PhUSE US Connect 2020

Paper DV14

Volcano Plot: Demystifying Why and How

Jyothi Nelakurthi, Merck & Co., Inc, Rahway, NJ, USA

Chandana Sudini, Merck & Co., Inc, Rahway, NJ, USA

ABSTRACT

Volcano plot is yet another two-dimensional space saving scatter plot to graphically present measure of strength versus significance of a statistical signal. It may be used to interpret multiplicity adjustments to reduce false positives. When events (AE) are displayed by time interval, it illustrates how risk changes over the course of a trial. It is widely used in genomics to analyze microarray expression data etc. It is a powerful tool to visually display the statistical significance as p-values on Y-axis against metrics like relative risk difference, odds ratio, or hazard ratio on X-axis for a given treatment comparison. This paper talks about the interpretation of volcano plot in safety evaluation and construction of graphic elements using SGPLOT procedure statements by providing an example from a late stage clinical trial.
Key Words: Volcano Plot, SGPLOT, Safety Evaluation

APPLICATION

Volcano plots were originally used in genomics to present large fold changes in genes, adjusted for multiplicity, from microarray experiments. Currently these are extended to other areas such as Adverse events, Interventions and Findings Analyses in clinical trials.

They are used in genomics to visually identify the meaningful changes between two conditions where there are many thousands of genes as replicate data points and displays large magnitude changes that are statistically significant.

In research areas with multiple endpoints, multiple treatment arm comparisons and analysis of multiple outcomes, volcano plots can be used to display the appropriate estimands and adjusted p-values to minimize the impact of inflated rate of false positive rate on conclusions due to testing of multiple hypothesis of interest.

Volcano plots have found their place in exploratory analysis to emphasize important differences between the adverse event profiles of two treatment arms. Presenting adverse events (AE) data can be challenging as a large amount of information has to be summarized succinctly. In general, the incidence of Adverse experience in safety reports are presented by Body System or Organ Class (AEBODSYS) along with preferred or decoded terms (AEDECOD) within a group. There are many graphs like rainfall or dot plots to display this information, but volcano plot can be beneficial as it helps in interpreting large amount of information in a spatially constrained manner.

It can be applied to other areas in safety analysis like summarizing frequency of laboratory abnormalities across study defined time intervals. In trials where subjects may be on many concomitant medications that are likely to impact treatment outcome, volcano plots allow a visual comparison of the difference in usage of each concomitant medication between the two treatment arms. This can help explain the possible impact on the final treatment outcome between the treatment arms.

CONSTRUCTION OF VOLCANO PLOT WITH CODE AND EXAMPLES

Creating a volcano plot is quite simple using PROC SGPLOT and combination of SCATTER, REFLINE and few other statements. In the below sections there are details about the input dataset followed by code for construction of the plot with explanation of the statements used in each step.

INPUT DATA:

Authors have included AE preferred terms of selective system organ classes from two treatment groups to generate all the graphs in this paper.

The key input variables required to generate the plot are listed on next page.

Body System or Organ Class	Preferred Term	Risk Difference	P-Value
Blood and lymphatic system disorders	Anemia	13.1	0.002
Infections and infestations	Cellulitis	-2.6	0.044
Nervous system disorders	Dizziness	-2.8	0.322
Gastrointestinal disorders	Abdominal pain	-4.5	0.138
Renal and urinary disorders	Acute kidney injury	6.3	0.015
Respiratory, thoracic and mediastinal disorders	Cough	0.7	0.829
Musculoskeletal and connective tissue disorders……	Flank pain…	-1.7…	0.18….

Display 1: Input data for the plot

Risk difference is calculated as the difference between the incidence proportion of an event in Drug A and the incidence proportion of the same outcome in the Drug B. P-values are calculated with respect to derived risk difference.

BASIC PLOT

The first and most basic plot displays the risk difference of AE's against the log transformed p-values, where each filled circle represents an AE preferred term.

