Paper 3419 - 2015

Forest Plotting Analysis Macro %FORESTPLOT

Jeffrey Meyers, Mayo Clinic, Rochester, Minnesota
Qian Shi, PhD, Mayo Clinic, Rochester, Minnesota

ABSTRACT

The forest plot is a powerful and versatile tool for visually presenting model estimates for multivariable analysis, or illustrating association measures of key interested factor across various models or subgroup analyses. The ability to see if two values are significantly different from each other, or if a covariate has a significant meaning on its own when compared to a reference value, is made much simpler in a forest plot rather than sifting through numbers in a report table. The amount of data preparation in order to build a high quality forest plot in SAS can be tremendous as the programmer will need to run analyses, extract the estimates to be plots, and structure the estimates in a format conducive to generating a forest plot. This code required for this process is often replicated repeatedly for multiple models, which is inefficient and prone to errors. While some SAS procedures can produce forest plots using the Output Delivery System (ODS) graphics automatically, the plots are not generally publication ready and are difficult to customize even if the programmer is familiar with the Graph Template Language. The macro FORESTPLOT is designed to efficiently and automatically perform all of the steps of building a high quality forest plot, and is currently designed to perform regression analyses common to the clinical oncology research areas, Cox proportional hazards and logistic models, as well as calculate Kaplan-Meier event-free rates and binomial success rates. Additionally to improve flexibility, the user can specify a pre-built data set to transform into a forest plot if the automated analysis options of the macro do not fit the user's needs.

INTRODUCTION

Although the forest plot is a common graphic data presentation in meta-analysis, it is also a powerful and versatile tool for visually presenting model estimates for multivariable analysis, or illustrating association measures of key interested factor across various models or subgroup analyses. For example, when analyzing time-to-event or binary outcomes across multiple subgroups, model estimates, ratios, and rates with confidence limits are graphically stacked vertically on the same plot, and the pattern of the model estimates can be evaluated easily across subgroups. For example, how they overlap with each other and/or with reference values such as 1 (i.e., no association) when hazard ratio or odds ratio are considered. The ability to see if two statistical estimates are significantly different from each other by comparing their confidence intervals, or if a covariate has a significant meaning on its own by comparing its confidence interval to the reference value, is made much simpler in a forest plot rather than trying to compare numbers in a report table. In association studies, it is common to assess whether the estimated association between the key interested factor and outcome vary across level of the third variable (i.e., testing interactions). A forest plot can place the outputs of these subgroup analyses next to each other for immediate comparison. All of the estimates from a multivariate model can also be listed together along with p-values to quickly assess significance and different behaviors across covariates.

The greatest challenge to building a forest plot is a large amount of data preparation which often requires repeating the same multiple steps. These steps include running analyses, extracting the relevant estimates to be plotted, structuring the estimates in a format conducive to generating a forest plot, and creating the plot. Manually repeating these codes is inefficient and can easily lead to programming errors, especially when updating the code. Therefore a streamlined and automatic procedure for creating a forest plot is needed.

The macro *FORESTPLOT* makes generating many types of forest plots simple, efficient, and quick. The macro can repeatedly estimate the statistical parameters required for plotting within subgroups defined by the levels of a variable (i.e. subgroup analysis), or by different models, then extract and put them into a dataset which is suitable for plotting. This macro can also use existing summary statistics as inputs for plotting a customizable graph. The analysis and plotting data set construction is all automated within the macro and customizable by macro parameters. The macro currently is set up to perform regression analyses common to the clinical oncology research areas, Cox proportional hazards and logistic models, as well as calculate event-free rates based on Kaplan-Meier curves and binomial success rates. Upon performing these analyses the macro will then generate a graph template using the Graph Template Language to create a high-quality SAS graphic that is extremely customizable with macro parameters.

SAMPLE DATASETS USED IN EXAMPLES

Time-to-Event Endpoint Analysis (Proportional Hazard Regression or Kaplan-Meier Methods)

All of the proportional hazard regression and Kaplan-Meier examples in this paper use the SASHELP.BMT data set, which is available in SAS 9.3 or later as an example data set. The data set contains survival data on bone marrow transplant patients and has three variables: GROUP, T, and STATUS. The GROUP variable is a discrete categorical variable containing three different disease groups: acute lymphoblastic leukemia (ALL), acute myeloid leukemia (AML)-High Risk, and AML-Low Risk. The T variable is the time from date of transplant to death or censor date, and the STATUS variable is the survival status (0=Alive, 1=Dead).

Table 1 shows the few selected rows of the SASHELP.BMT data set.

Group	T	Status
ALL	2081	0
ALL	1602	0
ALL	1496	0
ALL	1462	0
…	…	…
AML-Low Risk	2569	0
AML-Low Risk	2506	0
AML-Low Risk	2409	0
AML-Low Risk	2218	0
…	…	…
AML-High Risk	2640	0
AML-High Risk	2430	0
AML-High Risk	2252	0
AML-High Risk	2140	0

Table 1. The GROUP variable is a discrete categorical character variable with three levels: ALL, AML-High Risk, and AML-Low Risk. The T variable is a numeric variable representing time from transplant to death or censor date. The STATUS variable is a numeric variable representing survival status where 0 means alive and 1 means dead.

Binary Endpoint Analyses (Logistic Regression or Binomial Success Rate)

All of the logistic regression eamples in this paper use the NEURALGIA dataset taken from the SAS webpage[1]. The dataset has five variables: TREATMENT, SEX, AGE, DURATION and PAIN. The TREATMENT variable is a three level categorical covariate with levels: P (placebo), A (treatment A) and B (treatment B). The SEX variable is a two level categorical covariate with levels: male (M) and female (F). The PAIN variable is a two level categorical covariate with levels of Yes and No.The AGE and DURATION variable are continuous covariates. The PAIN, a binary variable, is considered as the dependent variable in the anlaysis.

Table 2 shows the first several rows of the SASHELP.BMT data set.

Treatment	Sex	Age	Duration	Pain
P	F	68	1	No
B	M	74	16	No
P	F	67	30	No
P	M	66	26	Yes
B	F	67	28	No
B	F	77	16	No
A	F	71	12	No
B	F	72	50	No
B	F	76	9	Yes
A	M	71	17	Yes
A	F	63	27	No
A	F	69	18	Yes
B	F	66	12	No

Table 2. The TREATMENT variable is a discrete categorical character variable with three levels: A, B and P.

The SEX variable is a discrete categorical variable with two levels: F and M. The AGE and DURATION variables are numeric variables.The PAIN variable is a discrete categorical variable with two levels: Yes and No. The PAIN variable is the primary outcome variable within this data set.

1.0 MACRO OVERVIEW

1.1 PLOT OVERVIEW

The output plot has three panels: the subtitle panel, the plot panel, and the statistical summary panel. The subtitle panel contains descriptive labels for included covariates and their categories which can be presented with flexible formats such as indenting and bold face fonts. The plot panel contains the actual forest plot, where the components of the scatterplot and confidence intervals of the model estimates, with the symbols, colors, and sizes being customizable along with the x-axis. The Third panel contains the summary statisics, which can include: number of patients, number of events, estimates, confidence bounds, concordance indexes, and p-values. There are also options to combine some of these statistics together to save space. For instance, number of patients and number of events can be combined into one column in three different ways. The columns in the statistical summary can be turned on and off and be displayed in any order. Footnotes for p-values are automatically generated by the macro.

