

## Butterfly Plot for Comparing Two Treatment Responses

Raghava Pamulapati, Liao Jane, Merck & Co., Inc.

### ABSTRACT

Butterfly plot is an effective graphical representation for comparing two treatment responses for the same subject across different time points. Butterfly plots are drawn across a centered axis to separate two treatment responses in one picture.

Butterfly plots accommodate a subject's two treatment responses in one plot, such as, current treatment response with prior study/non-study treatment response or active treatment response with control treatment response or mono therapy response with combination therapy response.

Butterfly plots can be created using SAS® PROC SGPLOT. The inclusion of the HBAR statement creates horizontal bars on the Y-axis. Each bar represents one subject. The RESPONSE option displays duration of study medications on the X-axis. The FILLATTRS option categorizes the response in colors by using variables that are derived for each treatment and response type with corresponding description displayed at plot legend using KEYLEGENT statement. More specific details on the PROC SGPLOT syntax and plot options will be presented in the body of the paper. Furthermore, steps to derive required variables and dataset pre-processing to categorize response will be discussed.

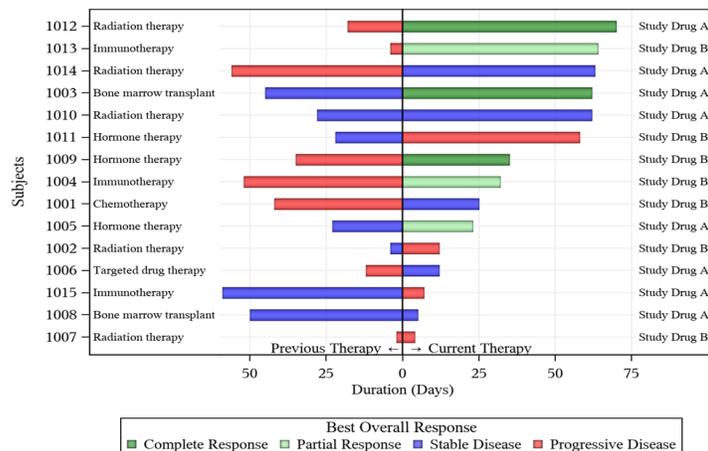
### KEYWORDS

PROC SGPLOT, PROC TEMPLATE, HBAR and DOT STATEMENTS, SGANNO, FILLATTRS and ODS IMAGENAME OPTIONS, ADRS, ADSL, BEST OVERALL RESPONSE and TREATMENT DURATION.

### INTRODUCTION

Most of the popular graphs in clinical research are designed to display subject single treatment response. In some studies, subject may receive study or non-study treatment before taking current treatment. Using Butterfly plot, we can display subjects' two different treatment response in one plot to make it easier for reviewers to compare two treatment responses.

Below example plot developed using subject's previous non-study treatment duration and response, current study period duration and response. More details will be discussed in later sections.



Butterfly plot can be developed using two methods, SAS® PROC SGPLOT and SAS® PROC TEMPLATE with PROC SGRENDER. Both methods produce butterfly plots with all features that we are going to discuss in this paper. One of the main differences between using SAS PROC SGPLOT and

PROC TEMPLATE is, SGPLOT does not allow more than one grouping variable, whereas in butterfly plot user is required to pass two grouping variables, each grouping variable represents one treatment response. PROC TEMPLATE, however, allows multiple grouping variables. Due to the limitation of using multiple grouping variables in SGPLOT, preprocessing is required that includes transforming the data into horizontal structure to avoid using multiple grouping variables.

This paper primarily focuses on developing butterfly plots using PROC SGPLOT with different options. PROC TEMPLATE is also briefly introduced.

Besides displaying two treatments' duration and response, butterfly plot also displays subject overall responses for each response assessment, study milestones like death or subject medication status in either or both treatment periods. In order to display this information, the corresponding ADaM dataset must have this this information populated.

## GATHERING INFORMATION IN ADaM DATASETS

Basic butterfly plots require subject level information present in baseline ADaM datasets like ADSL or ADBASE. If a Butterfly plot is intended to display overall response in each response assessment, then the corresponding information must be available in the ADaM datasets.

### REQUIRED INFORMATION

Basic butterfly plots display two pieces of information

- Subject's best overall response for two treatment periods.
- Treatment duration for those two periods.

