Waterfall Charts: Play the Mario with your data

Niraj J. Pandya, Element Technologies Inc., NJ
Sapan M. Raval, Independent Consultant, CT

ABSTRACT

Waterfall chart is a data visualization technique that depicts how a value increases or decreases for parameter of interest. Waterfall chart is gaining recognition in today's world because of its simple yet powerful representation of results. It can be used for performance or response analysis, especially for understanding or explaining the overall response of a parameter which can vary with the effect of multiple factors. Application of such charts can be applied across various domains such as Finance, Health care, Manufacturing, Clinical data analysis etc. SAS® offers multiple solutions to produce Waterfall Charts and in this article we will focus on the use of SGPLOT procedure. The example given in this paper is specific to clinical data analysis.

INTRODUCTION

Application of waterfall charts is different in various industries and is often presented/interpreted differently in different type of data analysis. It is gaining popularity because of its ability to include large data spectrum and its ability to present complete response and trend of interested parameter. In finance industry, it is often used to determine cumulative effect of different positive and negative intermediate values on an initial value of a parameter and how it leads to a final value. Initial and final values of an entity are presented by full bars on the chart and intermediate values are presented by floating bars. Because of this, such chart is also called 'Floating Bar Chart' or 'Mario Chart'. Sometimes, it is also referred as 'The Bridge'. In pharmaceutical industry, this chart is usually used in oncology clinical trials data analysis. It is used for performance analysis of drug. In such chart, each patient in the trial is presented by a vertical bar on the chart which represents the change or % change from baseline in the diameter of tumor. In controlled trials, such chart can be used to compare the performance of active drug versus placebo.

WATERFALL CHART

When a patient is enrolled in a clinical trial for the treatment of tumor, the diameter of tumor site is measured on different visits. Then collected data is compared to the data taken at baseline to determine if drug has some activity or not. Also each patient is assigned in to different categories based on overall response. Based on response and patient status, patient can be assigned in to any of the categories mentioned in following table.

CR	Complete Response
PR	Partial Response
PD	Progressive Disease
SD	Stable Disease
ED	Early Death

Plotting the maximum change(+ve or –ve) in the diameter of tumor site for each patient and color code each patient in to different groups based on above mentioned categories gives a very clear picture about the performance of drug. When the patient population in a particular trial is large, waterfall chart is the perfect presentation to give a good idea about drug activity.

To get the effect of waterfall, the flow of the plot has to be from worst value among the population on left to best value on the right side of the plot (descending order). As mentioned earlier, with some programming efforts, SAS offers multiple solutions to get required waterfall chart, but currently there is no inbuilt procedure available that can create waterfall chart readily. We will look at the use of PROC SGPLOT in this paper to generate waterfall plot. For that, the required input dataset should contain one record for each patient with value to be plotted in the vertical bar.

We will use following dataset named LSPCHG for the illustration purpose in this article.

USUBJID	LSPCHG	LSRESP
1001	10	PD
1002	12	PD
1003	-60	PR
1004	-20	SD
1005	-72	PR
1006	-42	PR
1007	-22	PR
1008	-66	CR
1009	-12	PR
1010	-14	SD
1011	-84	CR
1012	-48	PR
1013	-36	PR
1014	44	SD
1015	-27	PR
1016	-29	PR
1017	-61	CR
1018	53	PD
1019	25	ED
1020	-24	PR
1021	-37	PR
1022	-43	PR
1023	50	PD
1024	-81	PR
1025	-44	CR
1026	-33	PR
1027	-83	PR

Dataset Continue →

USUBJID	LSPCHG	LSRESP
1028	-34	PR
1029	-94	PR
1030	-14	SD
1031	-16	SD
1032	-86	PR
1033	-89	PR
1034	-13	SD
1035	-24	PR
1036	13	SD
1037	-55	PR
1038	-86	CR
1039	-27	SD
1040	-20	SD
1041	-13	PR
1042	-27	PR
1043	16	PD
1044	-26	PR
1045	-76	CR
1046	-82	ED
1047	-83	PR
1048	-41	PR
1049	-47	PR
1050	-50	CR
1051	6	PR
1052	0	SD
1053	-2	PR
1054	-9	PR