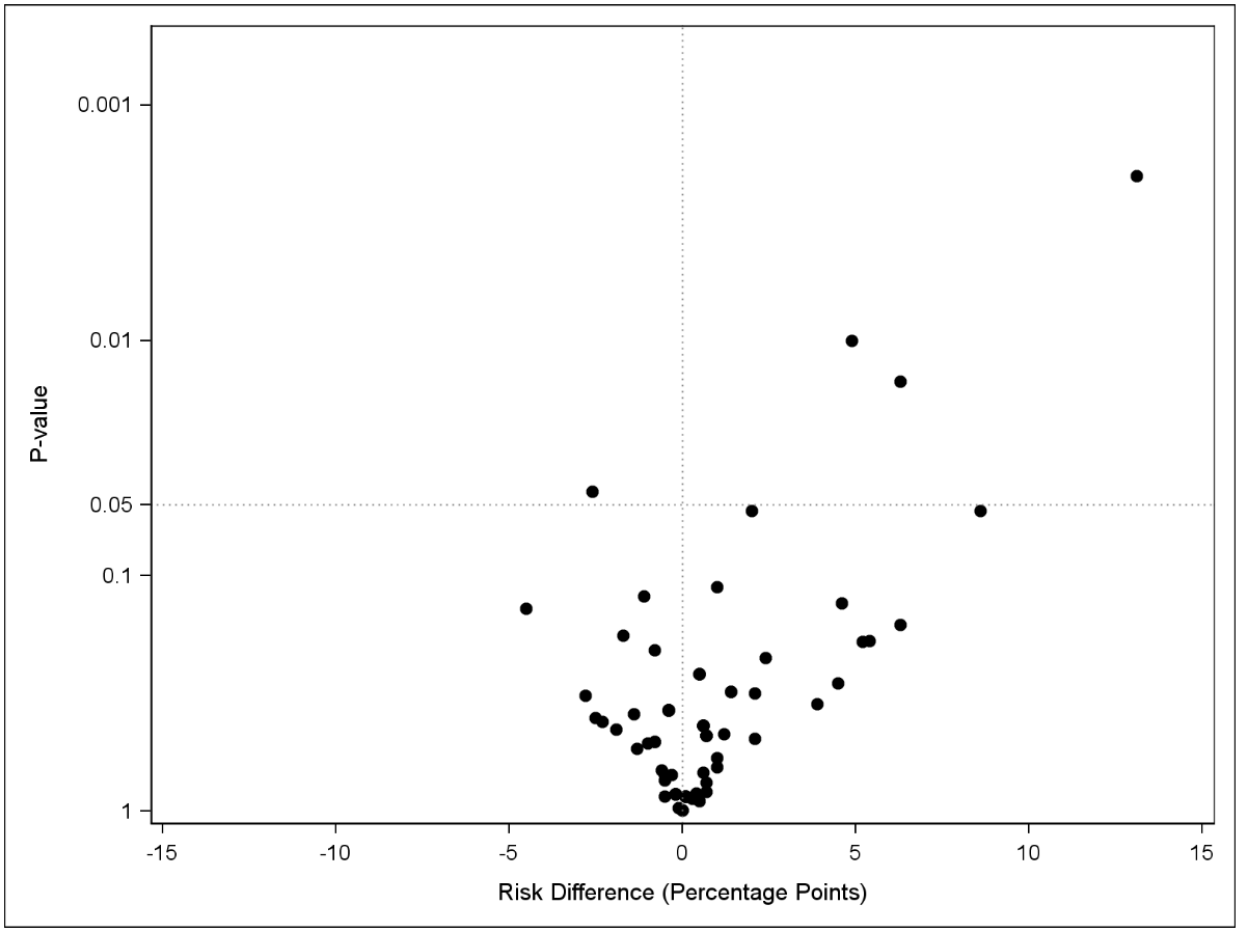

Figure 1: Basic Volcano Plot

This plot can simply be generated by using following PROC SGPLOT statements: SCATTER, REFLINE, LABEL, YAXIS and XAXIS. Code and a short description of each statement and corresponding arguments are below.

CODE:

```
proc sgplot data=vol1v3;
       label Risk Difference='Risk Difference (Percentage Points)';
       label P-value='P-value';
       scatter x=Risk y=P-value /markerattrs= (symbol=circlefilled color=black);
       refline 0/axis=x lineattrs=(pattern=dot);
       refline 0.05/axis=y lineattrs=(pattern=dot);
       yaxis reverse type=log values= (1.0 0.1 0.05 0.01 0.001) offsetmin=0.1;
       xaxis values= (-15 -10 -5 0 5 10 15);
run;
```

Statement: SCATTER

The SCATTER statement generates a scatter plot. They are used to plot data where there is no direct relationship between the data points, but in some cases where there may be a correlation. These plots are useful for displaying how data is distributed. Additional arguments on the SCATTER in the afore-mentioned graph are:

- MARKERATTRS: This option controls the appearance of marker symbols. The marker attributes most commonly specified are COLOR, SIZE, SYMBOL.

