Data Library Comparison Macro %COMPARE_ALL

Jeffrey Meyers, Mayo Clinic:

Statistical Programmer Analyst III within Mayo Clinic's Cancer Center Statistics team.

Research focus on gastrointestinal and respiratory cancers as well as meta-analysis.

10 years of SAS programming experience focusing on macros, graphics, SQL and reports.

Data Library Comparison Macro %COMPARE_ALL

Jeffrey Meyers, Mayo Clinic

%COMPARE_ALL

- Creates a report comparing all datasets within two libraries
  - Outputs to an Excel file (requires SAS 9.4+)
- 8 parameters (3 required, 5 optional)
- Creates four different types of worksheets in the report
- Contains navigation to easily move from one summary to another
- Contains error checking, documentation, and cleans up after itself

Macro Parameters

- Required
  - BASE: designates the first library (libname not path) to be compared
  - COMPARE: designates the second library for comparison
  - OUTDOC: full file path and file name of output Excel document
- Optional
  - ID: designate one or more variables as unique identifiers (example: Patient ID). Each dataset takes any ID variables that in this list that exist within it. They are sorted by the ID variables from left to right in the list.
  - SELECT: optionally selects specific datasets from libraries for comparison
  - CROSSTAB_THRESHOLD: cut-point for displaying all unique values for a variable in comparison table
  - IDSUMTABLE: determines how many ID variables are used for change summary table
  - DEBUG: turns on options for debugging macro issues

Macro Call Example

```
Libname frz18 ‘file-path 2018 version’;
Libname frz19 ‘file-path 2019 version’;

%COMPARE_ALL(
    BASE=frz18,
    COMPARE=frz19,
    OUTDOC=compare_all_example.xlsx,
    ID=patient_id cycle);
```

ID Parameter

Example: ID=patient_id cycle eval_dt

Dataset 1: Includes variables patient_id, age, sex, performance_score

Dataset 2: Includes variables patient_id, cycle, agent, dose

Dataset 3: Includes variables patient_id, eval_dt, toxicity, grade

ID Parameter

Example: ID=patient_id cycle eval_dt

Dataset 1: Includes variables patient_id, age, sex, performance_score

Uses only patient_id as ID variable

Dataset 2: Includes variables patient_id, cycle, agent, dose

Uses patient_id and cycle as ID variables in that order

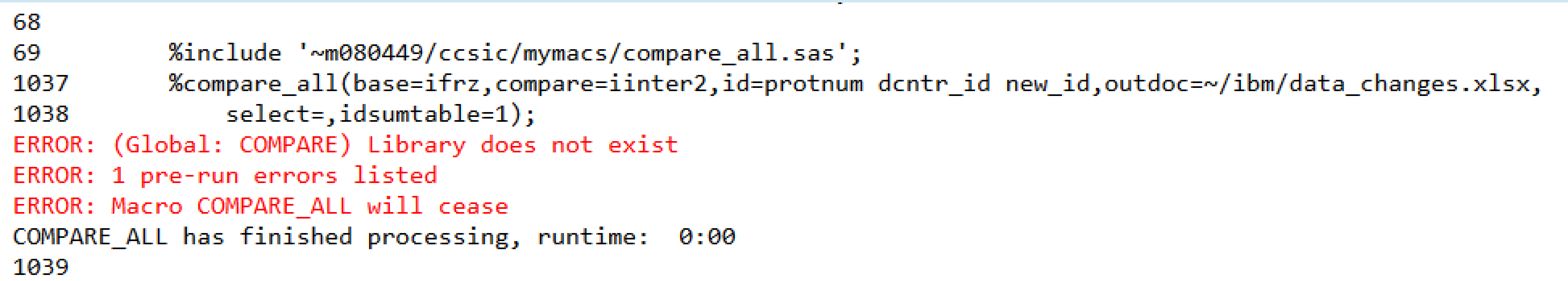
Dataset 3: Includes variables patient_id, eval_dt, toxicity, grade

