Creating Forest Plots Using SAS/GRAPH and the Annotate Facility

Amanda Tweed, Millennium: The Takeda Oncology Company, Cambridge, MA

ABSTRACT

Forest plots have become common in the pharmaceutical industry as a way of comparing the relative strength of treatment effect by providing a visual representation of the amount of variation between different groups. At present, SAS/GRAPH® does not have a stand alone procedure that can be called to generate these plots. Creation in SAS requires additional processing of the data and use of the annotate facility.

This paper demonstrates the development and application of a simple SAS program for generating forest plots as part of a Phase III clinical trial. It describes the annotate facility and how it was applied to achieve the desired result. It also highlights some of the interesting features that were added in response to requests from study clinicians and statisticians.

INTRODUCTION

Interest in forest plots has increased in recent years as clinicians and reviewers have begun to recognize their value when assessing trends across multiple groups. A forest plot displays the results, by group, as a horizontal line, representing the 95% confidence interval, and a single dot, representing the point estimate of the outcome variable. The horizontal line provides a measure of the precision of the estimate, with line length being directly proportional to the variability in the data. The visual representation allows for easy review and comparison across many factors.

Traditionally, forest plots have been used in meta-analysis to demonstrate variability across multiple studies. More recently, they have been used within larger clinical studies to examine differences in treatment effect between and within sub-groups and across risk factors. The utility of forest plots is highly dependent on the size of the sample, as smaller samples will have greater variability (hence, wider confidence intervals) and less meaningful point estimates. As such, they may not be relevant in a clinical program until pivotal studies are reached. Phase III studies and integrated summaries of efficacy present additional opportunities for displaying the data in this manner.

At present, no procedures are available in SAS/GRAPH to automatically generate a forest plot. Creation requires data manipulation, application of the annotate facility to the relevant output and use of the GPLOT procedure. While other software packages may be able to generate the output more simply, in the regulated environment of industry, all of the clinical programming must be done in a validated software package. At many pharmaceutical companies, this package is SAS.

This paper will review and demonstrate:

- the dataset format and structure used to generate the forest plot
- creation of the annotate dataset
- application of the annotate dataset in the GPLOT procedure

THE REQUEST

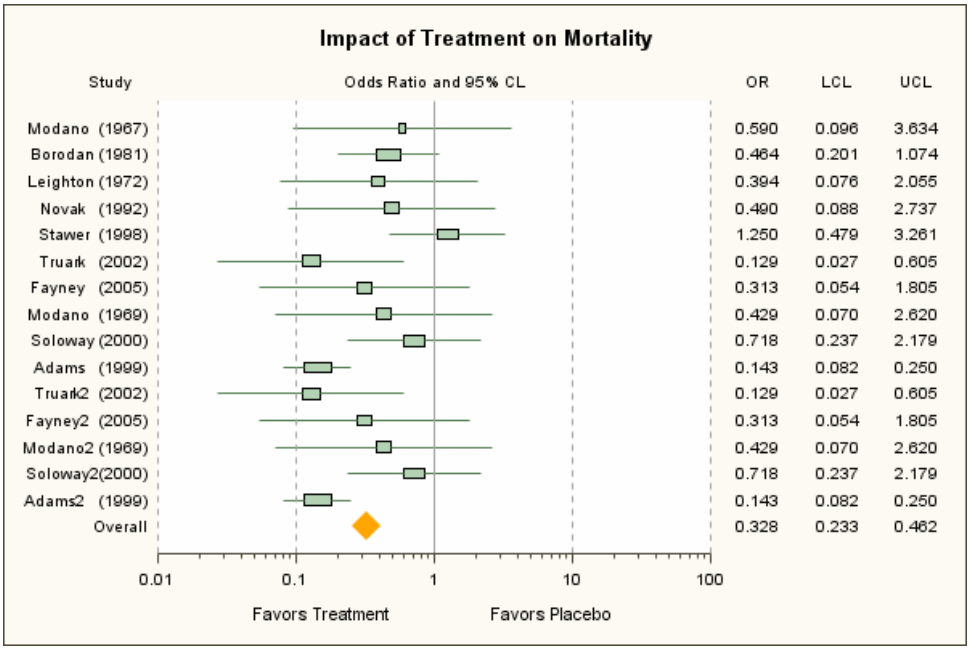
As part of a Phase III study at our company, the statisticians included a graphic of a forest plot in the table, listing and figure (TLF) shells for the Clinical Study Report. The purpose of the graph was to compare the risk difference and associated variability in achieving pain-free status between the placebo and treatment arms across a variety of demographic characteristics and categorical factors associated with disease severity. In this analysis, the risk difference represented the absolute difference in percent of people achieving a pain-free state between those in the placebo and active treatment groups. For example, if a pain-free state was achieved by 35% of people in the placebo group and 50% of people in the active group, the risk difference would be 0.15 (0.50-0.35=.15). A risk difference of 0 would indicate that there was no discernable difference between the groups. Hence, if the level of variability was such that the 95% confidence interval included 0 it would indicate that there was not a statistically significant difference between the two groups at the α=0.05 level of significance.