Figure 1 is an example plot image showing a number of the customization options available in *FORESTPLOT*

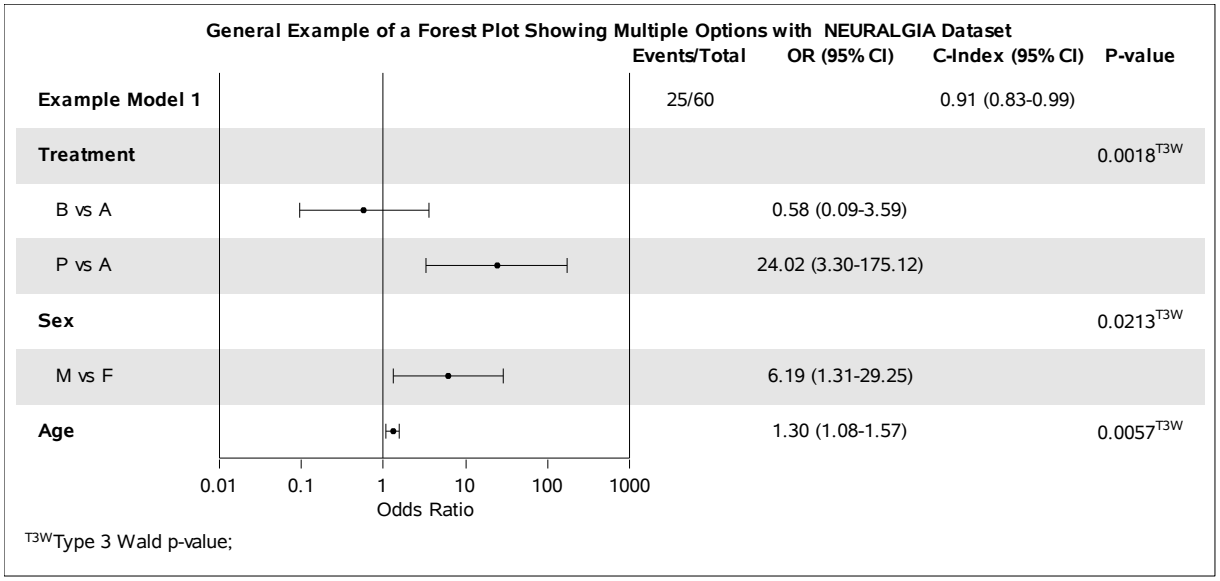

Figure 1. This is an example of showing the output of a single multivariate logistic regression model which assesses the adjusted associations between the covariates (treatment, sex, and age) included in the model and the dependant variable (pain outcome). The events and total column shows the number of events for the entire model. The odds ratio column shows the adjusted odds ratios for each covariate, i.e., comparing pain outcome between patients received treatment A and B to placebo, comparing males to females, and two patients with one year old apart. The concordance index is for the entire model. The p-value column shows the type 3 p-values for each model covariate.

```
%forestplot(DATA=neuralgia,METHOD=LOGISTIC,NMODELS=1,LOGPROC=logistic,
        EVENTCOV=pain,EVENT=Yes,CATCOV=treatment sex,CONTCOV=age,
        MDISPLAY=cat1 cat2 cont1,GPATH=~/ibm/,PLOTTYPE=emf,
        MTITLE=Example Model 1,DISPLAY=ev_t est_range c_est_range pval,
        t3pval=wald,REFLINE=1,HEIGHT=4in,WIDTH=8.5in,CWEIGHTS=0.17 0.35 0.48,
        TITLE=General Example of a Forest Plot Showing Multiple Options with
        NEURALGIA Dataset,LABEL=Odds Ratio,XAXISTYPE=LOG);
```

1.2 MACRO PROCESS OVERVIEW

1.2.1 Step 1: Error Checking

The macro contains a large amount of parameters, so it is necessary to add an error checking code throughout the macro to try to identify inappropriate macro parameter inputs before they cause errors in the SAS session. The error checking code makes sure variabes exist, that required parameters are entered, and that proper values are entered. For instance, if a parameter has a designated list of values, the macro will check whether the user entered an appropriate value. If the user entered value does not match the list, then the macro stops, displays an error message, and provides the list of allowed values.

1.2.2 Step 2: Automated Analysis

Creating a forest plot that compares multiple models requires a large amount of code replication. Nearly identicle code is repeated to run each model, and nearly identicle code is used to extract and combine the results from each model. Programmers not familiar with writing macros will need to spend a great amount of time writing out many of these near duplicate sections of code to create one forest plot. The chance of a programming error also increases when duplicating the same code, especially if later the code needs to be modified. The *FORESTPLOT* macro removes the time investment and risk by fully automating several different types of analyses (based on the *METHOD* parameter): Cox proportional hazards regression, logistic regression, binomial success rate, and Kaplan-Meier event-free rate estimates. Each of these analyses is customizable with macro parameters, and the detail of each method is detailed in section 4. The *FORESTPLOT* macro writes the analysis program for you so that a program that could have taken an hour to write and debug instead takes several minutes to set up and less than a minute to run.

Note that if *METHOD*=DATA, the macro skips the automated analysis section.

1.2.3 Step 3: Plot Data Set Construction

The macro generates a data set that is condusive to creating a forest plot based on either the automated analyses extracted data or a pre-generated data set provided by the user.

1.2.4 Step 4: Generate the Plot

The Graph Template Language within the TEMPLATE procedure is used to set up the plot with a combination of the variables in the plot data set and macro variables derived from the plot data set. The template generated is different depending on whether the user is using SAS 9.2 or SAS 9.3+ due to the evolution of the Graph Template Language between versions. The actual image is then created using the SGRENDER procedure in combination with ODS Graphics option settings. The image can be a number of file types including PNG, EMF, PDF, JPEG, TIFF, and SVG, and can be embedded into RTF, HTML or PDF destinations.

2.0 AUTOMATED ANALYSES

2.1 COX PROPORTIONAL HAZARDS REGRESSION

The PHREG procedure is used to perform Cox proportional hazards regression modeling. Models can also be stratified using the *STRATA* parameter.

2.1.1 Hazard Ratios

Hazard ratios are generated using the HAZARDRATIOS statement and are output with an ODS OUTPUT statement specifying the HAZARDRATIOS data set. The reference group for the categorical covariates is determined by the *CATREF* parameter, and the step size for continuous covariates is determined by the *CONTSTEP* covariate.

2.1.2 Number of Patients and Events

The number of patients are pulled in two different ways. There is a cumulative number of patients and events for the entire model that is generated from the ODS OUTPUT statement specifying the CENSOREDSUMMARY data set. The macro can also calculate the number of patients and events for each level of a categorical covariate by pulling these numbers in a SQL procedure query.

2.1.3 P-Values

There are three levels of p-values that can be calculated by the macro: global model test, type 3 tests for each covariate, and pair-wise comparison within a given covariate when one level is set to be the reference group. The Score, likelihood-ratio, and Wald p-values are available for the global model test, and are ouput using the ODS OUTPUT statement specifying the GLOBALTESTS data set. The Score, likelihood-ratio, and Wald p-values are available for the type 3 tests, and are output using the ODS OUTPUT statement specifying either the TYPE3 data set or the MODELANOVA data set depending on the SAS version (a later release of 9.4 changes the dta set name). The Wald p-value is available to the individual covariate tests, and is output using the ODS OUTPUT statement specifying the PARAMETERESTIMATES data set. Stratified p-values are automatically used when the *STRATA* parameter is used.