Uses patient_id and eval_dt as ID variables in that order

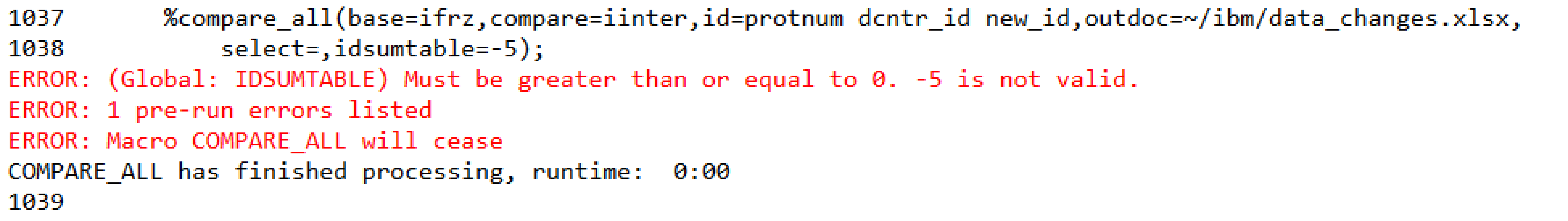
Macro Error Checking

- The following items are checked by the macro:
  - If the *BASE* and *COMPARE* libraries exist and have been assigned
  - If OUTDOC is missing
  - If the *CROSSTAB_THRESHOLD* and *IDSUMTABLE* parameters are not a number greater than 0

Macro Error Checking



```
68
69         %include '~m080449/ccsic/mymacs/compare_all.sas';
1037       %compare_all(base=ifrz,compare=iinter2,id=protnum dcntr_id new_id,outdoc=~/ibm/data_changes.xlsx,
1038           select=,idsumtable=1);
ERROR: (Global: COMPARE) Library does not exist
ERROR: 1 pre-run errors listed
ERROR: Macro COMPARE_ALL will cease
COMPARE_ALL has finished processing, runtime:  0:00
1039
```

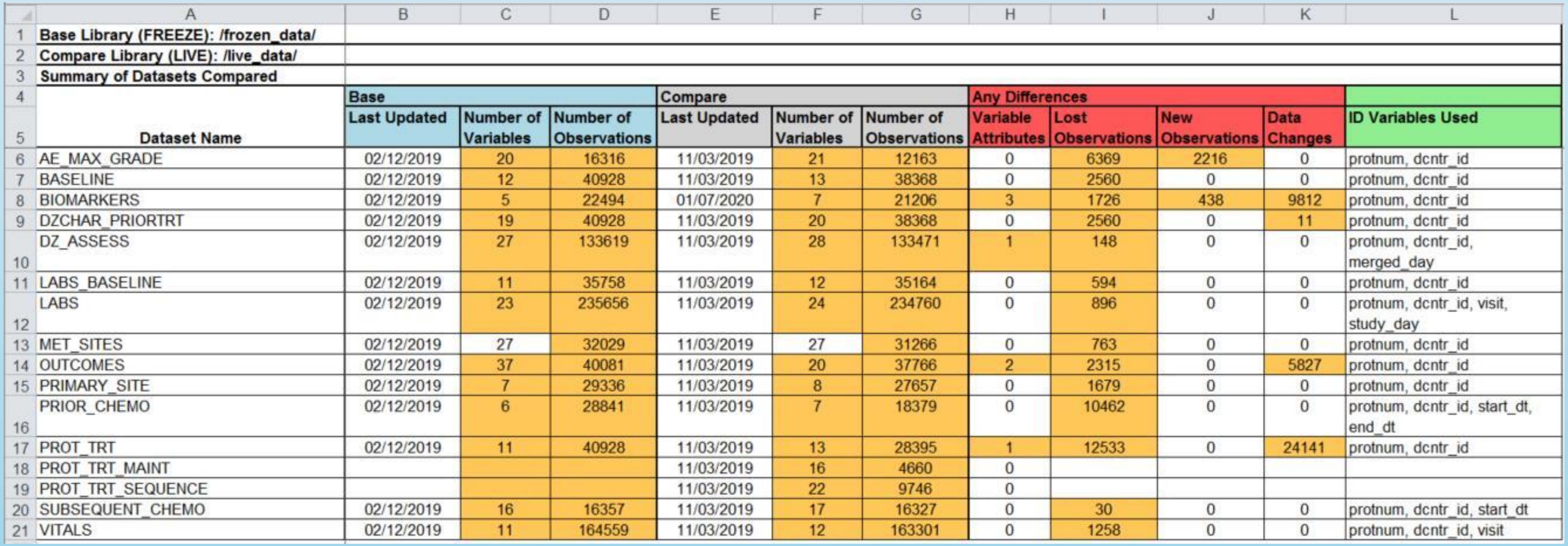


```
1037       %compare_all(base=ifrz,compare=iinter,id=protnum dcntr_id new_id,outdoc=~/ibm/data_changes.xlsx,
1038           select=,idsumtable=-5);
ERROR: (Global: IDSUMTABLE) Must be greater than or equal to 0. -5 is not valid.
ERROR: 1 pre-run errors listed
ERROR: Macro COMPARE_ALL will cease
COMPARE_ALL has finished processing, runtime:  0:00
1039
```