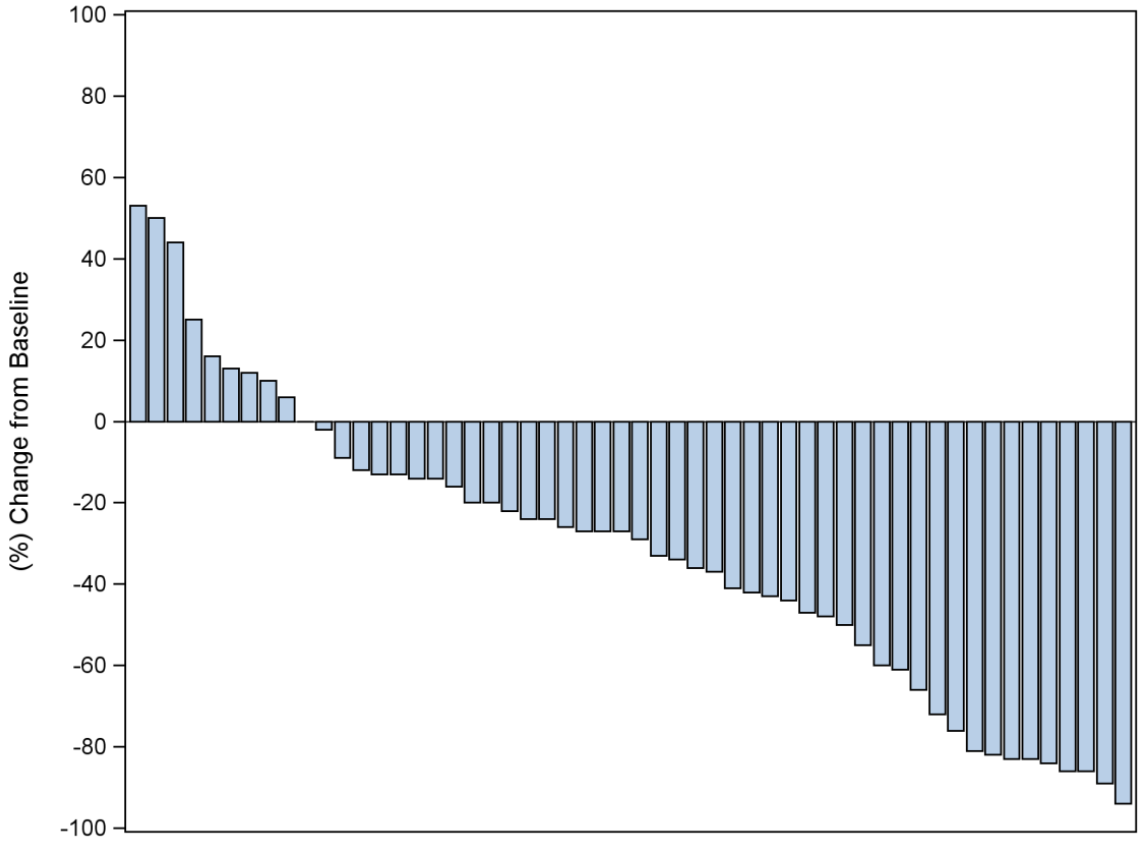
In this dataset, LSPCHG variable contains the value of % change from baseline in the tumor size for each patient. Variable LSRESP specifies the category of overall response for each patient.

Following piece of code explains how easily we can create a waterfall chart from above dataset using PROC SGPLOT.

```
/* Sort the input dataset in descending order for variable lspchg */

Proc Sort Data = lspchg;
     By descending lspchg;
Run;


Data lspchg;
     Set lspchg;

     n = _n_; /* Create variable n that contains observation number */

Run;

/* Create waterfall chart */
```

```
Proc Sgplot Data=lspchg;
  Vbar n / response=lspchg;

  Xaxis display=none;
  Yaxis values=(-100 to 100 by 20) label="(%) Change from Baseline";

Run;
```