Best overall response is collected at the subject level and depending on the type of treatment the data can be located in different domains. For instance, if the desired plot is for comparison with previous treatment, this information should be in baseline ADaM datasets like ADSL or ADBASE.

SUBJID	PBORINV	PDURN
1001	PD	42
1002	SD	4
1003	SD	45
1004	PD	52
1005	SD	23

In above example, PBORINV is subject previous treatment period best overall response and PDURN is previous treatment period duration in days. Previous treatment duration is calculated based on time since first previous treatment date to last date of the previous treatment.

The study treatment(s)' best overall response is usually present in the ADRS dataset along with overall responses. In below example, subject current treatment period best overall confirmed response is presented in "BORCFINV" parameter. TRTEDY column from ADSL is used as duration of current treatment in days.

SUBJID	PARAMCD	AVAL	SUBJID	TRTEDY
1001	BORCFINV	SD	1001	70
1002	BORCFINV	PD	1002	47
1003	BORCFINV	CR	1003	75
1004	BORCFINV	PR	1004	46
1005	BORCFINV	PR	1005	54

Note: Duration of response columns PDURN and TRTEDY are derived in days in both datasets. Based on the scale of X-axis for plot, the duration can be recalculated in weeks/months/years according to the requirement.

## OPTIONAL INFORMATION

- Subject's overall response in each response assessment and time to response.
- Other subject level labeling.

### 1. Subjects' overall response at each evaluation

Displaying overall response provides insight of how subject is responding to the treatment over a given period. Swimmer plots are ideal to display subject overall response in each period. Using same approach, butterfly plot can also display the overall response in both treatment or treatment of interest.

Response datasets like ADRS contains subject study treatment overall response. Below ADRS contains subjects' overall response at each visit in current treatment period with relative day since first treatment date.

SUBJID	PARAMCD	AVAL	ADY
1001	ORINV	PD	21
1001	ORINV	SD	42
1001	ORINV	SD	63
1002	ORINV	SD	25
1002	ORINV	PD	46
1003	ORINV	SD	21
1003	ORINV	PR	43
1003	ORINV	CR	69
1004	ORINV	SD	26
1004	ORINV	PR	39
1005	ORINV	SD	21
1005	ORINV	PR	49

### 2. Subject level or treatment level information

Labelling the subjects is very important to give more insight into subject details. Most required label information is, subject number, treatment, cohort, dose or route of intake. Butterfly plot have the flexibility to display labels inside and outside of the plot on Y-axis.

Following example is from ADSL dataset. PTRT indicates subject previous treatment information, TRT01A has current treatment information and EOTSTT is current treatment status.

SUBJID	PTRT	TRT01A	EOTSTT	TRTEDY
1001	Chemotherapy	Study Drug B	Ongoing	70
1002	Radiation therapy	Study Drug B	Discontinued	47
1003	Bone marrow transplant	Study Drug A	Ongoing	75
1004	Immunotherapy	Study Drug B	Ongoing	46
1005	Hormone therapy	Study Drug A	Ongoing	54
1006	Targeted drug therapy	Study Drug A	Discontinued	63
1007	Radiation therapy	Study Drug B	Discontinued	20
1008	Bone marrow transplant	Study Drug A	Ongoing	56
1009	Hormone therapy	Study Drug B	Discontinued	79
1010	Radiation therapy	Study Drug A	Discontinued	31
1011	Hormone therapy	Study Drug B	Discontinued	24
1012	Radiation therapy	Study Drug A	Ongoing	74
1013	Immunotherapy	Study Drug B	Ongoing	48
1014	Radiation therapy	Study Drug A	Discontinued	46
1015	Immunotherapy	Study Drug A	Discontinued	26

## PREPROCESSING

Preprocessing includes merging all related information into one dataset and formatting the dataset specific for SGPLOT or TEMPLATE procedures.

As discussed at the beginning of the paper, butterfly plot can be created using two different methods, PROC TEMPLATE and PROC SGPLOT. Butterfly plot displays two response information and it is required to use two grouping columns, whereas SGPLOT requires consistent grouping variable for all HBAR statements in the plot. If more than one grouping variable used in HBAR statement, then report will not be produced with below error message.

```

ERROR: The same category variable must be used for summarized plots.
NOTE: The SAS System stopped processing this step because of errors.
NOTE: There were 5 observations read from the data set WORK.ATTRMAP.
NOTE: PROCEDURE SGPLOT used (Total process time):
      real time          0.00 seconds
      cpu time           0.01 seconds

```

To avoid using multiple grouping variables, HBAR needs to run for each treatment and response. In order to run for each treatment and response, input ADaM dataset needs to be transformed into a horizontal structure and add default or user specified colors to each response using FILLATTRS option and justify this response in legend using KEYLEGEND statement.