Top Summary Worksheet

- Overview of all datasets in either library
  - If SELECT parameter is used then only the datasets in the SELECT list will be shown
- Meta data is shown for both BASE and COMPARE versions
  - Date last updated, number of observations, number of variables
- High-level summary of differences
  - How many variables have had an attribute such as type changed
  - How many observations (based on matching ID variables) are in BASE but not COMPARE
  - How many observations (based on matching ID variables) are in COMPARE but not BASE
  - How many total variable values changed (based on matching ID variables)
- ID Variables Used for each dataset’s comparison
- Libraries being compared are listed at top of sheet with file path
- Clicking a dataset name will navigate to that dataset’s summary page
- BASE, COMPARE, Differences, and ID variables are colored consistently throughout report
- Differences are highlighted with color

Top Summary Worksheet

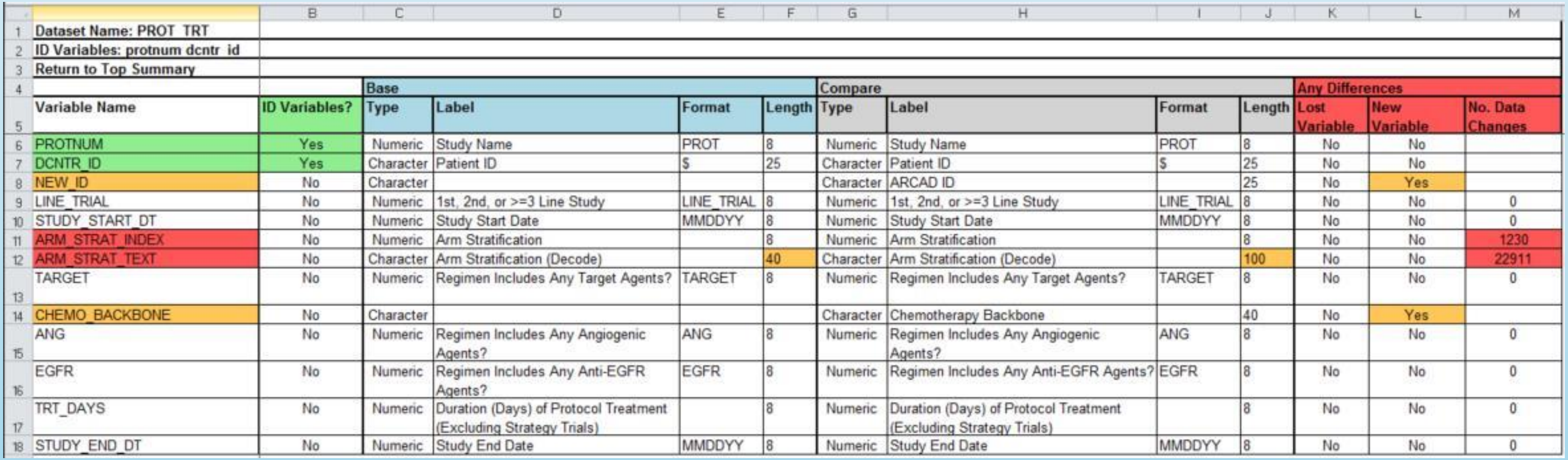


	A	B	C	D	E	F	G	H	I	J	K	L
1	**Base Library (FREEZE): /frozen_data/**											
2	**Compare Library (LIVE): /live_data/**											
3	**Summary of Datasets Compared**											
4		**Base**			**Compare**			**Any Differences**				
5	**Dataset Name**	**Last Updated**	**Number of Variables**	**Number of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

Dataset Summary Worksheet

- Overview of variables in each dataset that exists in both libraries
- Displays basic meta data for each variable
  - Type, length, label and format
  - Highlights any meta data that has changed
- Marks each variable being used as an ID variable in green
- Summary of differences by variable
  - Whether variable exists in BASE but not COMPARE
  - Whether variable exists in COMPARE but not BASE
  - Whether any observations (based on matching ID variables) have changed values
- Any variables with differences are marked in red
  - Clicking a red variable will navigate to that variable’s summary page
- Navigation to top summary worksheet is in the header