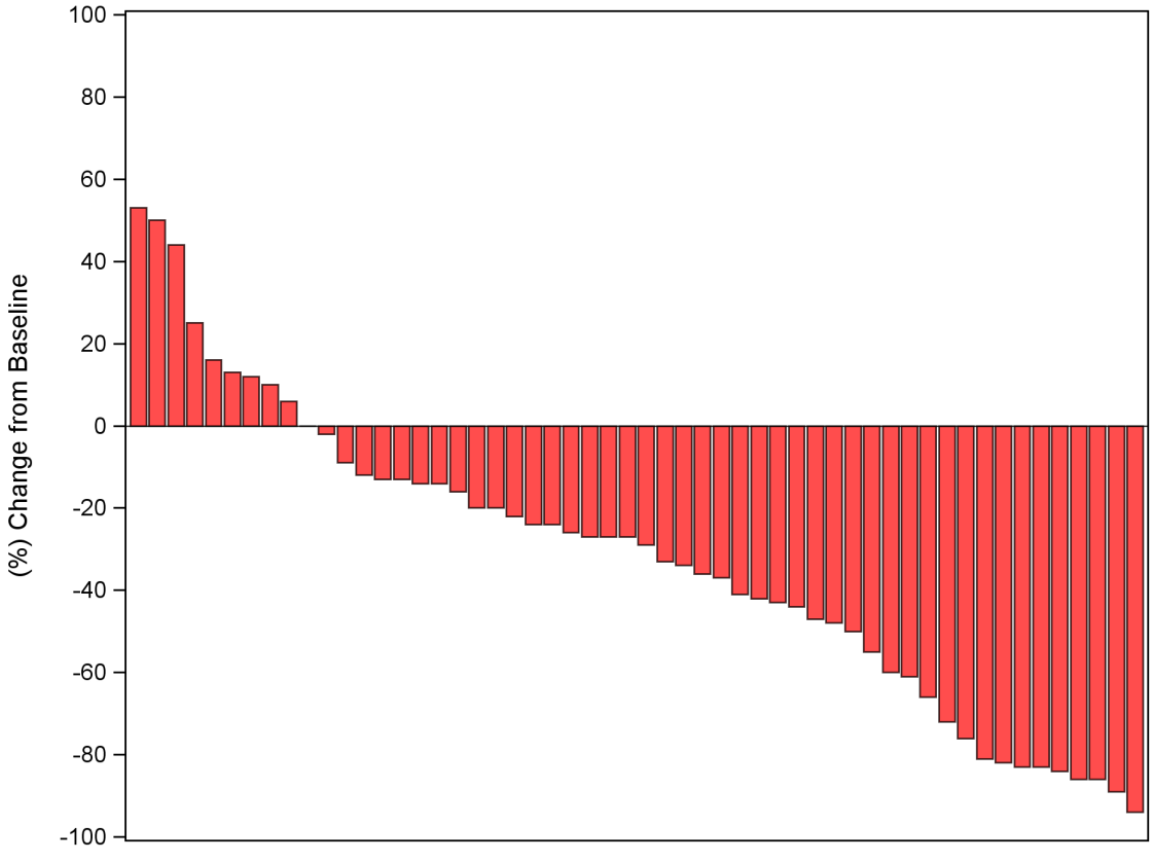
The output from above code looks like below figure. In above code, if you notice, no color is specified for the fill in the vertical bars. In this case, since we did not group the data, SGPLOT procedure picked the default color from COLOR attribute of the GraphDataDefault style element of current style.



Now let's say we do not want the default color for the bars and we want to change the color for the bars and also the transparency of the fill. Following piece of code explains how this can be done with some minor addition in the code and additional code is in bold font.

```
/* Create waterfall chart */

Proc Sgplot Data=lspchg;
  Vbar n / response=lspchg FILLATTRS = (COLOR=red) TRANSPARENCY = 0.3;

  Xaxis display=none;
  Yaxis values=(-100 to 100 by 20) label="(%) Change from Baseline";


Run;
```

In above code, we are using FILLATTRS option to change the color of the fill for bars. Since the data is ungrouped, we can easily change the color of the fill by using COLOR sub-option and specifying required value for this sub-option. We are also using TRANSPARENCY option to change the transparency of the fill color of the vertical bars. This option specify the degree of the transparency and valid value ranges from 0.0 to 1.0 where 0.0 being completely opaque and 1.0 being completely transparent. The default value for this option is 0.0. In this example, we are using the transparency of 0.3 with fill color of RED. The output from above code will look like below figure.



Now let's say we want to enhance this plot further, we want to customize it in a way so that it makes more sense. We want to create this plot by grouping different patients based on their overall response category and fill the bars of such patients with different colors so it is easy to identify different groups. At the same time, we want to increase the space between the bars. We also want to highlight some special cases.