The categorical groupings being examined in this analysis were mutually exclusive and are shown in the following table:

Group	Category
1: Age	Age<40

	Age>=40
2: Sex	Sex: Female
	Sex: Male
3: BMI	BMI <25
	BMI >=25
4: Duration of Disease	Duration of Disease: <2 years
	Duration of Disease: >=2 - <5 years
	Duration of Disease: >=5 - <10 years
5: Race	Race: Asian
	Race: Black or African American
	Race: White
	Race: Other
6: Country	Country: United Kingdom
	Country: Brazil
	Country: Canada
	Country: Sweden
	Country: United States
7: Heart Rate	Heart Rate: <=60
	Heart Rate: >60
8: Systolic Blood Pressure	Systolic Blood Pressure: <=120
	Systolic Blood Pressure: >120
9: Temperament	Temperament: Type A
	Temperament: Type B
	Temperament: Type C
	Temperament: Type D

A similar type of figure had been generated by a partner company for another of our company's products and was used in the TLF shells to demonstrate the desired result. The figure that was presented to us looked similar to the example shown below, available as sample 35143 within SAS Knowledge Base online.



This example illustrates a meta-analysis across 15 different studies. Here, the outcome of interest is the odds ratio; a value less than 1.0 represents that the treatment has a favorable effect on mortality. The vertical reference line at 1 allows easy visualization of the effect for each study at the 95% level of confidence. In this example, only 4 of the studies demonstrate that treatment had a significant favorable effect on mortality (Truark, Adams, Truark2 and Adams2).

Since none of the programmers in our company were familiar with generating forest plots, we requested our partner company's code so that we could understand how the graph was created and build our code accordingly. Unfortunately, the code given to us was written in a software package that is not validated at our company so we were forced to begin anew in SAS.

Our strategy was to: determine the parameters necessary for display, create a simple program to manipulate the data to a format compatible with the annotate facility, and use the GPLOT procedure to output the results in a simple linear format.

DATASET FORMAT AND STRUCTURE

To generate the forest plot, we established that we needed a dataset containing the following variables (variable names are given in parentheses):

- Category for analysis (SECTION)
- Total Number of People in Placebo Arm (TOTPLAC)
- Total Number of People in Treatment Arm (TOTTRT)
- Percent of Patients who are Pain Free in Placebo Arm (PERCPLAC)
- Percent of Patients who are Pain Free in Treatment Arm (PERCTRT)
- Risk Difference Comparing Placebo and Treatment Arms (_RDIF2_)
- Lower Bound of 95% Confidence Interval for Risk Difference (L_RDIF2)
- Upper Bound of 95% Confidence Interval for Risk Difference (U_RDIF2)

Using simple SAS code to generate counts and frequencies (PROC SQL) and the risk difference and associated confidence interval (PROC FREQ) by treatment group and factor, we were able to build a macro that could quickly generate our required dataset with minimal code. The factors of interest were included as parameters in the macro call along with a numeric equivalent to facilitate ordering in the final graph. The macro was called once for each grouping of factors. The table below represents the dataset (DSETA1) that resulted from applying the macro to all nine groupings:

RDIF2	L_RDIF2	U_RDIF2	_SECTION	PERCTRT	TOTPLAC	PERCPLAC	TOTTRT	SECTION
0.05	-0.05	0.16	1	42	219	36	140	Age<40
-0.04	-0.49	0.41	2	26	8	28	7	Age>=40
0.08	-0.08	0.24	3	44	74	36	65	Sex: Female
0.04	-0.09	0.17	4	40	153	36	82	Sex: Male
0.07	-0.09	0.23	5	42	95	34	55	BMI <25
0.04	-0.09	0.17	6	40	132	36	92	BMI >=25
0.05	-0.18	0.29	7	38	43	32	25	Duration of Disease: <2 years
0.50	0.01	0.99	8	50	4			Duration of Disease: >=2 - <5 years
0.04	-0.07	0.15	9	40	179	36	119	Duration of Disease: >=5 - <10 years
-0.15	-0.53	0.22	10	38	13	54	13	Race: Asian
0.17	0.00	0.35	11	40	64	24	43	Race: Black or African American
0.07	-0.12	0.25	12	44	72	36	44	Race: White
-0.02	-0.20	0.16	13	38	78	40	47	Race: Other
0.08	-0.08	0.25	14	42	85	34	56	Country: United Kingdom
-0.03	-0.27	0.20	15	32	34	36	28	Country: Brazil
0.08	-0.25	0.42	16	66	24	58	12	Country: Canada
-0.03	-0.33	0.26	17	32	22	36	17	Country: Sweden
0.06	-0.13	0.25	18	36	62	30	34	Country: United States
0.08	-0.11	0.26	19	42	60	34	44	Heart Rate: <=60
0.02	-0.11	0.14	20	40	154	38	94	Heart Rate: >60
-0.05	-0.30	0.19	21	42	34	46	30	Systolic Blood Pressure: <=120
0.06	-0.05	0.18	22	40	180	34	108	Systolic Blood Pressure: >120
0.01	-0.27	0.28	23	30	30	30	17	Temperament: Type A
0.07	-0.09	0.23	24	42	93	36	57	Temperament: Type B
0.10	-0.21	0.40	25	42	28	34	15	Temperament: Type C
0.05	-0.12	0.22	26	42	75	38	58	Temperament: Type D

THE ANNOTATE DATASET

Once we had generated our summary dataset (DSETA1), we were ready to annotate. To do so, we needed to build an annotate dataset. The ANNOTATE facility in SAS/GRAPH allows a programmer to customize graphical output by specifying position (horizontal and vertical coordinates on a page) and instruction for what function should be performed at that position (e.g. DRAW, create a LABEL, etc.) in the annotate dataset. The ANNOTATE facility requires that specific variable names contain instructions that can be translated into actions during the graphing procedure. Annotate variables that were needed in the creation of our forest plot are shown in the table below:

FUNCTION	MOVE - Moves to a point, generally in preparation for drawing DRAW - Draws a line LABEL - Adds text to graph
STYLE	Specifies font for text
COLOR	Specifies the color when drawing/labeling
SIZE	Specifies the size when drawing/labeling
X	Specifies the X coordinate where a function should be performed
Y	Specifies the Y coordinate where a function should be performed
YC	Character equivalent of Y that is used when coordinate system is based on data values and the Y axis values are character (rather than numeric)
XSYS	Defines the area and coordinate system used by X
YSYS	Defines the area and coordinate system used by Y
HSYS	Defines the coordinate system used by SIZE
POSITION	Defines placement of text in LABEL function
TEXT	Specifies text to use in the label

The results dataset, DSETA1, served as the base for building the annotate dataset. It provided the x and y coordinates, the labels for the tick marks on the y axis (SECTION), and the results for display. For each SECTION (e.g. subgroup), multiple actions were described to instruct SAS/GRAPH on where to position parameters, how to draw lines, and how to apply labels when the GPLOT procedure was called. The steps for each SECTION (reflected in the SAS code that follows) were to:

1. Draw the horizontal line from the lower bound (L_RDIF2) to the upper bound (U_RDIF2) of the confidence interval.
2. Draw the tick line for the lower bound (L_RDIF2)value
3. Draw the tick line for the upper bound (U_RDIF2) value
4. Print the risk difference as text
5. Print the 95 percent confidence interval as text
6. Print subgroup name on left side of graph as text
7. Print number of patients in placebo arm as text
8. Print number of patients in treated arm as text
9. Print percent of patients who were pain-free in placebo arm as text
10. Print percent of patients who were pain-free in treated arm as text