Dataset Summary Worksheet

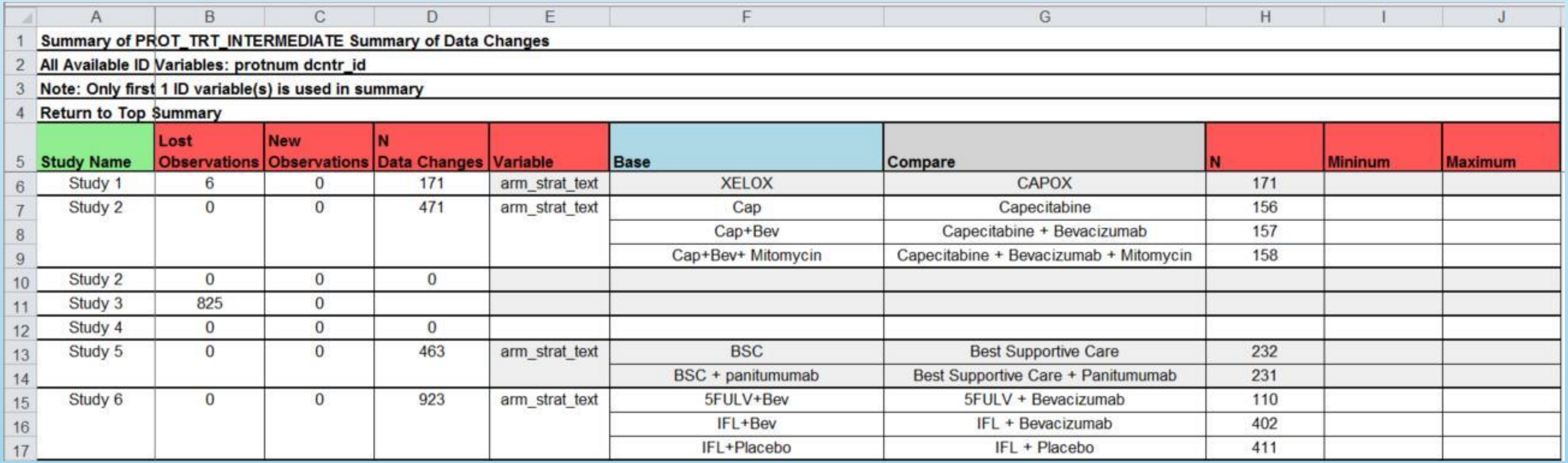


		B	C	D	E	F	G	H	I	J	K	L	M
1	Dataset Name: PROT_TRT												
2	ID Variables: protnum dcntr_id												
3	Return to Top Summary												
4			Base				Compare				Any Differences		
5	Variable Name	ID Variables?	Type	Label	Format	Length	Type	Label	Format	Length	Lost Variable	New Variable	No. Data Changes
6	PROTNUM	Yes	Numeric	Study Name	PROT	8	Numeric	Study Name	PROT	8	No	No	
7	DCNTR_ID	Yes	Character	Patient ID	$	25	Character	Patient ID	$	25	No	No	
8	NEW_ID	No	Character				Character	ARCAD ID		25	No	Yes	
9	LINE_TRIAL	No	Numeric	1st, 2nd, or >=3 Line Study	LINE_TRIAL	8	Numeric	1st, 2nd, or >=3 Line Study	LINE_TRIAL	8	No	No	0
10	STUDY_START_DT	No	Numeric	Study Start Date	MMDDYY	8	Numeric	Study Start Date	MMDDYY	8	No	No	0
11	ARM_STRAT_INDEX	No	Numeric	Arm Stratification		8	Numeric	Arm Stratification		8	No	No	1230
12	ARM_STRAT_TEXT	No	Character	Arm Stratification (Decode)		40	Character	Arm Stratification (Decode)		100	No	No	22911
13	TARGET	No	Numeric	Regimen Includes Any Target Agents?	TARGET	8	Numeric	Regimen Includes Any Target Agents?	TARGET	8	No	No	0
14	CHEMO_BACKBONE	No	Character				Character	Chemotherapy Backbone		40	No	Yes	
15	ANG	No	Numeric	Regimen Includes Any Angiogenic Agents?	ANG	8	Numeric	Regimen Includes Any Angiogenic Agents?	ANG	8	No	No	0
16	EGFR	No	Numeric	Regimen Includes Any Anti-EGFR Agents?	EGFR	8	Numeric	Regimen Includes Any Anti-EGFR Agents?	EGFR	8	No	No	0
17	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
18	STUDY_END_DT	No	Numeric	Study End Date	MMDDYY	8	Numeric	Study End Date	MMDDYY	8	No	No	0