Statement: REFLINE

The REFLINE statement draws reference lines, horizontal or vertical, at a specified set of values. Additional arguments on the REFLINE are:

- AXIS: specify which axis the reference lines relate to (X|Y).
- LINEATTRS: specifies attributes of the reference lines.
    - PATTERN: specifies the line style – continuous, dotted, dashed, etc.,

Statement: LABEL

The LABEL statement specifies a label for a variable. Typically PROC SGPLOT uses this option to label the appropriate axis when the variable is plotted.

Statement: YAXIS

The YAXIS specifies the axis options to control the features of the axis like TYPE, REVERSE, OFFSETMIN/ OFFSETMAX, VALUES.

- TYPE: specifies type of the axis like LOG|TIME|LINEAR. In our case we are using LOG, which specifies the logarithmic scale for the axis.
- REVERSE: specifies the tick values to be displayed in reverse order.
- OFFSETMIN/ OFFSETMAX: specifies to leave 10% of the unused axis at either end, so that the values do not extend beyond the ends of the axis.
- VALUES: creates ticks for specified values.

Statement: XAXIS

Like the YAXIS statement the plot code also requires an XAXIS statement, with similar arguments as the YAXIS described above.

CUSTOMIZED PLOTS

The basic plot is a good place to start but we have further provided options to generate enhanced graphs to display all the information of interest, to specifically identify the AE terms with significant risk difference.

PLOT 1:

Below graph, we have displayed all AE's pertaining to a Body System or Organ Class with a single color and labeled them. The dashed line on y-axis shows $p = 0.05$ with points above the line having $p < 0.05$ and points below the line having $p > 0.05$.