## Converting left side treatment duration into negative values

Butterfly plot cannot display two response duration columns along X-axis in different direction from 0 if both duration values are in positive format. Technically SAS allows only negative values in left side of the plot. Hence, we need to derive left side duration values into negative values in ADaM level or temporarily convert positive values into negative values in preprocessing. However, at plot development section we will apply picture format to display left side treatment duration in positive values to avoid confusion.

SUBJID	LDUR1	LDUR2	LDUR3	LDUR4	RDUR1	RDUR2	RDUR3	RDUR4
1001	.	.	.	-42	.	.	25	.
1002	.	.	-4	.	.	.	.	12
1003	.	.	-45	.	62	.	.	.
1004	.	.	.	-52	.	32	.	.
1005	.	.	-23	.	.	23	.	.
1006	.	.	.	-12	.	.	12	.
1007	.	.	.	-2	.	.	.	4
1008	.	.	-50	.	.	.	5	.
1009	.	.	.	-35	35	.	.	.
1010	.	.	-28	.	.	.	62	.
1011	.	.	-22	.	.	.	.	58
1012	.	.	.	-18	70	.	.	.
1013	.	.	.	-4	.	64	.	.
1014	.	.	.	-56	.	.	63	.
1015	.	.	-59	.	.	.	.	7

### Displaying overall response optional item

Like above step, if overall response needs to be displayed then overall response also needs to be transformed into horizontal structure to create one column for each treatment and response.

Below table is the final dataset contains one column per treatment and response. Depending on the maximum number of each response, corresponding columns will be created. In below example, subject 1001 has two SD values which is the maximum number of SDs per subject in overall dataset, hence total two SD columns has created SD1-SD2. Below dataset derived using only current treatment overall response from ADRS dataset along with best overall response duration columns from both treatments.

SUBJID	CR1	CR2	PD1	PD2	PR1	PR2	SD1	SD2	LDUR1	LDUR2	LDUR3	LDUR4	RDUR1	RDUR2	RDUR3	RDUR4
1001	.	.	21	.	.	.	42	63	.	.	.	-42	.	.	70	.
1002	.	.	46	.	.	.	25	.	.	.	-4	.	.	.	.	47
1003	69	.	.	.	43	.	21	.	.	.	-45	.	75	.	.	.
1004	.	.	.	.	39	.	26	.	.	.	-52	.	46	.	.	.
1005	.	.	.	.	49	.	21	.	.	.	-23	.	54	.	.	.
1006	.	.	.	.	.	.	20	51	.	.	.	-12	.	.	63	.
1007	.	.	20	.	.	.	.	.	.	.	-2	.	.	.	.	20
1008	.	.	.	.	.	.	26	51	.	.	-50	.	.	.	56	.
1009	46	73	.	.	.	.	32	.	.	.	-35	79	.	.	.	.
1010	.	.	.	.	.	.	26	.	.	.	-28	.	.	.	31	.

### DATA preparation for PROC TEMPLATE with PROC SGRENDER

On other hand, using PROC TEMPLATE, we can have multiple grouping variable, thus it's not required to transpose the data. PROC TEMPLATE is very flexible, giving many options to customize the plot. Using TEMPOUT option in PROC SGPLOT we can get basic PROC TEMPLATE code and modify the template code based on requirement.

Two steps are needed to process data for PROC TEMPLATE

1. Merge ADSL and ADRS data by SUBJID
2. Populate data as desired (the data in red box).
  - For best overall responses (i.e. Best overall responses for Prior/Post treatment), set value for one observation only,
  - For other overall responses set values for those observations (i.e. PARAMCD='ORINV')

Here is the example dataset after process and ready to use by PROC TEMPLATE

subjid	paramcd	aval	ady	paval	pady	baval	bady
1001	BORCFINV	SD		PD	-42	SD	70
1001	ORINV	PD	21				
1001	ORINV	SD	42				
1001	ORINV	SD	63				
1002	BORCFINV	PD		SD	-4	PD	47
1002	ORINV	SD	25				
1002	ORINV	PD	46				
1003	BORCFINV	CR		SD	-45	CR	75
1003	ORINV	SD	21				
1003	ORINV	PR	43				
1003	ORINV	CR	69				
1004	BORCFINV	PR		PD	-52	PR	46
1004	ORINV	SD	26				
1004	ORINV	PR	39				