Data Changes Summary Worksheet

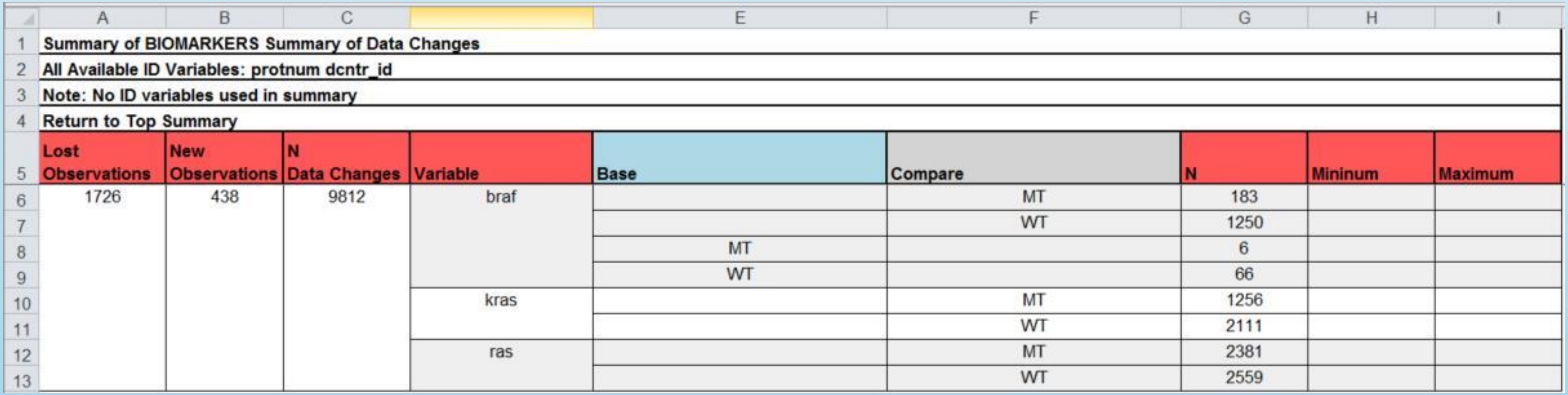
- Summary of changes within dataset, follows Dataset Summary worksheet
- Summarizes lost observations, new observations, and number of data changes across specified ID variables
  - Maximum number of ID variables used is determined by IDSUMTABLE
- Each variable with data changes is listed along with a summary based on variable type and CROSSTAB_THRESHOLD
  - If number of unique changes <=CROSSTAB_THRESHOLD then each unique change is listed along with a count
  - If number of unique changes > CROSSTAB_THRESHOLD then values are summarized as “Non-missing” or “Missing” with counts
  - Numeric variables also have minimum and maximum changed value
- Navigation to top summary and dataset summary worksheet is in the header

Data Changes Summary Worksheet



	A	B	C	D	E	F	G	H	I	J
1	Summary of PROT_TRT_INTERMEDIATE Summary of Data Changes									
2	All Available ID Variables: protnum dcntr_id									
3	Note: Only first 1 ID variable(s) is used in summary									
4	Return to Top Summary									
5	**Study Name**	**Lost Observations**	**New Observations**	**N Data Changes**	**Variable**	**Base**	**Compare**	**N**	**Mininum**	**Maximum**
6	Study 1	6	0	171	arm_strat_text	XELOX	CAPOX	171		
7	Study 2	0	0	471	arm_strat_text	Cap	Capecitabine	156		
8						Cap+Bev	Capecitabine + Bevacizumab	157		
9						Cap+Bev+ Mitomycin	Capecitabine + Bevacizumab + Mitomycin	158		
10	Study 2	0	0	0						
11	Study 3	825	0							
12	Study 4	0	0	0						
13	Study 5	0	0	463	arm_strat_text	BSC	Best Supportive Care	232		
14						BSC + panitumumab	Best Supportive Care + Panitumumab	231		
15	Study 6	0	0	923	arm_strat_text	5FULV+Bev	5FULV + Bevacizumab	110		
16						IFL+Bev	IFL + Bevacizumab	402		
17						IFL+Placebo	IFL + Placebo	411		