Let's say we want to highlight patients who are in the category of 'Early Death'. We want to highlight the patients in the category of 'Complete Response' and who got decrease in the tumor size equal to or more than 75 %. We also want to highlight cases where tumor size increased by equal or more than 50 %.

Following piece of code demonstrates how we can customize the plot based on our requirement.

```
/* Define the variables with values to be plotted on Y axis for special cases */

Data lspchg;
     Set lspchg;

     If lsresp = 'CR' and lspchg <= -75 then
```

```
        CR_Y = lspchg + ((lspchg/abs(lspchg))*2);

     If lsresp = 'ED' then ED_Y = lspchg + ((lspchg/abs(lspchg))*2);

     If lspchg >= 50 then GE50_Y = lspchg + ((lspchg/abs(lspchg))*2);
Run;

/* Create waterfall chart */

Proc Sgplot Data=lspchg;
  Vbar n / response=lspchg GROUP=lsresp BARWIDTH = 0.6;

  Vline n / response=CR_Y MARKERS
            MARKERATTRS=(SYMBOL=plus SIZE=4 COLOR=black)
            LINEATTRS=(THICKNESS=0);

  Vline n / response=ED_Y MARKERS
            MARKERATTRS=(SYMBOL=star SIZE=4 COLOR=blue)
            LINEATTRS=(THICKNESS=0);

  Vline n / response=GE50_Y MARKERS
            MARKERATTRS=(SYMBOL=triangle SIZE=4 COLOR=red)
            LINEATTRS=(THICKNESS=0);

  Xaxis display=none;
  Yaxis values=(-100 to 100 by 20) label="(%) Change from Baseline";

Run;
```

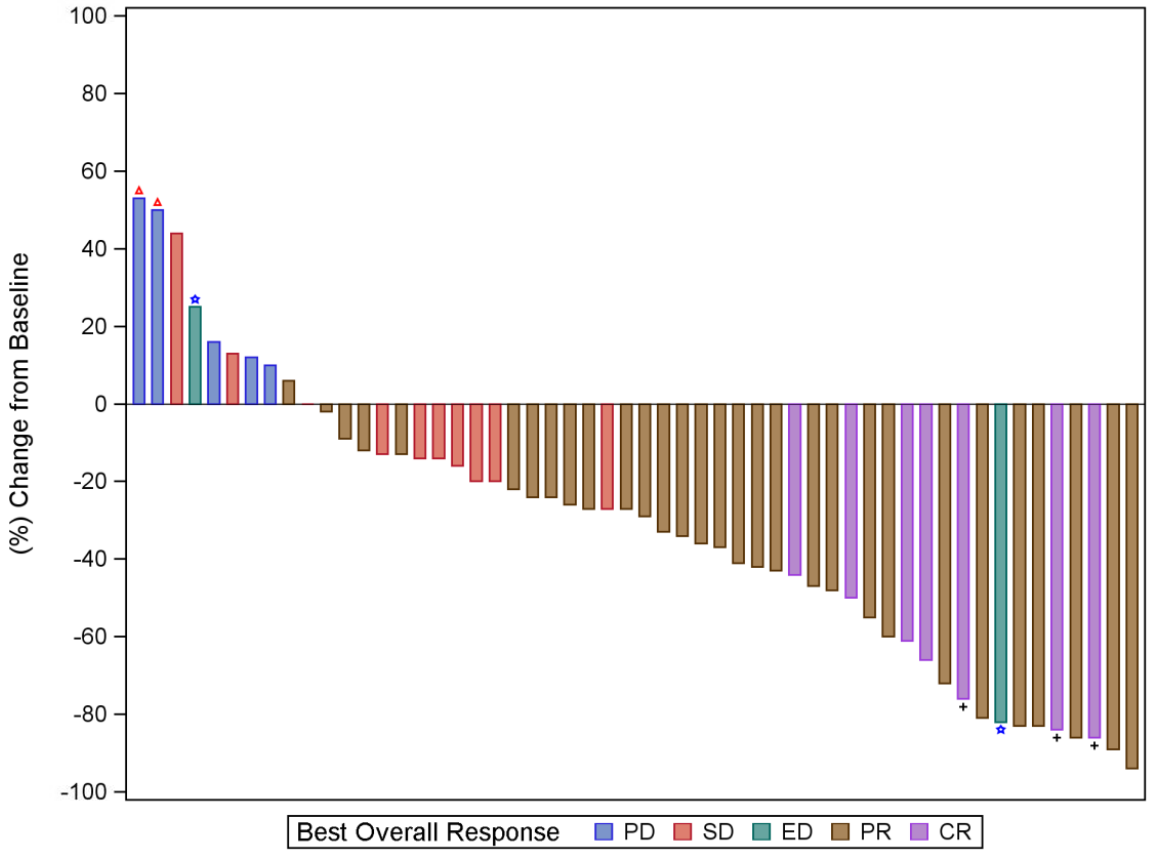
In above code, different variables (CR_Y, ED_Y, and GE50_Y) with the values to be plotted on Y scale are created first in the input dataset for special cases that we want to highlight. Later, in PROC SGPLOT, these variables are used as response variables in VLINE statements to create different symbols to highlight special cases.