## DEVELOPING PLOT

### ANNOTATION DATASET

It is recommended to annotate the plot by displaying the treatment period information inside or outside the plot. If the plot creating by PROC TEMPLATE then DRAWTEXT option can replace annotation dataset, whereas PROC SGPLOT required to have annotation dataset. Below annotation dataset is created to display the treatment period labels inside the plot and duration denominator outside the plot. Previous Therapy label display on left side bottom, Current Therapy label display on right side bottom and Duration(days) label display outside center of the plot.

LABEL	ANCHOR	FUNCTION	X1SPACE	Y1SPACE	WIDTH	X1	Y1
Previous Therapy ("ESC"){unicode '2190x}	right	text	datavalue	wallpercent	100000	0	2
("ESC"){unicode '2192x} Current Therapy	left	text	datavalue	wallpercent	100000	0	2
Duration (Days)	center	text	datavalue	wallpercent	100000	0	-10

### APPLYING FORMAT

Since we have derived negative duration response for previous treatment, we can apply below picture format to show previous treatment duration values in positive.

```
proc format;
  picture positive low-<0="000,000"
  0<-high="000,000";
run;
```

### RUNNING SGPLOT

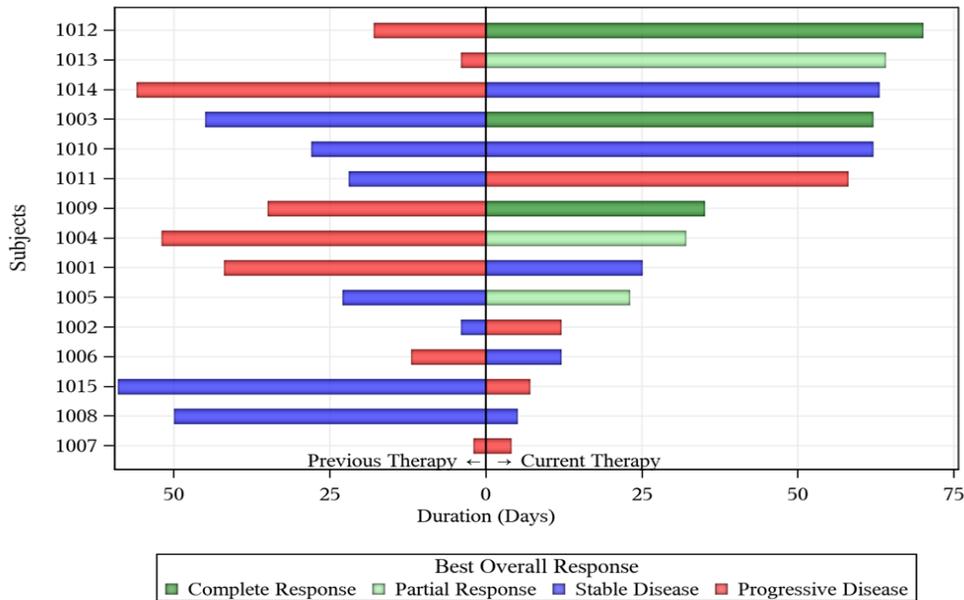
Below SGPLOT creates Butterfly plot using custom colors. HBAR statement runs for each response variable created in preprocessing section. Annotation dataset "sgannods" will be passed into SGPLOT using SGANNO option. KEYLEGEND display unique response values at plot legend section using LDUR1-RDUR4 labels. YAXIS LABEL option display SUBJID numbers for each bar in Y-axis.

SAS applies default colors based on the order of the data. We can also specify user defined colors using FILLATTRS to fill the colors inside the bar and OUTLINEATTRRS to select the color of bar lines. Below plots we have both color and black and white shades. Also, using DATASKIN option to change the visual effects to the plot. In below example "pressed" option has used for DATASKIN.