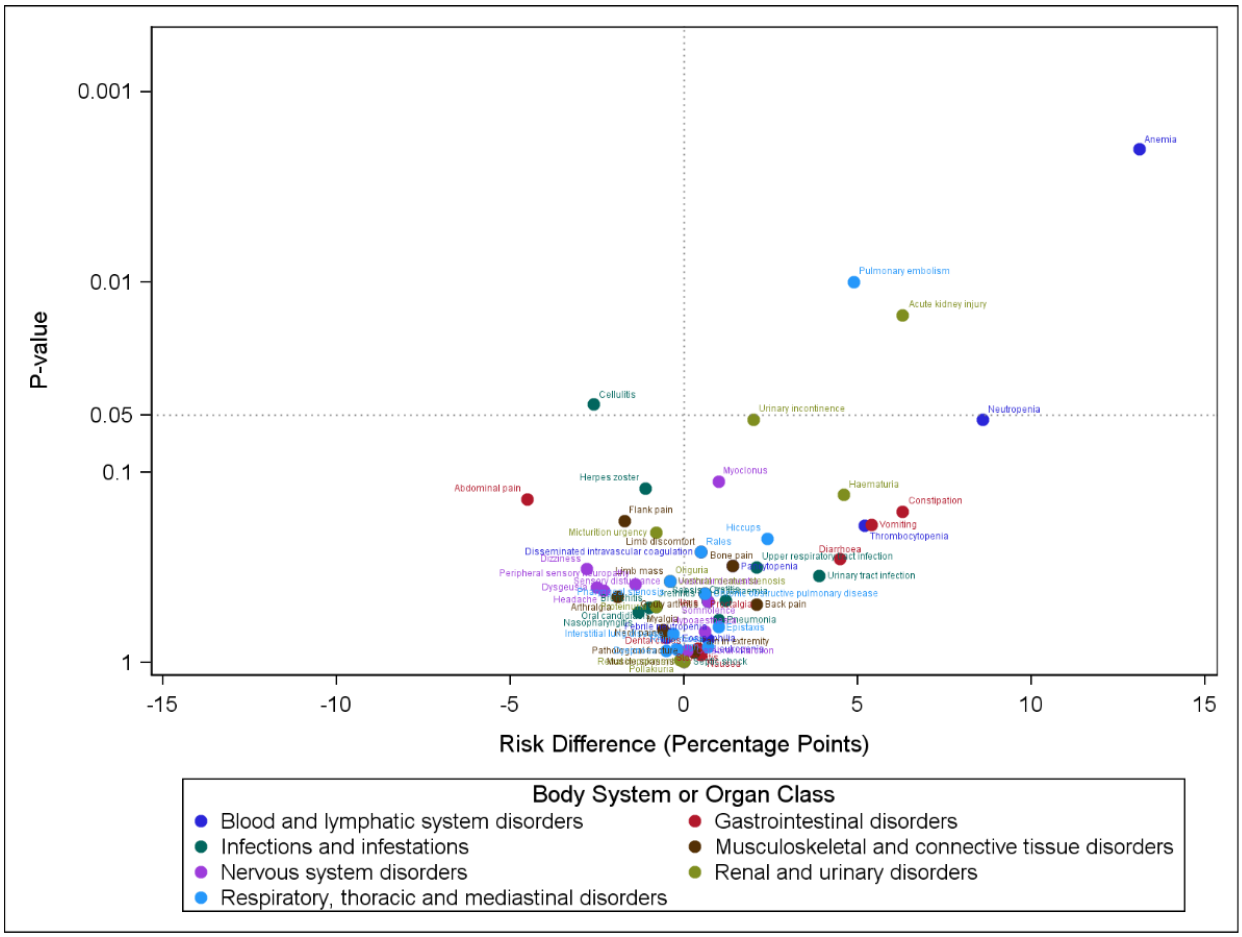

Figure 2: Customized Plot 1

Each circle represents an AE preferred term. All the circles/PTs pertaining to a Body System or Organ Class are assigned same color. Code and a short description of additional statements and corresponding arguments used for this graph are below.
Note: Additional options used in below code are displayed in bounded boxes

CODE:

```
proc sgplot data=vol1v32;
       label risk ='Risk Difference (Percentage Points)';
       label pval='P-value';
       scatter x=risk y=pval / GROUP=aebodsys DATALABEL=aedecod
       markerattrs= (symbol=circlefilled);
       refline 0/axis=x lineattrs=(pattern=dot);
       refline 0.05/axis=y lineattrs=(pattern=dot);
       yaxis reverse type=log values= (1.0 0.1 0.05 0.01 0.001) offsetmin=0.1;
       xaxis values= (-15 -10 -5 0 5 10 15);
run;
```

SCATTER:

- GROUP**:** specifies a grouping variable. Observations with the same value of the grouping variable will be plotted with the same symbol/color.
- DATALABEL**:** specifies a label variable, whose value is to be plotted adjacent to each point.

In this plot we have grouped body system or organ class and displayed preferred terms as data labels. Labeling all the AE preferred terms made the above graph look inept, so in the next section of the paper, we have provided options to display labels for AEs of significance only.

PLOT 2:

This graph displays labels for AE terms with p-values below 0.05 only and shows the favorability of an AE towards a drug. Data points with low p values (highly significant) will appear toward the top of the plot, those AEs are our point of interest for safety evaluation.