2.1.4 Concordance Indexes

The concordance index for Cox proportional hazards regression is not automatically computed with any SAS procedure, and there is not a universally accepted macro to be used for this purpose. The method for calculating concordance indexes described in the survConcordance[1] package from R developed by Therry Therneau is a widely recognized method, and thus was chosen to develop calculation codes in SAS which is included in this macro. The method uses a binary tree approach to calculating the weights, sum of squares, and eventual standard error. The model predicted values used in the binary tree method are taken from the OUTPUT statement defining the XBETA

variable.

2.2 LOGISTIC REGRESSION

Either the LOGISTIC procedure or the GENMOD procedure can be used for logistic regression. The procedure used is determined by the *LOGPROC* parameter.

2.2.1 Odds Ratios

There are two different methods that are used depending on whether the covariate is categorical or continuous. Odds ratios for categorical covariates are calculated with the LSMEANS statement along with the DIFF, CL, and EXP options. These odds ratios are then output using an ODS OUTPUT statement specifying the DIFFS data set. Odds ratios for continuous covariates are calculated with the ESTIMATE statement along with the CL (if LOGISTIC procedure) and EXP options. The odds ratios are then output using an ODS OUTPUT statement specifying the ESTIMATES data set.

2.2.2 Number of Patients and Events

The number patients are pulled in two different ways. There is a cumulative number of patients and events for the entire model that is generated from the ODS OUTPUT statement specifying the RESPONEPROFILE data set. The macro can also calculate the number of patients and events for each level of a categorical covariate by pulling these numbers in a SQL procedure query.

2.2.3 P-Values

The p-values available depend on which procedure is used for analysis.

Type 3 tests for each covariate and pairwise comparisons within a given covariate are available when using the GENMOD procedure. The likelihood-ratio and Wald p-values are are available for the type 3 tests, and are output using the ODS OUTPUT statement specifying either the TYPE3 data set or the MODELANOVA data set depending on the SAS version (a later release of 9.4 changes the data set name). The Wald p-value is available to the individual covariate tests, and is output using the ODS OUTPUT statement specifying the PARAMETERESTIMATES data set. Stratified p-values are not available when using the GENMOD procedure in this macro because the GENMOD procedure can only do exact stratified analyses which this macro does not automate.

Global tests, type 3 tests, and individual covariate tests are available when using the LOGISTIC procedure. The likelihood-ratio and Wald p-values are available for the global model test, and are ouput using the ODS OUTPUT statement specifying the GLOBALTESTS data set. The Wald p-value is available for the type 3 tests, and is output using the ODS OUTPUT statement specifying either the TYPE3 data set or the MODELANOVA data set depending on the SAS version (a later release of 9.4 changes the dta set name). The Wald p-value is available to the individual covariate tests, and is output using the ODS OUTPUT statement specifying the PARAMETERESTIMATES data set. Stratified p-values are automatically used when the *STRATA* parameter is used.

2.2.4 Concordance Indexes

The method for calculating concordance indexes follows the methods described in a paper by JA Hanley and BJ McNeil[2]. While the concordance index for logistic regression can be automatically output from the LOGISTIC procedure, the standard error is not. Without a standard error the confidence bounds for the concordance index cannot be calculated. The paper by Hanley and McNeil provide a method for calculating the standard error that has been commonly used within the Biomedical Statistics and Informatics division at Mayo Clinic. The model predicted values used in this method are taken from the OUTPUT statement defining the XBETA variable.

2.3 KAPLAN-MEIER EVENT-FREE ESTIMATIONS

The LIFETEST procedure is used to calculate the event-free rates.

2.3.1 Event-Free Rates

The TIMELIST option is used to specify the time-point for the event-free rate, and the OUTSURV option in combination with the REDUCEOUT option is used to output the event-free rates to a data set.

2.3.2 Number of Patients and Events

An ODS OUTPUT statement specifying the CENSOREDSUMMARY data set is used to output the number of patients and number events.

2.3.3 P-Values

The logrank test and the Wilcoxon test can be called within the macro for Kaplan-Meier event-free rates when a *BY* parameter is specified. These are generated with a TEST option within a STRATA statement, and output with an

ODS OUTPUT statement specifying the HOMTESTS data set. Stratified versions of these p-values are generated when the *STRATA* parameter is used. These are calculated slightly differently by running a second LIFETEST procedure call and specifying the *STRATA* variables within the STRATA statement. A GROUP option is then used specifying the *BY* variable. The p-values are output in the same method as the unstratified versions.

2.4 BINOMIAL SUCCESS RATE

The FREQ procedure is used to calculate binomial success rates

2.4.1 Success Rates

The TABLES statement is used with the BIN option to generate the binomial success rates. A data step is utilized before hand to create a variable that contains the counts for each level of the binomial variable. This variable is then used in a FREQ statement with the ZEROS option (which forces PROC FREQ to include counts of zero) to avoid errors that can arise with zero percent success rates and 100 percent success rates. The estimates are then output using an OUTPUT statement with the BIN option.

2.4.2 Number of Patients and Successes

The dataset created in section 2.4.1 also contains the number of patients, and using this, along with the success rate, the number of paitents and number of successes can both be calculated.

2.4.3 P-Values

When a *BY* parameter is specified, a p-value can be computed to test significance of the different success rates across groups. Either a Chi-square p-value or a Fisher exact p-value can be calculated, and these are computed by adding the CHISQ and FISHER options to the TABLES statement. These are output to the same table mentioned in 2.4.1 by adding the CHISQ and FISHER options to the OUTPUT statement.

2.5 PRE-GENERATED DATA

The statistics displayed for pre-generated data all come from the inputted data itself. The variables containing these statistics are pointed to using several macro parameters. The format that the input data set is required to be in is described in section 5.

2.5.1 Estimates

The *VESTIMATE*, *VLCL*, and *VUCL* parameters point to the estimate, lower confidence limit and upper confidence limit respectively. These must be numeric variables.

2.5.2 Number of Patients and Events

The *VTOTAL* and *VEVENTS* parameters are optional and point to the number of patients and the number of events respectively. These must be numeric columns.

2.5.3 P-values

The *VPVAL* parameter is an optional parameter that points to p-values. This can be a numeric or character variable.

2.5.4 Other Statistics

The *VOTHER* parameter exists to give the flexibility to add statistics not covered by the macro to the plot. These can be numeric or character columns, and mutiple variables can be pointed to by specifying a list separated by spaces. The column headers for these variables is determined by the *VOTHERLABELS* parameter.

3.0 PLOT DATASET CONSTRUCTION

3.1 HOW THE DATASET IS CONSTRUCTED

The plot dataset is constructed using the SQL procedure. A blank data set is created with the columns needed using a CREATE TABLE statement, and then macro loops are combined with INSERT statements to add rows to the plot data set. In the case of calculated analysis, the INSERT statements make use of the SET option in combination with subqueries to pull the outputted data sets from section 4. When specifying pre-generated data, the INSERT statements make use of the SELECT statement to pull a query from the inputted data set.

3.2 VARIABLE COMPONENTS

The variables that make up the plot data set are separated into different subtypes.