For colored plots used Dark Green, Light Green, Blue and Red colors for Complete Response, Partial Response, Stable Disease and Progressive Disease respectively for both treatment responses.

```
ods graphics / reset=all noborder imagename="butterfly" width= 7.0in
                height= 5.0in;

proc sgplot data=og0stemleaf0plot sganno=sgannods;
  format ldur: positive.;
  hbar id / response=ldur1 barwidth=.5 transparency=.25 name='a1'
  fillattrs=(color=Green) dataskin =pressed;
  hbar id / response=ldur2 barwidth=.5 transparency=.25 name='a2'
  fillattrs=(color=LightGreen) dataskin =pressed;
  hbar id / response=ldur3 barwidth=.5 transparency=.25 name='a3'
  fillattrs=(color=Blue) dataskin =pressed;
  hbar id / response=ldur4 barwidth=.5 transparency=.25 name='a4'
  fillattrs=(color=Red) dataskin =pressed;
  hbar id / response=rdur1 barwidth=.5 transparency=.25 name='a1'
  fillattrs=(color=Green) dataskin =pressed;
  hbar id / response=rdur2 barwidth=.5 transparency=.25 name='a2'
  fillattrs=(color=LightGreen) dataskin =pressed;
  hbar id / response=rdur3 barwidth=.5 transparency=.25 name='a3'
  fillattrs=(color=Blue) dataskin =pressed;
  hbar id / response=rdur4 barwidth=.5 transparency=.25 name='a4'
  fillattrs=(color=Red) dataskin =pressed;
  keylegend "a1" "a2" "a3" "a4" / across=4 position=bottom border
  location=outside
  title = "Best Overall Response";
  xaxis label=" " grid;
  yaxis label="Subjects" grid discreteorder=data;
run;
```



Above report will be produced after executing the Butterfly plot basic syntax. Each bar represents one

subject. Durations in X-Axis calculated using corresponding treatment start date to last dosing dates. For Previous Therapy, x=0 means the starting point of most recent prior therapy; for Current Therapy, x=0 means the starting point of the study therapy

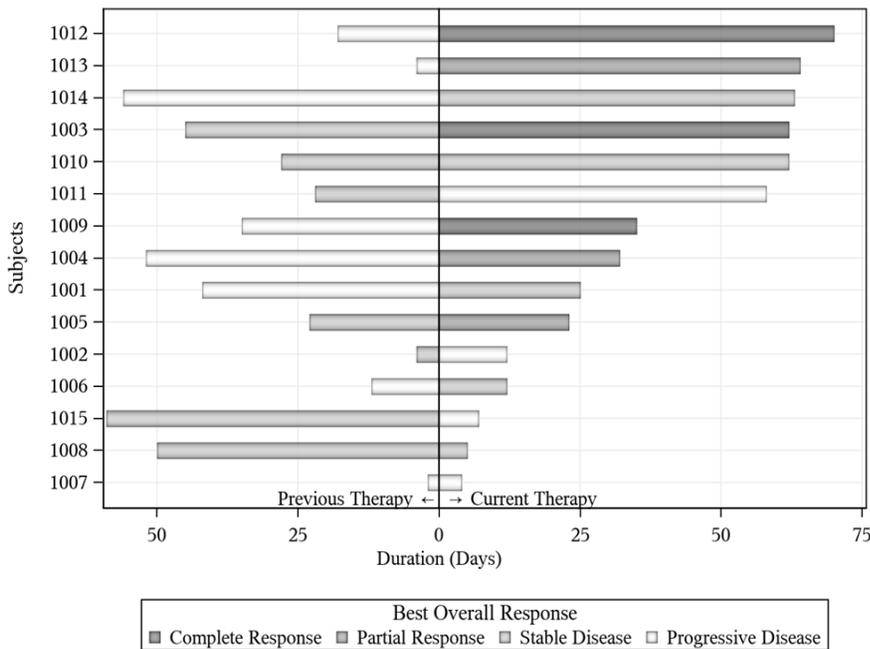
Please note, when passing response variable into HBAR make sure to have at least one non-missing record in each response column otherwise SAS produce a warning message.

```
WARNING: Y=_SUM1_LDUR1_ is invalid. The option expects that the column not contain all missing values.
WARNING: Y=_SUM2_LDUR2_ is invalid. The option expects that the column not contain all missing values.
```

Above warning message indicating that LDUR1 and LDUR2 have all blank values.. LDUR1 and LDUR2 are intermediate duration columns for previous therapy “CR” and “PR” response respectively. Since subjects are not responded to previous therapy these columns having blank values.

In order to avoid this warning message, it’s recommended to execute HBAR statement for individual response column when at least one non-missing value is present.