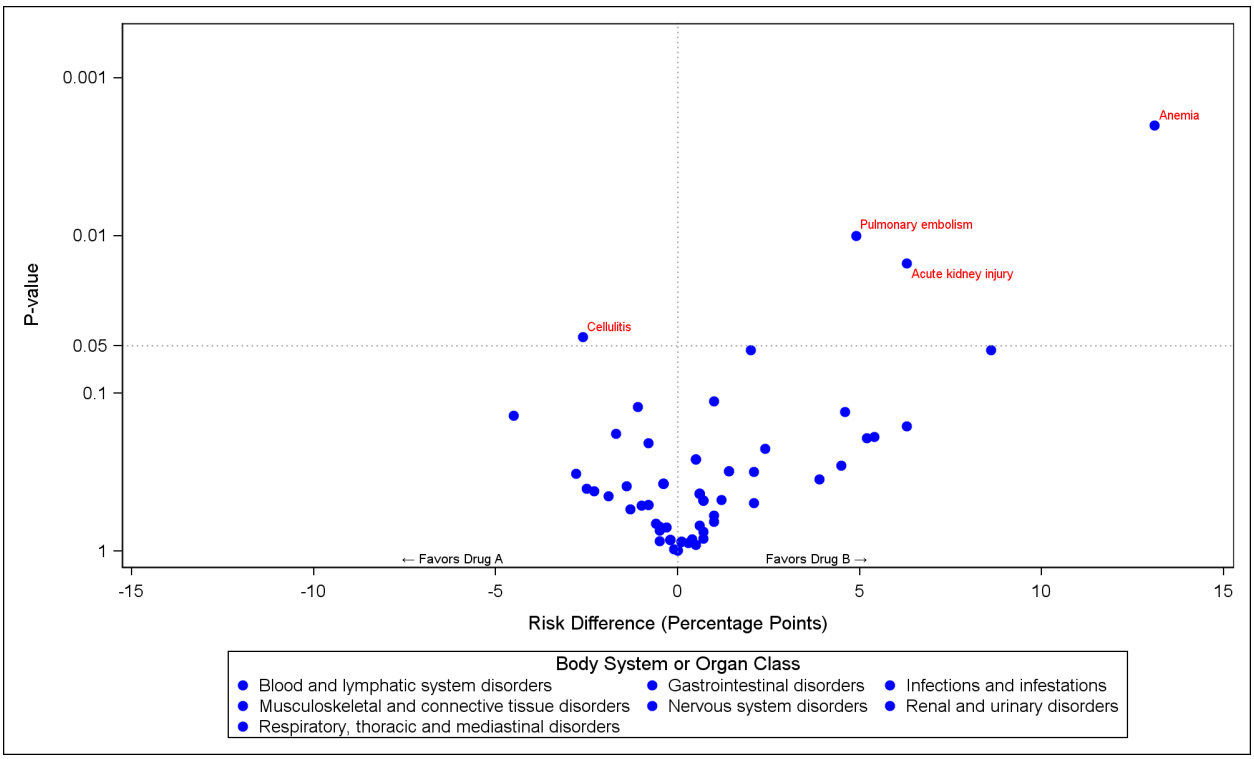

Figure 3: Customized Plot 2

From the above graph it is evident that Anemia (1), Pulmonary Embolism (2), Acute Kidney Injury (3) and Cellulitis (4) are AEs with significant risk difference as they are plotted above the 0.05 p-value line. Of which, 1,2,3 are favoring Drug B and 4 is favoring Drug A.
Code and a short description of additional statements and corresponding arguments used for this graph are below.
Note: Additional options used in below code are displayed in bounded boxes.

CODE:

```
proc format;
  value $fav
  "A" = "(*ESC*) {Unicode '2190'x} Favors Drug A"
  "B" = "Favors Drug B (*ESC*) {Unicode '2192'x}";
run;

data vol1v32;
  set vol1v3 end=last;
  if pval < 0.05 then label=aedecod;
 output;
  if last then do;
  ylbl=1.0; xlbl=-5; text="A"; output;
    ylbl=1.0; xlbl=5; text="B"; output;
  end;
```

```
run;

        ods graphics / width=10in height=6in maxlegendarea=100;
        proc sgplot data=vol1v32;
                format text $fav.;
                label risk='Risk Difference (Percentage Point)';
                label pval='P-value';
                scatter x=risk y=pval /group=aebodsys datalabel=label
                markerattrs= (symbol=circlefilled color=blue) datalabelattrs=(color=red);
                refline 0/axis=x lineattrs=(pattern=dot);
                refline 0.05/axis=y lineattrs=(pattern=dot);
                text x=xlbl y=ylbl text=text / position=bottom contributeoffsets=(ymax);
                yaxis reverse type=log values= (1.0 0.1 0.05 0.01 0.001) offsetmin=0.1;
                xaxis values= (-15 -10 -5 0 5 10 15);
        run;
```

Statement 1:TEXT

It displays the associated text values at (X, Y) locations in the graph. This statement can be used to rotate text at any angle, manage the text position, split the text into multiple lines. It can also display a bounding box around the specified text and add a back-light effect to the text.

- DATALABELATTRS: specifies the appearance of the labels in the plot when using DATALABEL= option.
- POSITION: specifies the position of the legend within the graph like BOTTOM | BOTTOMLEFT | LEFT | RIGHT | TOP and so on and so forth.
- CONTRIBUTEOFFSETS**:** specifies whether the plot's space requirements contribute to the calculation of the axis offsets. This value determines which axis offsets can be affected by the plot.

