a quick way to summarize all variable level changes when there are many changes across numerous ID variable levels.

TABLE TYPE 4: VARIABLE CHANGE SUMMARY

The fourth table summary is similar to the output given by the COMPARE procedure. Each variable that has data changes will have its own worksheet to show individual observation changes.

Figure 5 displays one of the Type 4 Tables.

Dataset Name: PROT_TRT						
Variable Name (label): arm_strat_text (Arm Stratification (Decode))						
Return to Top Summary						
ID Variables						
Study Name	Patient ID	Observation	Base	Compare	Absolute Change	Percent Change
E3200	1	1	FOLFOX4 - bevacizumab	FOLFOX4 + Bevacizumab		
E3200	101	4	bevacizumab	Bevacizumab		
E3200	102	5	FOLFOX4 - bevacizumab	FOLFOX4 + Bevacizumab		
E3200	105	8	FOLFOX4 - bevacizumab	FOLFOX4 + Bevacizumab		
E3200	106	9	FOLFOX4 - bevacizumab	FOLFOX4 + Bevacizumab		
E3200	108	11	FOLFOX4 - bevacizumab	FOLFOX4 + Bevacizumab		
E3200	109	12	FOLFOX4 - bevacizumab	FOLFOX4 + Bevacizumab		
E3200	11	13	bevacizumab	Bevacizumab		

Figure 5. Displays a listing of every observation with a variable that changed value.

The listing will include each ID variable, observation number in the data set, the base data set value, the compare data set value, and the absolute and percent changes if both variables are numeric and non-missing. The header of the table lists the data set, variable name and label. The listing is straightforward and mimics the output of the COMPARE procedure.

NAVIGATING THE REPORT

The report has the potential to create a large amount of worksheets which can make it difficult to navigate quickly to the desired summary. The COMPARE_ALL macro accounts for this by inserting hyperlinks into each table to allow the user to jump to the appropriate summary and back. Examples of the hyperlinks are:

- Within the Type 1 worksheet the user can click the name of any data set to jump to that data set's Type 2 data set summary page.
- Within each Type 2 data set summary page any variable that is marked red for data changes can be clicked on to jump to that variable's type 4 variable listing.
- Each Type 2, 3, and 4 worksheet includes a link in the header to either jump back to the Type 1 worksheet or to drill back up a level (e.g. Type 4 listing back to the Type 2 data set listing)

These hyperlinks allow the user to navigate the report much more efficiently.

CALLING THE COMPARE_ALL MACRO

The COMPARE_ALL macro itself is straightforward available due to the small amount of parameters needed. Using the macro requires access to ODS EXCEL within SAS. There are three required parameters and four optional parameters:

- Required parameters:
    - BASE: Library name to be considered the old data for comparisons. Note that the libname must already be specified prior to the macro.

  - COMPARE: Library name to be considered the new data for comparisons. Note that the libname must already be specified prior to the macro.
  - OUTDOC: The destination filepath and filename of the XLSX file that will contain the report

- Optional parameters:
  - ID: A space delimited list of variable names to be used as ID variables. Note that not all data sets in the same library have to contain all of the ID variables. The macro will match any available ID variable in the list to a given data set in the order they are listed. ID variables not in a given data set are ignored for that comparison.
  - IDSUMTABLE: Determines how many of the available ID variables are used to produce the Table 3 Summary
  - SELECT: Allows the user to specify which data sets in the library are included in the report. The user can specify a space delimited list of the desired data sets. If this option is used then a message is shown in the Type 1 worksheet to indicate that not all of the data sets in the libraries are being shown. The data sets in the SELECT statement must exist in both the BASE and COMPARE libraries.
  - CROSSTAB_THRESHOLD: Determines the threshold for the number of unique variable value changes before summarizing the changes as non-missing and missing to save vertical space. The value defaults to 15 and must be greater than or equal to 1.

The following is the macro call that leads to the images in figure 1 to figure 4 (file paths and study names have been masked for confidentiality):

```
libname live '/live_data/';
libname freeze '/frozen_data/';
options fmtsearch=(live work library);

%compare_all(
    base=freeze,
    compare=live,
    id=protnum dcntr_id merged_day visit study_day start_dt end_dt,
    outdoc=database_changes.xlsx,
     select=,
    idsumtable=1);
```

The libraries are predefined in LIBNAME statements. The FMTSEARCH option is enabled in order to activate the formats of the library in order to make the values in the comparisons make sense. The most complicated parameter is the ID option. There are seven ID variables listed, but a majority of the data sets will only use the first two listed ID variables due to not having the other five. There are several data sets in these libraries where there are multiple rows per patient and times such as MERGED_DAY and START_DT are necessary to compare the appropriate observations. The macro will search for all seven listed ID variables in each data set and subset down to only the variables that exist in both the base and compare data sets.

CONCLUSION

The COMPARE_ALL macro is a powerful tool to efficiently compare two SAS libraries versus running multiple COMPARE Procedure calls for each data set in the libraries. The generated report is clean and straightforward to read with built in navigation for ease of use. The COMPARE_ALL macro will be useful for any programmer working with clinical trial data and is available to be shared.

CONTACT INFORMATION

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

Name: Jeffrey Meyers
Enterprise: Mayo Clinic
Address: 200 First Street SW
City, State ZIP: Rochester, MN 55905
Work Phone: 507-266-2711
E-mail: Meyers.jeffrey@mayo.edu / jpmeyers.spa@gmail.com
Website (Macro Download Available): SAS Communities Page