Similarly, black white plot below is created using Dark gray, Medium gray, Light gray and white shades for Complete Response, Partial Response, Stable Disease and Progressive Disease respectively for both treatment periods.

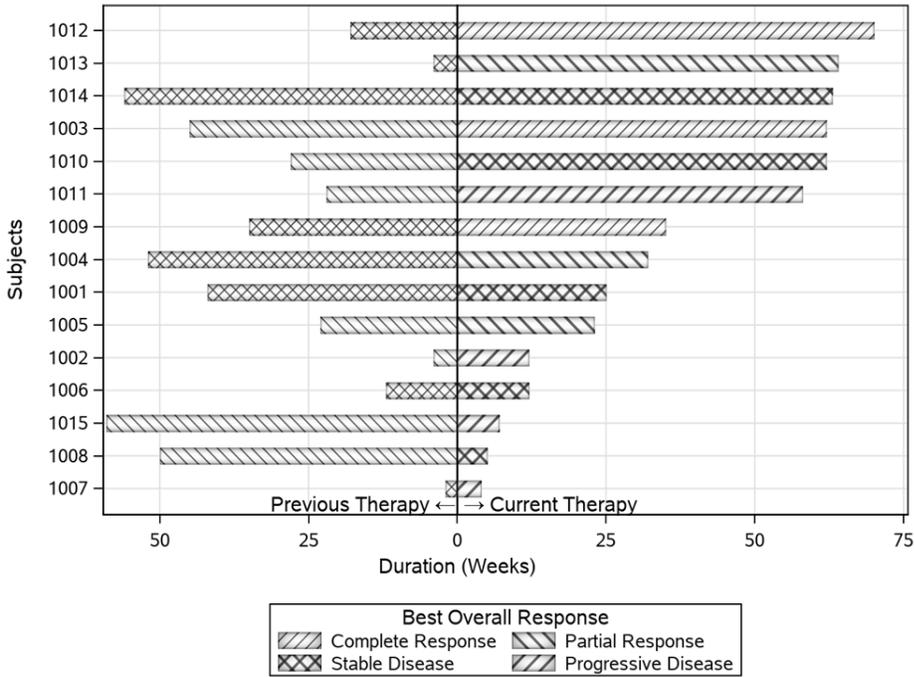


Like other bar charts, Butterfly plot can be created using patterns. For applying pattern TEMPLATE style dataset needs to be created to support PROC SGPlot. After creating the style dataset, it will be passed into RTF table Style option and FILLATTERN colors should be White and OUTLINEATTRS color should be Black.

```

proc template;
  define style styles.mylisting;
  parent = styles.listing;
  style GraphBar from GraphComponent / displayopts = "fillpattern";
  style GraphData1 from GraphData1 / fillpattern = "L1" contrastcolor=black;
  style GraphData2 from GraphData2 / fillpattern = "X1" contrastcolor=black;
  style GraphData3 from GraphData3 / fillpattern = "R1" contrastcolor=black;
  style GraphData4 from GraphData4 / fillpattern = "L2" contrastcolor=black;
  end;
run;