3.2.1 Row headers

Each row of the forest plot has a rowheader, or subtitle, to show the variable label or value. The row header is contained within one variable:

- SUBTITLE: Contains the row header such as variable label, variable level, group label, group level, or model title.

3.2.2 Estimates and Confidence Limits

There are numeric variables for the each calculated estimate, such as hazard ratio or odds ratio, and the upper and lower confidence limits. These variables are used for making the scatterplot and error bars in the plot.

- ESTIMATE: Calculated ratio/estimate/rate
- LCL: Calculated lower 95% confidence limit
- UCL: Calculated upper 95% confidence limit

When running a logistic model odds ratios and concordance indexes are both calculated, so in the dataset there are additional variables created to contain both of these statistics. These variables are the same but with a prefix. For example, the variables for hazard ratios would be HR_ESTIMATE, HR_LCL, and HR_UCL. Whichever statistic is being plotted is then copied into the ESTIMATE, LCL and UCL variables. Each of these variables also has a character variable equivalent. The character versions of these columns are specially formmated to be used for displaying statistics in the plot. There are additional variables that are combinations of these variables:

- RANGE: LCL - UCL. Example: 0.25-4.20
- EST_RANGE: ESTIMATE (LCL - UCL). Example: 1.34 (0.25-4.20)

3.2.3 Number of Patients and Events

The section for number of patients and events has numeric versions of these counts as well as character versions of these counts to use in the summary statistics panel of the plot. In addition to these are additional variables that can be used in the summary statistics panel with special formats to save space:

- TOTAL: Total number of patients withing model/group
- EVENTS: Number of events or successes (for binomial) within model/group
- EV_T: Events/Total. Example: 245/300
- PCT: Percentage %. Example: 60%
- EV_T_PCT: Events/Total (Percentage %). Example: 245/300 (81.6%)

3.2.4 P-Values

There is one variable in the plot data set for the p-values. This is a character variable, and the macro can recognize that a footnote exists if the variable is in the following format:

- #.###"{sup "4"}". Example: 0.0032”{sup “F”}” will turn into 0.0032^{F}

3.2.5 Plot Indicators

There are three different indicator variables in the plot data set to serve the following purposes:

- SUBIND: determines the number of indentations of the subtitle
- BOLDIND: determines if the subtitle is bold or normal weight
- SHADEIND: determines if the row in the plot has a shade background when *SHADING*=2

These variables are either automatically calculated by the macro for automated analyses, or can be provided in the input data set for *METHOD*=data by using the *VSUBIND*, *VBOLDIND*, and *VSHADEIND* parameters.

3.3 REASONING FOR THIS FORMAT

The data set is designed to be a tabular view of the forest plot in that each row in the data set corresponds to a row in the graph and the columns of the data set align with the columns of the graph (see figure 2)

Figure 1Figure 2 is highlighted in red with the column names from the data set used to create it.

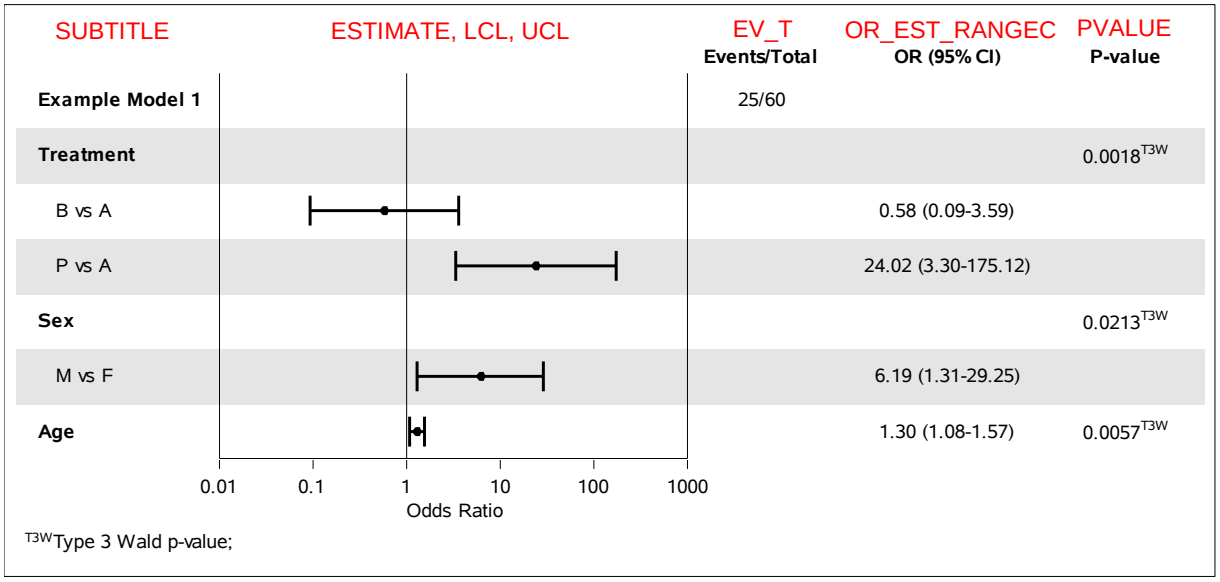

Figure 2. The Y column is the row number within the data set. SUBTITLE gives the row headers. ESTIMATE, LCL, and UCL give the values for the scatterplot and confidence bounds. EV_T, OR_EST_RANGEC, and PVALUE give the values for the statistics

Table 3 shows part of the data set used to make figure 2

y	subind	boldind	subtitle	estimate	lcl	ucl	ev_t	or_est_rangec	pval
1	0	1					Events/Total	OR (95% CI)	P-value
2	0	1	Example Model 1				25/60		
3	0	1	Treatment						0.0018 "{sup "T3W"}"
4	1	0	B vs A	.5784	.09323	3.589		0.58 (0.09-3.59)	
5	1	0	P vs A	24.0218	3.29514	175.121		24.02 (3.30-175.12)	
6	0	1	Sex						0.0213 "{sup "T3W"}"
7	1	0	M vs F	6.1937	1.31161	29.248		6.19 (1.31-29.25)	
8	0	1	Age	1.3034	1.08001	1.573		1.30 (1.08-1.57)	0.0057 "{sup "T3W"}"

Table 3. SUBIND determines the number of indentations for the SUBTITLE column in the plot. BOLDIND determines if the SUBTITLE is bold font. As can be seen when comparing the plot to the data set, the headers for the statistics tables are actually pulled from the first row of the data set and graph starts at the second row of the data set.

4.0 PLOT GENERATION

4.1 GRAPH TEMPLATE CONSTRUCTION

The macro uses the Graph Template Language within the TEMPLATE procedure in combination with the SGRENDER procedure to produce the final graph. Due to the evolution of the Graph Template Langauge from SAS 9.2 to SAS 9.3 and 9.4, it is much more difficult to produce a customized forest plot in SAS 9.2 than the later versions. Due to the inefficient process the macro uses for SAS 9.2, an alternate version was designed for SAS 9.3 or later. The 9.2 version of the template remains in the macro however to increase the versatility of the macro. Both versions of the template split the graph into three panels: the subtitle panel, the plot panel, and the statistical summary panel.