If you notice, we are using GROUP option now in VBAR statement to group different patients by their overall response. Since we are grouping the data, different colors are used by SGPLOT procedure from COLOR attribute of GraphData1, GraphData2…….GraphData**n** style elements of the current style. To over write the default colors in this case, previously mentioned FILLATTRS option and COLOR sub-option will not work. To do so, we can use PROC TEMPLATE to define a new style by changing the COLOR attribute of GraphData style elements from default style and use this newly created style.

Also note that we are using BARWIDTH option in VBAR statement to increase the space between two bars. This option specifies the width of the bars as a ratio of the maximum possible width. The maximum possible width here is the distance between the center of each bar and the centers of the adjacent bars. The valid values for this option are from 0.1 to 1. Value of 1 leaves no space between the bars. The default value of this option is 0.8 and in this example, we are using the value of 0.6.

To highlight special cases on the chart, as mentioned earlier, we are using previously created variables as response variables in different VLINE statements. For every type of special case, we have added one VLINE statement in above code. In VLINE statements, we are making use of MARKERS and MARKERATTRS options to specify different symbols with different colors to highlight special cases. At the same time we are using LINEATTRS options and THICKNESS sub-option with the value of 0 to suppress the line created by VLINE statement. This will put different symbols at the end of the vertical bar for patients that are special cases.

The output from above code will look like below figure. In this figure, bars with blue star at the end of the bar are the cases with early death. Bars with black '+' sign are the cases from 'Complete Response' group with decrease in tumor size equal to or more than 75 %. Bars with red triangle are the cases that have increase in tumor size equal to or more than 50%.



In above figure, the label of legend is coming from the label of variable used as grouping variable which is LSRESP in our case.

Some other noteworthy options in VBAR statement that can come handy to change the appearance of the bars are FILL, NOFILL, OUTLINE and NOOUTLINE. FILL and OUTLINE are the defaults and can be changed using NOFILL and NOOUTLINE respectively. FILL/NOFILL controls whether the bars have fill and OUTLINE/NOOUTLINE controls whether the bars have outline. Please note that NOFILL and NOOUTLINE can not be used in the same VBAR statement.

One another useful option that has been added recently in SGPLOT procedure is DATALABEL. This option in VBAR statement adds data labels for bars. The values of the response variable appear at the end of the bars which adds readability in the waterfall chart and is very useful particularly when large data is being presented in the chart. This option is available in SAS release 9.2 Phase 2 and later.

CONCLUSION

Waterfall chart is a very useful tool for various types of quantitative analysis. Area of application is vast for such charts and can be used in various analysis done in accounting, inventory analysis, performance analysis, test analysis, sales analysis etc. Such charts can be created using various graphical procedures in SAS (PROC GCHART, PROC GPLOT) but customization of such charts may require lots of annotation based on the requirements. Recently introduced PROC SGPLOT provides lots of options and features that can be used to create such charts with very little to no annotation efforts.

REFERENCES

SAS/GRAPH V9.2 Online Documentation -
http://support.sas.com/documentation/cdl/en/grstatproc/62603/HTML/default/viewer.htm#sgplot-chap.htm

CONTACT INFORMATION

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

Name: NIRAJ J PANDYA
Phone: 201-936-5826
E-mail: npandya@elementtechnologies.com

Name: SAPAN RAVAL
Phone: 848-565-4097
E-mail: sapan.gateway@gmail.com