```

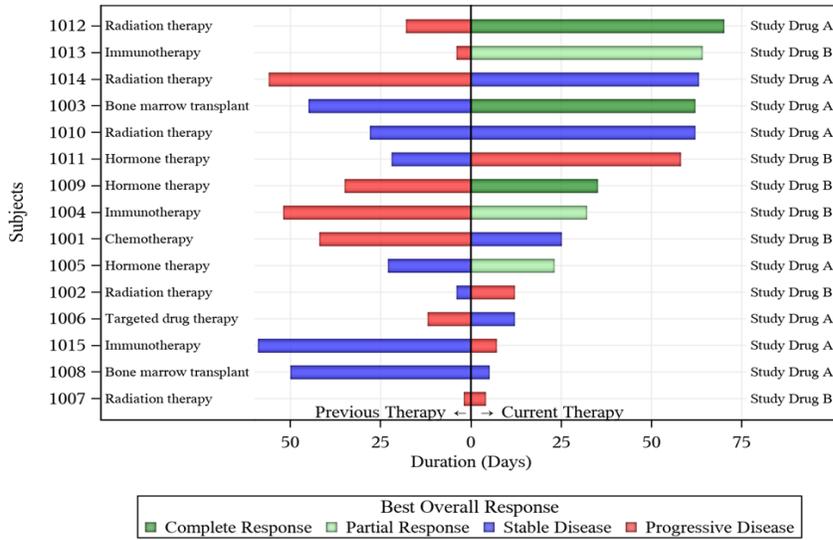


Besides displaying Subject numbers in Y-axis, Stem and Leaf plot also allow label the bars inside the plot. YAXISTABLE statement helps to display labels either or both sides of the plots. In below example, previous treatment and current treatment information used for labeling the bars respectively.

```

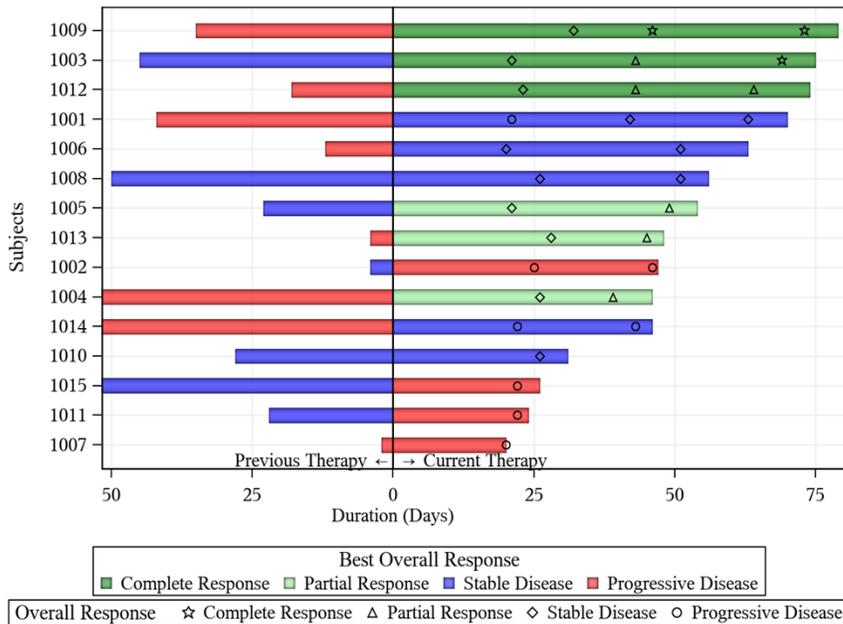
yaxistable ptrt / location=inside position=left label='';
yaxistable trt01a / location=inside position=right label='';

```



## DISPLAYING MULTIPLE OVERALL RESPONSES

Like swimmer plot, butterfly plot also allows displaying multiple overall response on both or either side of the bars. DOT statement helps to assign symbols for each response type and symbols will be placed on X-axis using corresponding duration values from ADY. Using KEYLEGND we can adjust the symbols in plot legend section.



As discussed earlier, due to the limitation of using multiple grouping variables in SGPLOT, we transposed overall response and create one column for each response. Below example given for running DOT statement on SD1 response column. User needs to run DOT statement for each response column and change the symbol and name accordingly.

```
dot id / response=SD1 markerattrs= (symbol=Diamond color=black) name="x3";
```

Using %DO loops conditional execution, we can avoid running DOT statement multiple times and reduce the length of the code and increase efficiency.

```
%let _respcnt = %nrstr(&_crcnt, &_prcnt, &_sdcnt, &_pdcnt);
%let _smb1    = %str(Star, Triangle, Diamond, Circle);
%let _resp    = %str(CR, PR, SD, PD);

%do _i = 1 %to 4;
%do _r= 1 %to %scan(&_respcnt, &_i, %str(,)) ;
dot subjid / response=%scan(&_resp, &_i, %str(,))&_r.
markerattrs=(symbol=%scan(&_smb1, &_i, %str(,)) color=black ) name="x&_i.";
%end;
%end;
```

&\_crcnt, &\_prcnt, &\_sdcnt, &\_pdcnt macro variables contains number of each response columns.

## USING PROC TEMPLATE

1. Using PROC TEMPLATE to define the template for graph
  - TEMPLOUT statement at PROC SGPLOT can write out the template
2. Using PROC SGRENDER to create the plot by calling the define template from step 1.