The subtitle panel contains the row headers for the other two panels and includes; the model titles, the covariate labels, and the levels of the covariate. The macro creates the forest plot in a row-by-row basis. Each row of the graph will have a row header (or subtitle), a possible scatterplot, and columns of summary statistics. Multiple models can be run and displayed in the plot, and model titles go before any covariates from a model are listed. Covariate labels are listed before the covariate levels. For example, for gender there would be a row with a label such as "Gender" followed by up to two rows (depending on display options) designating a "Male" row and a "Female" row.

4.1.1 SAS Version 9.2

4.1.1.1 Subtitle Panel

One of the key requirements for making the subtitle panel is the ability to make the text left aligned and to be able to indent the text. Unfortunately in SAS 9.2 there was not a data driven plot that could produce left aligned text at specified coordinates. Block plots are close to being able to do so, but there are currently glitches that prevent block plots with a CLASS statement from filling space properly to align with the y-axis. After numerous attempts at circumventing this problem, the solution was to split the scatterplot into a lattice of cells, where the number of rows is equal to the number of rows in the plot data set and there are two columns (see figure x). The reasoning behind this is that there are graph objects called ENTRY text strings that can be aligned within a plot window at any of the 9 anchor points (topleft, top, topright, etc.), and these objects can be left aligned and spaces can be added to the front of strings to create indents. By anchoring these ENTRY text strings to the left of the row cell, they become vertically centered and are left aligned on the row. The reason there are two columns even though there are three panels is that the minimum x-axis offset is used instead to separate the subtitle panel from the plot panel. An axis offset creates a space that the plot cannot enter, however this space is still available to ENTRY statements. The benefit that this gives is that the text from the ENTRY statement can cross over into the plot window if the string is too long for the text allocated. If instead separate columns were used then text strings that were too long would be cut off by the plot panel.

Figure 3a shows the 9 anchor points available to an ENTRY statement

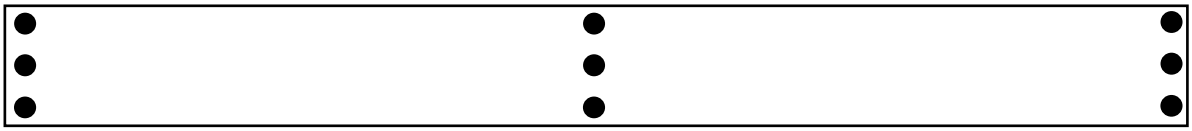

Figure 3a. Text is aligned differently depending on where it is anchored into the graph space. If anchored to the left three dots, the text will be left aligned at the anchor point. If anchored to the middle three dots, the text will be center aligned around the anchor point. If anchored to the right three dots, the text will be right aligned and the anchor point.

Figure 3b shows a lattice of graph cells stacked into a column with multiple rows.

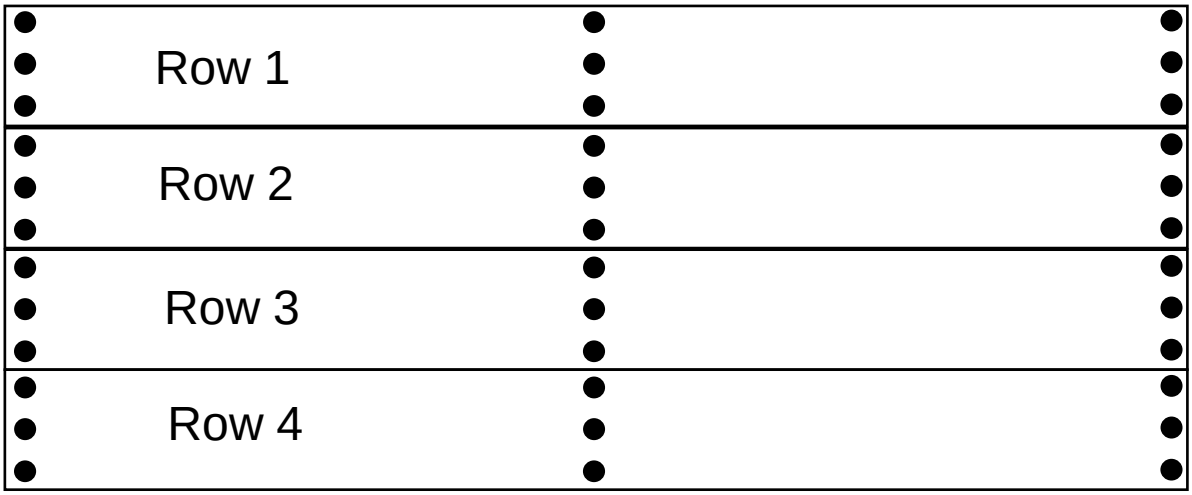

Figure 3b. Each row of the lattice column has its own set of anchor points

Figure 3b Focusing on the left center dot reveals how text can be center aligned across rows of the plot when each row of the lattice has a uniform span for the Y-axis.

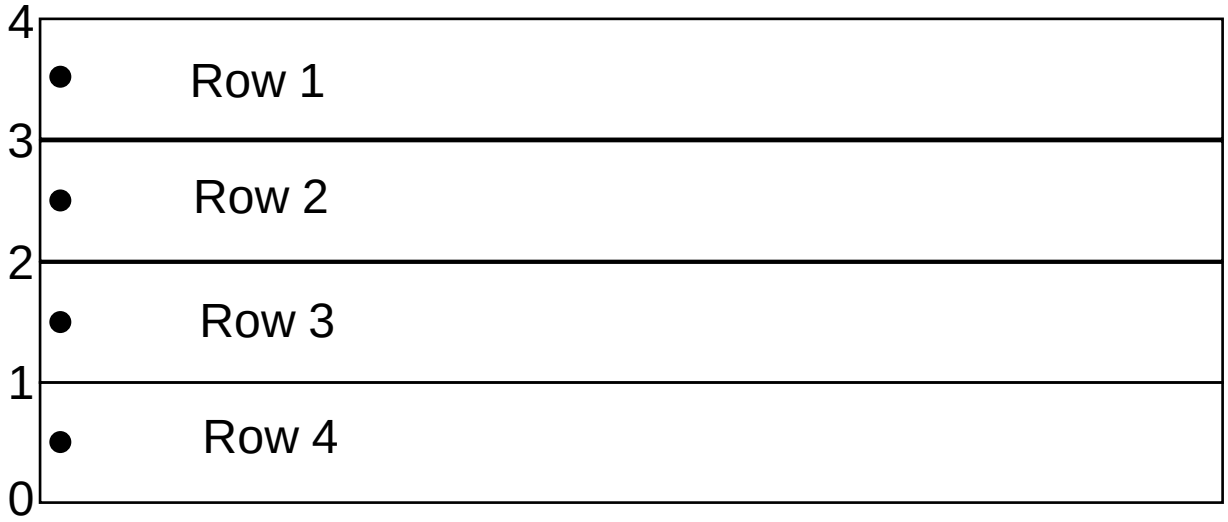

Figure 3c. In this example the y-axis is set up for each individual row of the lattice. Row 4 goes from 0 to 1, row 3 from 1 to 2, row 2 from 2 to 3, and row 4 from 3 to 4. By setting the scatterplot within each row to be plot on the midway point (0.5 in row 4 for example) the left middle anchor point will align perfectly with the scatterplot.

Figure 3d Finally, by removing the borders around the individual rows the graph will take shape.