PLOT 3:

Below is a bubble plot where each bubble represents an adverse event and bubble size indicates the total number of adverse events that occur in both treatments combined. Radius of any bubble is proportional to the square root of the event frequency.

INPUT DATA

Body System or Organ Class	Preferred Term	Risk Difference	P-Value	Bubble Size
Blood and lymphatic system disorders	Anemia	13.1	0.002	70
Infections and infestations	Cellulitis	-2.6	0.044	2
Nervous system disorders	Dizziness	-2.8	0.322	11
Gastrointestinal disorders	Abdominal pain	-4.5	0.138	13
Renal and urinary disorders	Acute kidney injury	6.3	0.015	9
Respiratory, thoracic and mediastinal disorders	Cough	0.7	0.829	15
Musculoskeletal and connective tissue disorders……	Flank pain…	-1.7…	0.18….	2….

Display 2: Input data for below bubble plot

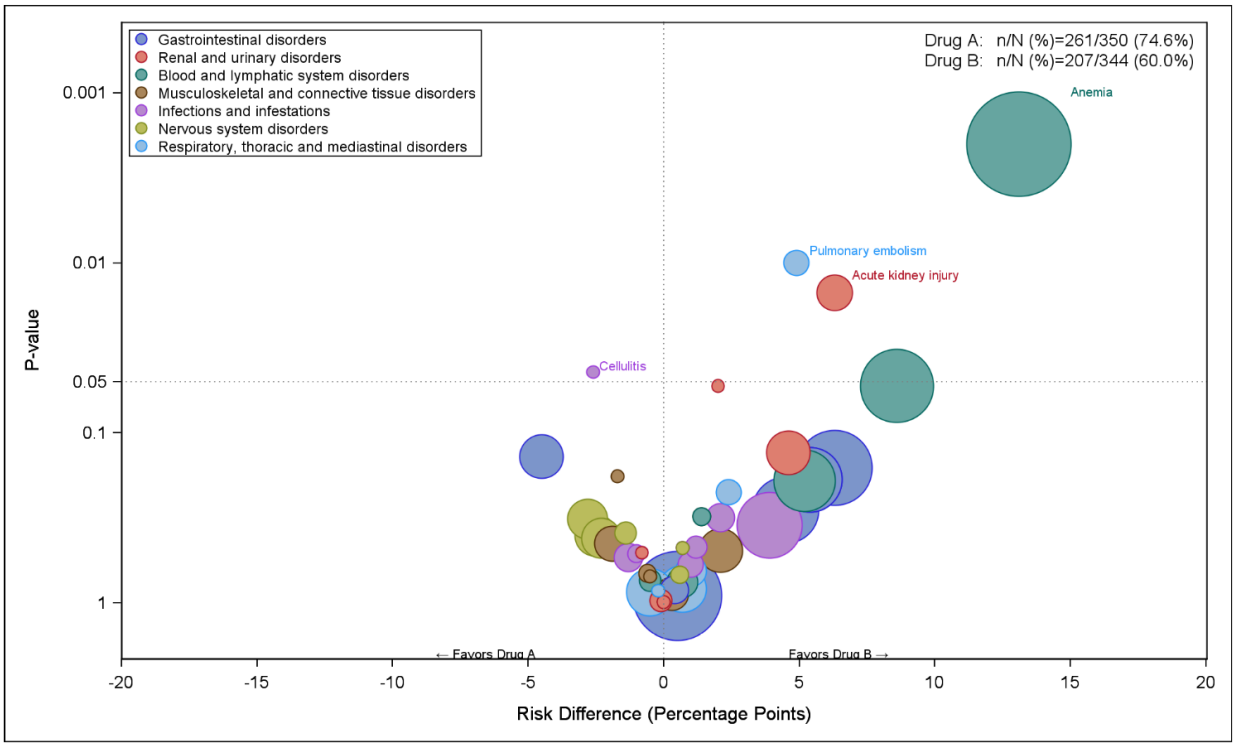

Figure 4: Bubble plot

As shown above, Anemia is the most commonly occurring significant adverse event during this trial and is favoring Drug B. Bubbles to the right indicate a higher risk for subjects in Drug A and bubbles to the left indicate a higher risk for subjects in Drug B. Names of events with p-value < 0.05 are displayed. Code and a short description of additional statements and corresponding arguments used for this graph are below.
Note: Additional options used in below code are displayed in bounded boxes.