Data Changes Summary Worksheet



Summary of BIOMARKERS Summary of Data Changes

All Available ID Variables: protnum dcntr_id

Note: No ID variables used in summary

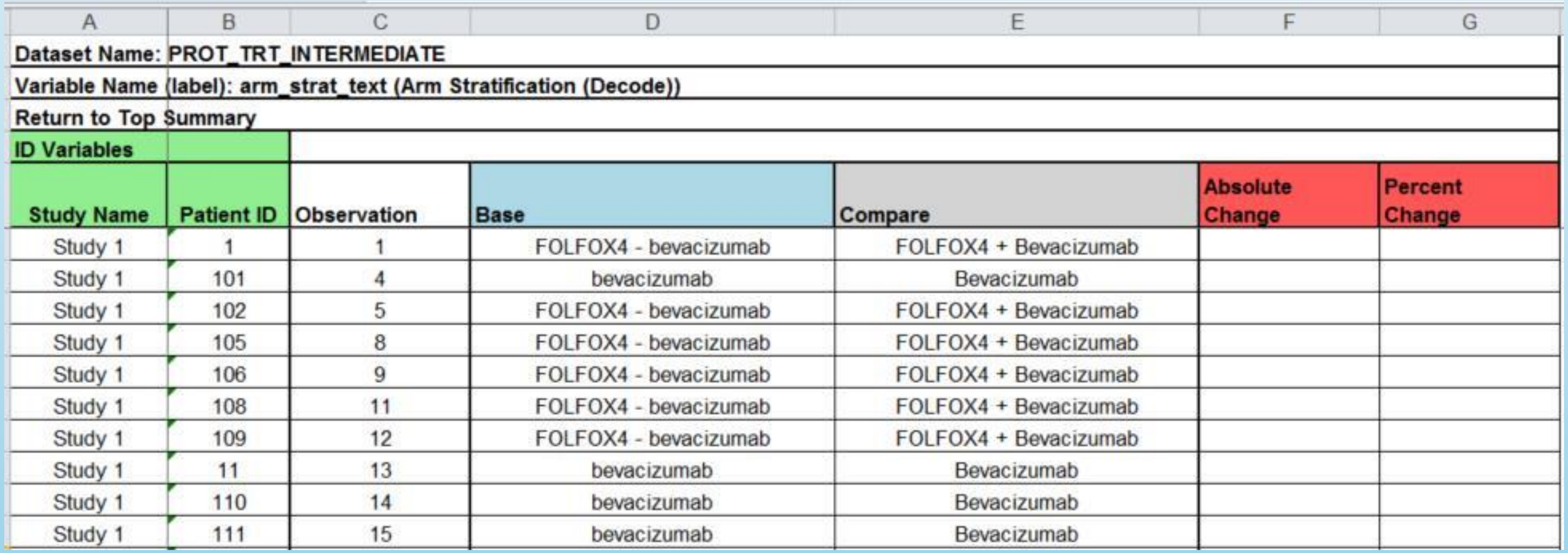
Return to Top Summary

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		

Variable Changes Summary Worksheet

- Created for each variable that had any value changes
- Replicates the output from traditional COMPARE procedure
  - Displays observations with data changes
  - Lists each ID variable, base and compare values, and absolute/percent change if the variable is numeric
- Navigation to top summary and dataset summary worksheet is in the header

Variable Changes Summary Worksheet



A	B	C	D	E	F	G
Dataset Name:	PROT_TRT_INTERMEDIATE					
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
Study 1	1	1	FOLFOX4 - bevacizumab	FOLFOX4 + Bevacizumab		
Study 1	101	4	bevacizumab	Bevacizumab		
Study 1	102	5	FOLFOX4 - bevacizumab	FOLFOX4 + Bevacizumab		
Study 1	105	8	FOLFOX4 - bevacizumab	FOLFOX4 + Bevacizumab		
Study 1	106	9	FOLFOX4 - bevacizumab	FOLFOX4 + Bevacizumab		
Study 1	108	11	FOLFOX4 - bevacizumab	FOLFOX4 + Bevacizumab		
Study 1	109	12	FOLFOX4 - bevacizumab	FOLFOX4 + Bevacizumab		
Study 1	11	13	bevacizumab	Bevacizumab		
Study 1	110	14	bevacizumab	Bevacizumab		
Study 1	111	15	bevacizumab	Bevacizumab		

Conclusion

- The COMPARE_ALL macro is a powerful tool for comparing multiple versions of the same library

- Creates an easy to read Excel report with built in navigation to easily jump to the needed worksheet

- The macro is available for download on the SAS Communities page

Name: Jeffrey Meyers
Organization: Mayo Clinic
E-mail: meyers.jeffrey@mayo.edu / jpmeyers.spa@gmail.com