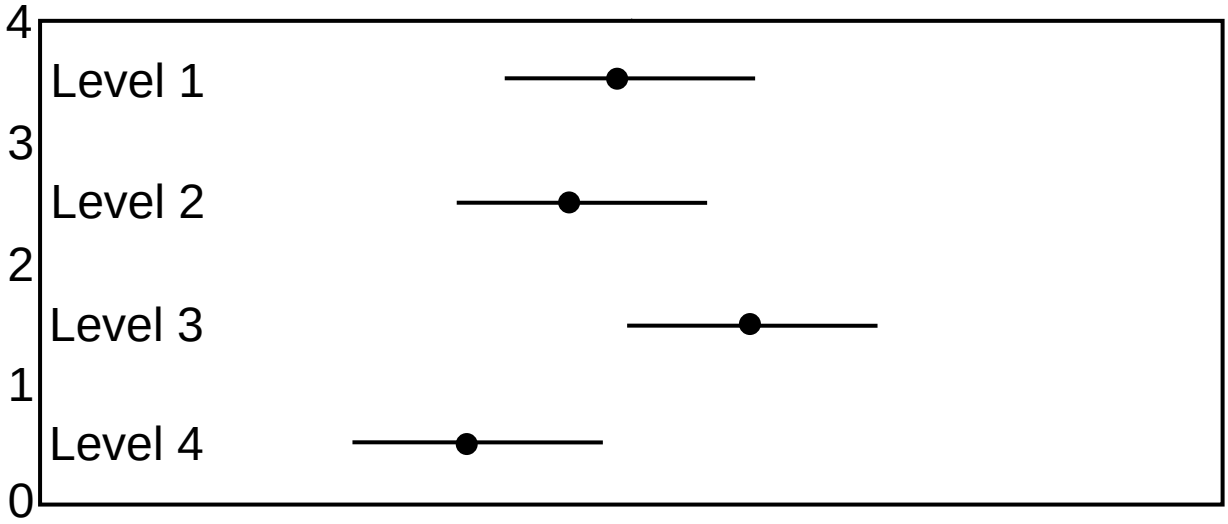

Figure 3d. The borders around the individual rows are removed making the graph appear to be one single graph space. The text "Level 1" through "Level 4" are added using ENTRY statements within each row anchoring at the left middle point. The scatterplots are then added into the plot within each row of the lattice, creating the final image.

4.1.1.2 Plot Panel

The plot panel is made up of four pieces: the scatterplot, the confidence bounds, the reference line and the plot walls. The scatterplot is what draws the forest plot, the reference line is commonly used with odds ratios and hazards ratios to help give a visual confirmation of significance. The scatterplot is drawn using a SCATTERPLOT statement, and there is one SCATTERPLOT statement in each row of the forest plot. The confidence bounds are plotted in two different ways depending on whether the user wants the confidence bounds to have "caps" on the end of the lines. If the user specifies to have caps, then the XERRORLOWER and XERRORUPPER options in the SCATTERPLOT statement. If the user does not want to have caps on the lines then a VECTORPLOT statement is used instead. The reference line and plot walls are drawn with VECTORPLOT statements. The walls are drawn with VECTORPLOT statements because of the offset to the x-axis. If the normal plot walls were used, the y-axis would be pushed to the far left of the plot, which would be to the left of the subtitles. Also, the x-axis is drawn with a VECTORPLOT within the bottom-most cell because the LATTICE layout is created with the COLUMNDATARANGE=UNION option. This

removes the x-axes for each of the internal cells in exchange for one external axis. However if the x-axis line is turned on, it will show the line within each cell. To avoid this, the x-axis line is turned off and a manual one is created with a VECTORPLOT instead.

4.1.1.3 Statistical Summary Panel

The statistical summary panel is contained in the second column of the lattice mentioned in section 4.1.1.1, and actually has a nested lattice within each row of this column (see figure 4). The number of columns within the nested lattice is determined by how many statistics are requested with the *DISPLAY* parameter. The text entries for these summary statistics are also created using ENTRY statements because they do not require axes to be defined within the cell, and can be easily anchored to the center of the cell. Entry statements are also one of the only ways in SAS 9.2 to generate text with unicode, superscript and subscript text within a graph. The widths for these columns are automatically calculated from the maximum width of text within each column divided among the total length available. The user can also specify a set of column weights with the *SUMMARYWEIGHTS* parameter.

Figure 4 is an example plot image showing the SAS 9.2 version of the forest plot

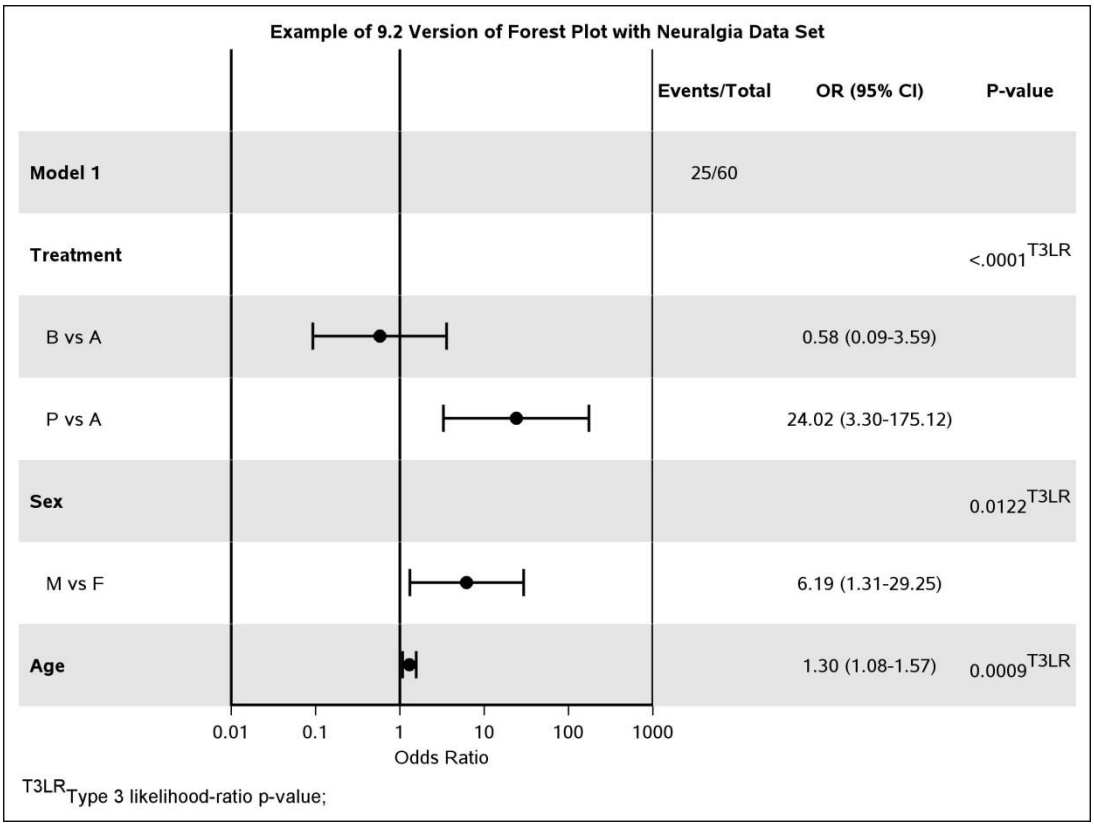

Figure 4. A example of a logistic regression model that is modeling pain against treatment, sex, and age that is displayed using the SAS 9.2 version of the forest plot template.

```
%forestplot (DATA=neuralgia,METHOD=LOGISTIC,NMODELS=1,LOGPROC=genmod,
   EVENTCOV=pain,EVENT=Yes,CATCOV=treatment sex,CONTCOV=age,
   MDISPLAY=cat1 cat2 cont1,CATDISPLAY=1,XAXISTYPE=log,LABEL=Odds Ratio,
   T3PVAL=lr,DISPLAY=ev_t est_range pval,CWEIGHTS=0.2 0.4 0.4,
   TITLE=Example of 9.2 Version of Forest Plot with Neuralgia Data Set,
   MTITLE=Model 1,SYMBOLSIZE=4pt,LINESIZE=0.5pt,
   REFLINE=1,PLOTNAME=9_2example,PLOTTYPE=jpg,DPI=300,GPATH=~/ibm/);
```

Figure 5 Shows example 4 with each row and column of the lattices outlined.