Each step was accomplished using combinations of the move, draw and label functions. Function records also needed to be generated to build the headers which explained the contents of the graph (e.g. “Estimate”, “95% Confidence Interval”, “Placebo”, “Treated”, etc.). The code below represents the creation of the actual annotate dataset:

```
%*** CREATE ANNOTATE DATASET:DRAW LINES FOR CONFIDENCE INTERVALS, APPLY SUBGRP LABELS;

data anno;
  length function style color $20 text $60 ;
  retain xsys ysys '2' when 'a';
  set dseta1;

  /* Draw the horizontal line from l_rdif2 to u_rdif2 */
  function='move'; xsys='2'; ysys='2'; yc=section; x=l_rdif2; color='black'; output;
  function='draw'; x=u_rdif2; color='black'; size=1; output;

  /* Draw the tick line for the l_rdif2 value */
  function='move';xsys='2'; ysys='2';yc=section; x=l_rdif2; color='black'; output;
  function='draw';x=l_rdif2; ysys='9'; y=+.5; size=1; output;
  function='draw';x=l_rdif2; y=-1; size=1;output;
```

```
/* Draw the tick line for the u_rdif2 value */
  function='move';xsys='2'; ysys='2'; yc=section; x=u_rdif2; color='black'; output;
  function='draw';x=u_rdif2; ysys='9'; y=+.5; size=1; output;
  function='draw';x=u_rdif2; y=-1; size=1; output;

  /* Print risk difference */
  function='label';   style="'Times New Roman'"; xsys='2'; ysys='2'; hsys='1';
       position='6'; size=3;x=-2.1; yc=section; _rdif2=round(_rdif2_,.001);
       text=strip(put(_rdif2,best.)); output;

  /* Print 95 percent CI. */
  function='label'; style="'Times New Roman'"; xsys='2'; ysys='2'; hsys='1';
       position='6'; size=3; x=-1.6; yc=section; _l_rdif2=round(l_rdif2,.001);
       _u_rdif2=round(u_rdif2,.001);   text= "("||strip(put(_l_rdif2,best.))||","
       ||strip(put(_u_rdif2,best.))||")"; output;

  /* Print header for estimate.  */
  function='label'; style="'Times New Roman'"; xsys='2'; ysys='2'; hsys='1';
       position='6'; size=3;x=-2.1; yc="Group"; text= "Estimate"; output;

  /* Print header for CI */
  function='label'; style="'Times New Roman'"; xsys='2'; ysys='2'; hsys='1';
       position='6'; size=3; x=-1.6; yc="Group"; text= "(95%CI)"; output;

  /* Print all subgroups on left hand side of graph */
  function='label'; style="'Times New Roman'"; xsys='2'; ysys='2'; hsys='1';
       position='6'; size=3; x=-4.6; yc=section; text=section; output;

  /* Print placebo for treatment group header */
  function='label'; style="'Times New Roman'"; xsys='2'; ysys='2'; hsys='1';
       position='6'; size=3; x=1.5; yc="Title"; text= "Placebo"; output;

  /* Print N for header */
  function='label'; style="'Times New Roman'"; xsys='2'; ysys='2'; hsys='1';
       position='6'; size=3; x=1.5; yc="Group"; text= "N"; output;

  /* Print percent in response for header */
  function='label'; style="'Times New Roman'"; xsys='2'; ysys='2'; hsys='1';
       position='6'; size=3; x=1.8; yc="Group"; text= "% Pain Free"; output;

  /* Print vedolizumab for treatment group header */
  function='label'; style="'Times New Roman'"; xsys='2'; ysys='2'; hsys='1';
       position='6'; size=3; x=2.5; yc="Title"; text= "Treated"; output;

  /* Print N for treatment group header. */
  function='label'; style="'Times New Roman'"; xsys='2'; ysys='2'; hsys='1';
       position='6'; size=3; x=2.5; yc="Group"; text= "N"; output;

  /* Print number of patients in placebo for given subgroup. */
  function='label'; style="'Times New Roman'"; xsys='2'; ysys='2'; hsys='1';
       position='6'; size=3; x=1.5; yc=section; text= strip(put(totplac,best.));
       output;

  /* Print percent of placebo patients pain-free for given subgroup. */
  function='label'; style="'Times New Roman'";xsys='2'; ysys='2'; hsys='1';
       position='6'; size=3;x=1.8; yc=section; text= strip(put(percplac,best.));
       output;

  /* Print number of patients in treated arm for given subgroup. */
  function='label'; style="'Times New Roman'"; xsys='2'; ysys='2'; hsys='1';
       position='6'; size=3;x=2.5; yc=section; text= strip(put(tottrt,best.)); output;

  /* Print percent of treated patients pain-free for given subgroup. */
  function='label'; style="'Times New Roman'"; xsys='2'; ysys='2'; hsys='1';
       position='6';size=3;x=2.8; yc=section; text= strip(put(perctrt,best.)); output;
run;
```