CODE:

```
proc format;
  value $fav
  "A" = "(*ESC*) {Unicode '2190'x} Favors Drug A"
  "B" = "Favors Drug B (*ESC*) {Unicode '2192'x}";
run;

%let N1= 261/350 (74.6%);
%let N2= 207/344 (60.0%);

data bubble1;
  set bubble1 end=last;
  if pval < 0.05 then label=aedecod;
  output;
  if last then do;
        ylbl=1.8; xlbl=-5; text="A"; output;
        ylbl=1.8; xlbl=8; text="B"; output;
  end;
run;

proc sgplot data=bubble1;
       format text $fav.;
       label risk='Risk Difference (Percentage Points)';
       label npval='P-value';
```

```
        bubble x=risk_disp y=npval size=size / group=aebodsys datalabel=label name='a'
        bradiusmin=0 bradiusmax=39;
        refline 0/axis=x lineattrs=(pattern=dot);
        refline 0.05/axis=y lineattrs=(pattern=dot);
        inset ("Drug A:" ="n/N (%) =&N1"
                "Drug B:" ="n/N (%) =&N2") / noborder position=topright;
        text x=xlbl y=ylbl text=text / position=bottom contributeoffsets=(ymax);
        yaxis reverse type=log values= (1.0 0.1 0.05 0.01 0.001);
        xaxis values= (-20 -15 -10 -5 0 5 10 15 20) offsetmin=0 offsetmax=0;
        keylegend 'a' / across=1 location=inside valueattrs=(size=8);
run;
```

Statement: BUBBLE

This statement creates a bubble on the plot, where (x, y) coordinate variables determine the location of the bubble center.

- SIZE**:** specifies to control the size of the bubble.
- BRADIUSMIN**:** specifies the size of the radius of the smallest bubble.
- BRADIUSMAX**:** specifies the size of the radius of the largest bubble.

Statement: INSET

This statement adds a text box inside the plot.

- NOBORDER/BORDER: Adds or removes the border around the legend.

Statement: KEYLEGEND

This statement adds a legend to the plot. By using this statement, the default legend can be modified according to the specifications.

- NAME: Assigns a name to the plot statement.
- ACROSS: specifies the number of the columns in the legend.
- LOCATION: specifies location of the legend inside or outside the plot.
- VALUEATTRS: specifies how to align the values when label-value pairs are provided.

Using KEYLEGEND statements and its options we have placed the legend with Body System or Organ Class information at the top left corner within the graph.

SUMMARY

In the drug development process patient safety is always a primary focus. Safety information is usually reported in tables which are voluminous and difficult to interpret; however graphical outputs can be beneficial in presenting concise summaries to communicate important implications. Volcano plot is one of the tools used to easily identify differences in large datasets and display number of data points in the spatially constrained manner. Though the volcano plot has the limitation that it allows only pairwise comparisons, it still emphasizes on important difference between the arms. PROC SGPLOT statements and options can be used to generate multiple variations of plots described above. Although, this paper discusses interpretation and implementation of volcano plot in the context of summarizing incidence of adverse events, it can be applied to other areas in safety analysis as well.

REFERENCES

1. Richard C Zink, Russell D Wolfinger and Geoffrey Mann. Summarizing the incidence of adverse events using volcano plots and time intervals. Clinical Trials 2013; 10:398-406
2. Mehrotra DV, Adewale AJ. Flagging clinical adverse experiences: Reducing false discoveries without materially compromising power for detecting true signals. Stat Med 2012; 31: 1918–30.
3. Sanjay Matange. https://blogs.sas.com/content/graphicallyspeaking/2016/05/23/ctspedia-clinical-graphs-volcano-plot/#prettyPhoto

ACKNOWLEDGEMENTS

The authors would like to thank statistician Uma Kher for her valuable input and feedback.

CONTACT INFORMATION

Your Comments and Questions are valued and encouraged. Contact the authors at:

Jyothi Nelakurthi

Merck &Co., Inc.

Rahway, NJ-07065

Email: Jyothi.nelakurthi@merck.com

Chandana Sudini

Merck &Co., Inc.

Rahway, NJ-07065

Email: chandana.sudini@merck.com