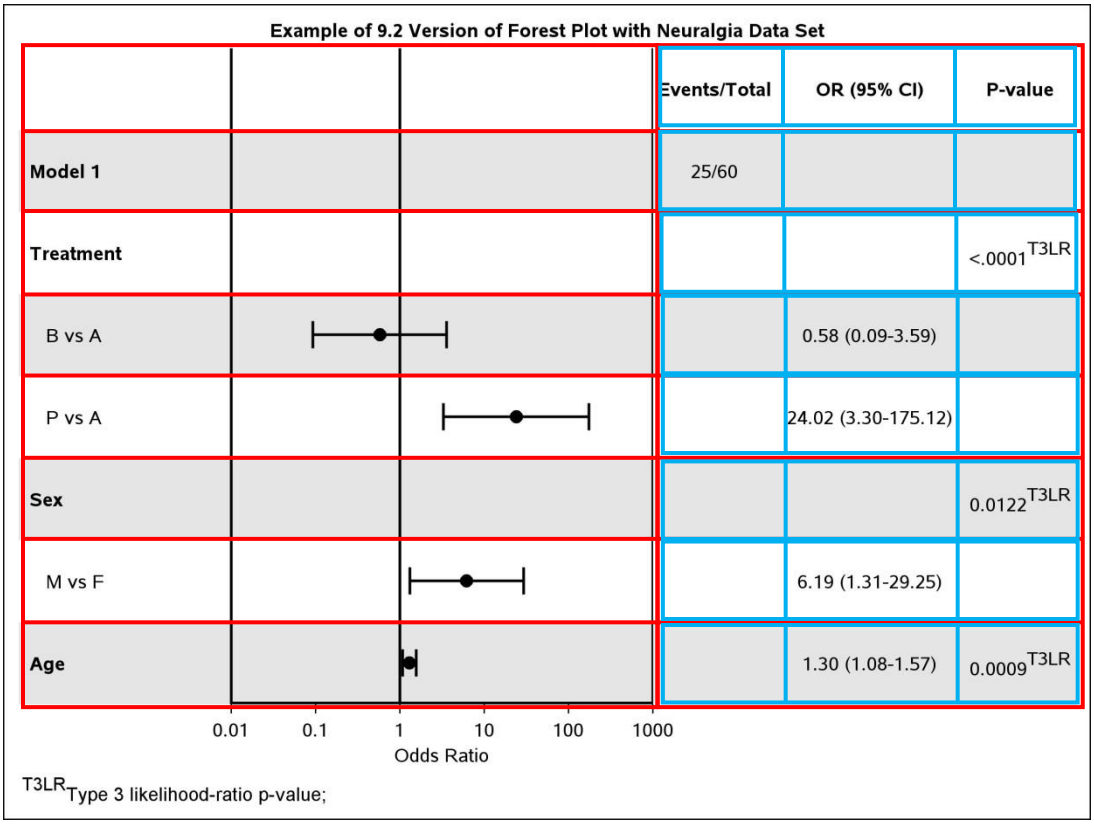

Figure 5. The red outlines show the outermost lattice columns and rows. There are two overall columns with the red outlines, which separate the plot from the statistics. The blue outlines show the inner lattice that is added to each row.

4.1.2 SAS Version 9.3 and 9.4

SAS 9.3 introduced a number of features that when combined with the macro facility streamline the process to making a forest plot. SAS 9.4 added even more features, but due to SAS 9.3 being much more prevalently used and the features in 9.3 still working in 9.4, this version of the template focuses on using 9.3 techniques. One of the most useful features added in 9.3 is the annotation facility for SG graphics. While this macro does not use the annotation facility, it does make use of the functions within GTL that came with it, the DRAW functions. There are numerous DRAW functions available such as DRAWTEXT, DRAWLINE, DRAWARROW, and DRAWPOLYGON. While these are not as useful in data-driven templates, they are incredibly powerful when used in the macro facility. These functions allow the user to manually draw items at specified locations with a multitude of options. Because of these functions, the need to separate each row of the forest plot into rows of a lattice is not necessary. Therefore, the graph space is separated into three panels as a one row three column lattice.

4.1.1.1 Subtitle Panel

The subtitles are drawn entirely with DRAWTEXT statements, which like ENTRY statements can be left aligned and can have spaces added in front for indentation. Unlike ENTRY statements however, the coordinates for the text can be specified with the DRAWTEXT statement. This allows the text to be aligned with the y-axis from the plot panel. Each subtitle has its own DRAWTEXT statement. DRAWTEXT statements are not bound by the walls of a layout or lattice cell, so the same effect of the text being able to overflow into the plot panel is present in this version of the template.

4.1.1.2 Plot Panel

There are only three components to the plot in this version of the template: the scatterplot, the confidence bounds, and the reference line. Because there is not offset in the x-axis, there is no need to draw the plot walls in this version. The reference line is drawn with a REFERENCELINE statement. The scatterplots are drawn with SCATTERPLOT statements, where there is one SCATTERPLOT statement per model called for analysis in the macro. This is done so that the attributes of each model's scatterplot can be altered separately. The confidence bounds are drawn in two ways depending on the *ERRORCAPS* parameter. If caps are requested for the confidence bounds lines, then the majority of the lines are drawn using the XERRORUPPER and XERRORLOWER options within the SCATTERPLOT statement. If the caps on the lines are not requested, then VECTORPLOTS are used to draw the confidence bounds.

4.1.1.3 Statistical Summary Panel

The statistical summary panel in this version does not have nested LATTICE layouts within it, and again relies on DRAWTEXT statements to write the statistics into the graph. There is no need for the nested layouts since the coordinates of the text can be specified when using the DRAWTEXT statements. An algorithm in the macro calculates where the text should be drawn to create the columns of text in the panel. The text is anchored in the center of this coordinate. One benefit of using this method is that the text will not get cut off by the walls of the layout cell like ENTRY statements do since the DRAW functions can write beyond layout barriers.

Figure 6 is an example plot image showing the SAS 9.3+ version of the forest plot.