Here is the example of the refined template code

- Two BARCHAT for prior/post treatment ( GROUP=PAVAL for prior treatment best overall response and GROUP=BAVAL for current treatment best overall response)
- One SCATTERPLOT for overall responses with GROUP=AVAL
- DrawText statement to insert text in the plot, no need to create the annotation dataset

\*\*\* Step 1: Define graph template (initial code can be created by SGPLOT with TEMPLOUT);

```
proc template;
define statgraph Graph;
begingraph;
  layout lattice / rowdatarange=data columndatarange=data rowgutter=10
  columngutter=10;
  layout overlay / xaxisopts=(label=" " griddisplay=on offsetmin=0.1
  offsetmax=0.1 type= linear linearopts=(tickvalueformat=data
  THRESHOLDMIN=0.6 THRESHOLDMax=0.6 tickvaluepriority=true))
  yaxisopts=(label=" " reverse=true display=(tickvalues line ticks)
  griddisplay=on offsetmin=0.1 offsetmax=0.1
  discreteopts=(tickvaluefitpolicy=none));
  barchart category=SUBJID response=pady / group=paval name='a'
  orient=horizontal barwidth=0.85 groupdisplay=stack;
  barchart category=SUBJID response=bady / group=baval orient=horizontal
  barwidth=0.85 groupdisplay=stack;
  scatterplot y=subjid x=ady / group=aval name='b';
```

```

innermargin / align=left;
    axistable y=SUBJID value=TRT01A / label=' ';
endinnermargin;
DrawText "Prior Treatment " {unicode '2190'x} / X=-3 Y=4
XSPACE=datavalue YSPACE=wallpercent ANCHOR=right WIDTH=100;
DrawText " " {unicode '2192'x} " Post Treatment" / X=3 Y=4
XSPACE=datavalue YSPACE=wallpercent ANCHOR=left WIDTH=100;
DrawText "Duration (days)" / X=0 Y=-9 XSPACE=datavalue
YSPACE=wallpercent ANCHOR=center WIDTH=100;
endLayout;
endlayout;
layout globalLegend / border=false displayclipped=true
    legendtitleposition=left;
    discretelegend 'a' / location=outside title='Best Overall Response';
    discretelegend 'b' / location=outside title='Overall Responses';
endLayout;
endgraph;
end;
run;

*** Step 2: Create graph output with SGRENDER;
proc sgrender data=vadrs template=Graph dattrmap=myattrmap;
    dattrvar paval="myid" baval="myid" aval='myid';
    format pady body ady positive.;
run;

```

Example to create the dataset to control the attributes

```

data myattrmap;
    length id $4 linecolor fillcolor $ 10 value $12;
    v="CR PR SD PD";
    c="Green Lightgreen Blue Red";
    m="Star Triangle Diamond Circle";
    show='attrmap';
    id='myid';
    do j=1 to 4;
        value=scan(v, j, ' ');
        linecolor=scan(c, j, ' ');
        fillcolor=linecolor;
        markersymbol=scan(m, j, ' ');
        markersize=9;
        markercolor='viro';
        label=value;
        output;
    end;
run;

```

## CONCLUSION

Butterfly plot is an effective graphical presentation of two response variables side by side. It will help reduce number of reports and easy to compare two treatment responses. Developing butterfly plot can be as easy as swimmer plot since it uses same ADaM datasets and PROC SGPLOT or PROC TEMPLATE procedures.

## REFERENCES

HBAR Statement - Creates a horizontal bar chart that summarizes the values of a category variable:

<https://documentation.sas.com/?docsetId=grstatproc&docsetTarget=n1wvtdp1zvu57un1wjyheyd61e3i.htm&docsetVersion=9.4&locale=en>

SGPLOT Procedure - DOT Statement:

<https://documentation.sas.com/?docsetId=grstatproc&docsetTarget=p1sk20hj2wvhnkn1q7t9jq0gvj0o.htm&docsetVersion=9.4&locale=en>

## ACKNOWLEDGMENTS

The authors would like to thank the management teams for their advice on this paper/presentation.

## CONTACT INFORMATION

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

Raghava Pamulapati  
Merck & Co., Inc.  
126 Lincoln Avenue  
Rahway, NJ 07065  
Phone: 732-594-5819  
e-mail: [raghava.pamulapati@merck.com](mailto:raghava.pamulapati@merck.com)

Jane Liao  
Merck & Co., Inc.  
770 Sumneytown Pike  
West Point, PA 19486  
Phone: 267-305-7514  
e-mail: [jane\\_liu2@merck.com](mailto:jane_liu2@merck.com)

## TRADEMARK

SAS and all other SAS Institute Inc. products or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.