APPLICATION OF THE ANNOTATE DATASET TO PRODUCE THE FOREST PLOT

Once the annotation dataset was created, the remainder of the code needed for generating the forest plot was very simple. We had created a macro variable (&DISPLAY) to represent all of the subgroups that had been analyzed for use in the order statement for the y axis. We plotted SECTION on the y axis and the risk difference point estimate on the x axis, and included the ANNOTATE option—applying the ANNO dataset—in the PLOT statement. To aide in interpretation, we added a vertical reference line at the position of 0 on the x axis (representing no risk difference between the groups) by specifying an option of HREF=0. The actual code used to generate the plot follows:

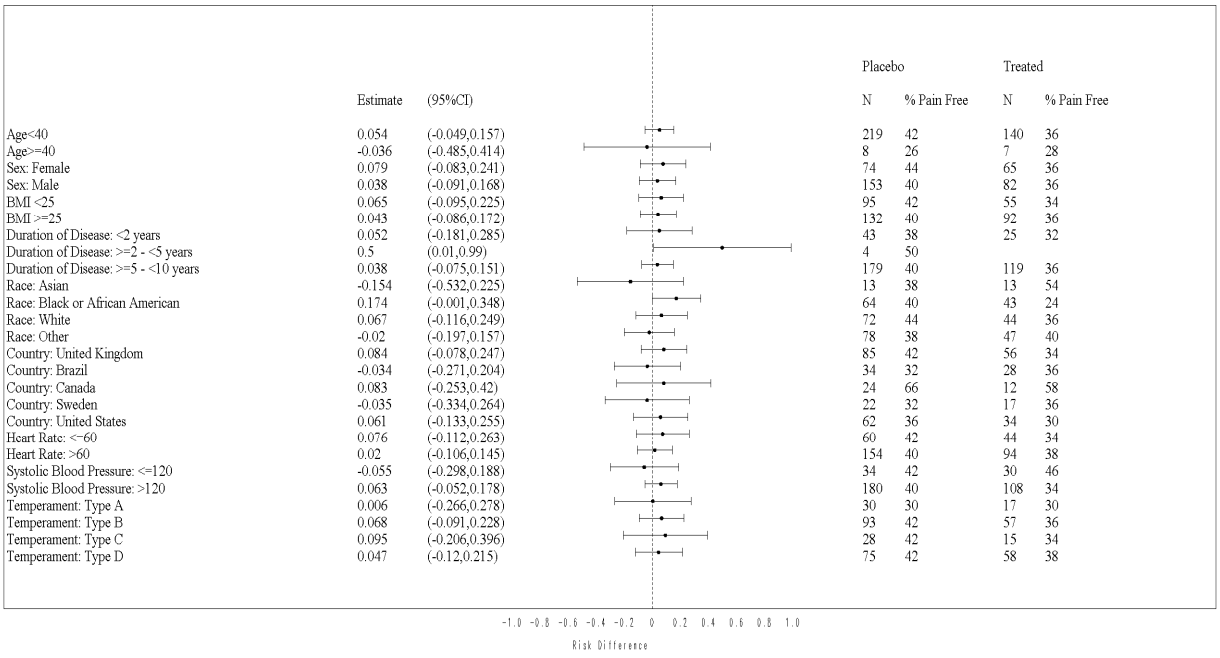
```
axis1 order=( &display  'Group' ' ' 'Title')
      label=none
      value=none
      major=none
      minor=none
      offset=(5,5);

axis2 order=(-4.6 to 4 by 0.2)
      label=('Risk Difference'
      font='Times New Roman'
      height=2.5)
      minor=none
      major=none;

proc sort data=dseta1;
    by _section;
run;

options nobyline;
proc gplot data=dseta1;
   plot section*_rdif2_ / annotate=anno
        nolegend
        vaxis=axis1
        haxis=axis2
        href = 0
        lhref = 2;
   symbol1 interpol=none color=black value=dot height=.5;
run;
quit;
```

THE RESULT: OUR FOREST PLOT



CONCLUSIONS

By using the annotate facility and PROC GPLOT, programmers can generate a forest plot comparing treatment effect across multiple groups. The forest plot may prove useful in later stages of clinical trials or at the time of a regulatory submission when comparisons need to be made across multiple studies. While the programming was written specifically to meet the needs of one clinical program at our company, it could easily be standardized and developed into a macro with wider applicability.

REFERENCES

SAS® Knowledge Base / Samples & SAS Notes—Sample 35143: Forest Plot: http://support.sas.com/kb/35/143.html

CONTACT INFORMATION

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

Amanda Tweed
Millennium: The Takeda Oncology Company
40 Landsdowne Street
Cambridge, MA 02140
Email: amanda.tweed@mpi.com