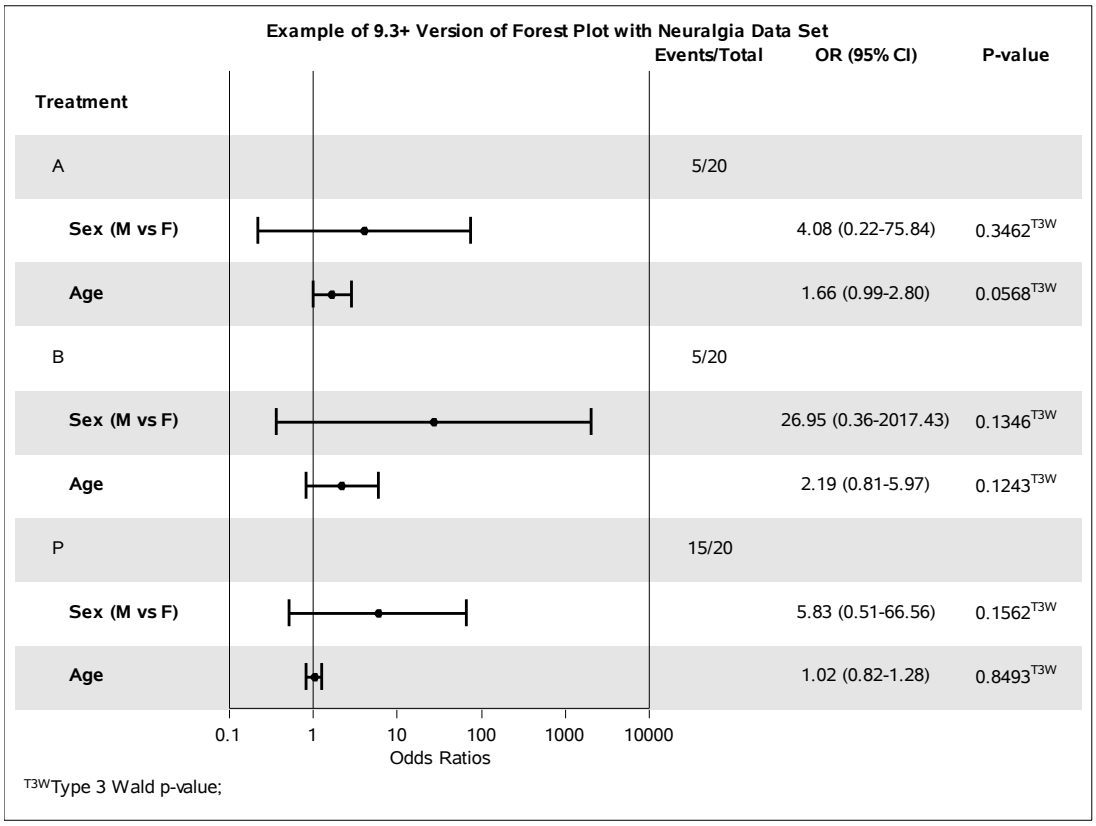

Figure 6. An example of a logistic regression model that is modeling pain against sex and age within the levels of treatment that is displayed using the SAS 9.3+ version of the forest plot template.

```
%forestplot (DATA=neuralgia,METHOD=LOGISTIC,NMODELS=1,LOGPROC=logistic,
   EVENTCOV=pain,EVENT=Yes,CATCOV=sex,CONTCOV=age,
   MDISPLAY=cat1 cont1,CATDISPLAY=2,XAXISTYPE=log,LABEL=Odds Ratios,BY=TREATMENT,
   T3PVAL=wald,DISPLAY=ev_t est_range pval,CWEIGHTS=0.2 0.4 0.4,
   TITLE=Example of 9.3+ Version of Forest Plot with Neuralgia Data Set,
   REFLINE=1,PLOTNAME=9_3example,PLOTTYPE=emf,GPATH=~/ibm/);
```

Figure 7 is an example plot image showing the SAS 9.3+ version of the forest plot with overlays showing the separate cells within the plot.

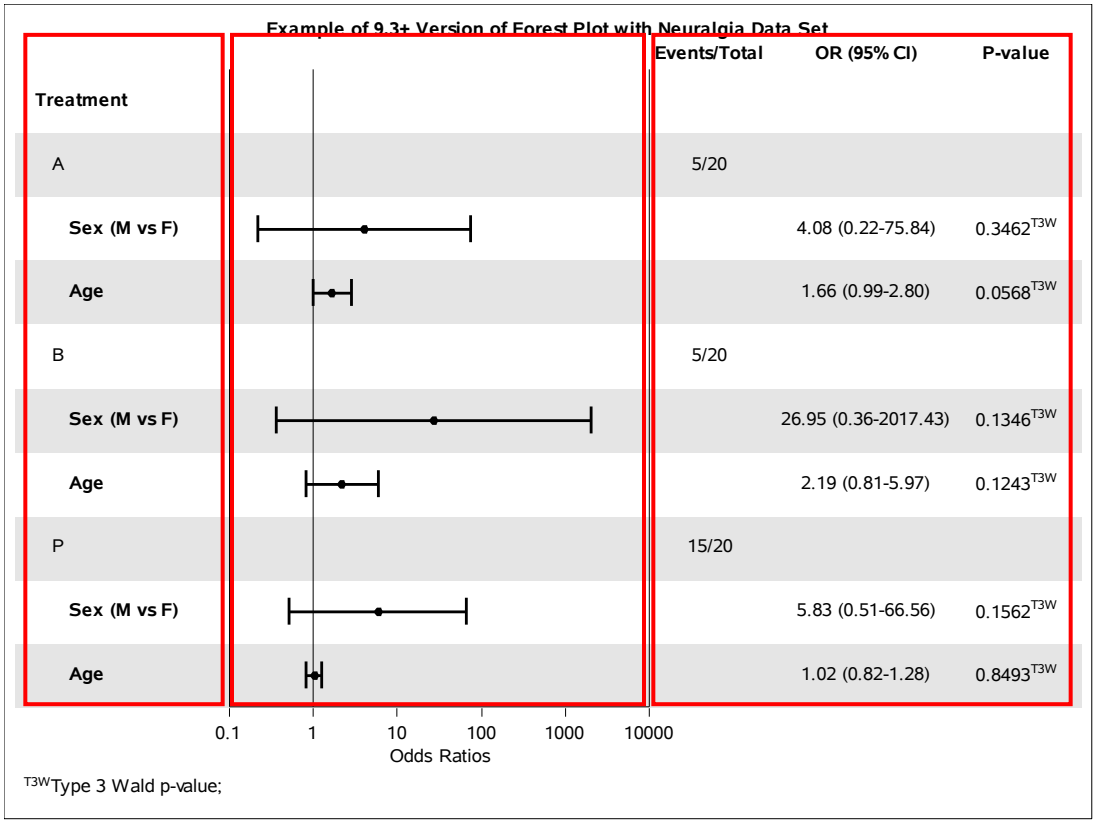

Figure 7. Shows the subtitle panel, the plot panel, and the statistical summary panel for the graph in figure 6 when using the SAS 9.3+ version of the forest plot.

5.0 EXAMPLE GRAPHS

5.1 BINOMIAL SUCCESS RATE EXAMPLE

The first example calculates the rate that patients experience pain in three different ways. The first is across the three treatment arms, the second across the two genders, and the third within all patients pooled together. The example shows off some of the customization options of the plot itself including changing colors, symbols and shading. The plot also turns off the walls separating the subtitle panel and statistical summary panel from the plot panel, and shows how some of the row labels and statistical summary column labels can be modified.

Figure 8 is an example image showing the success rate of a patient having pain across treatment groups and sex, and then listing the overall rate (when ignoring the status of these factors) as a separate symbol.

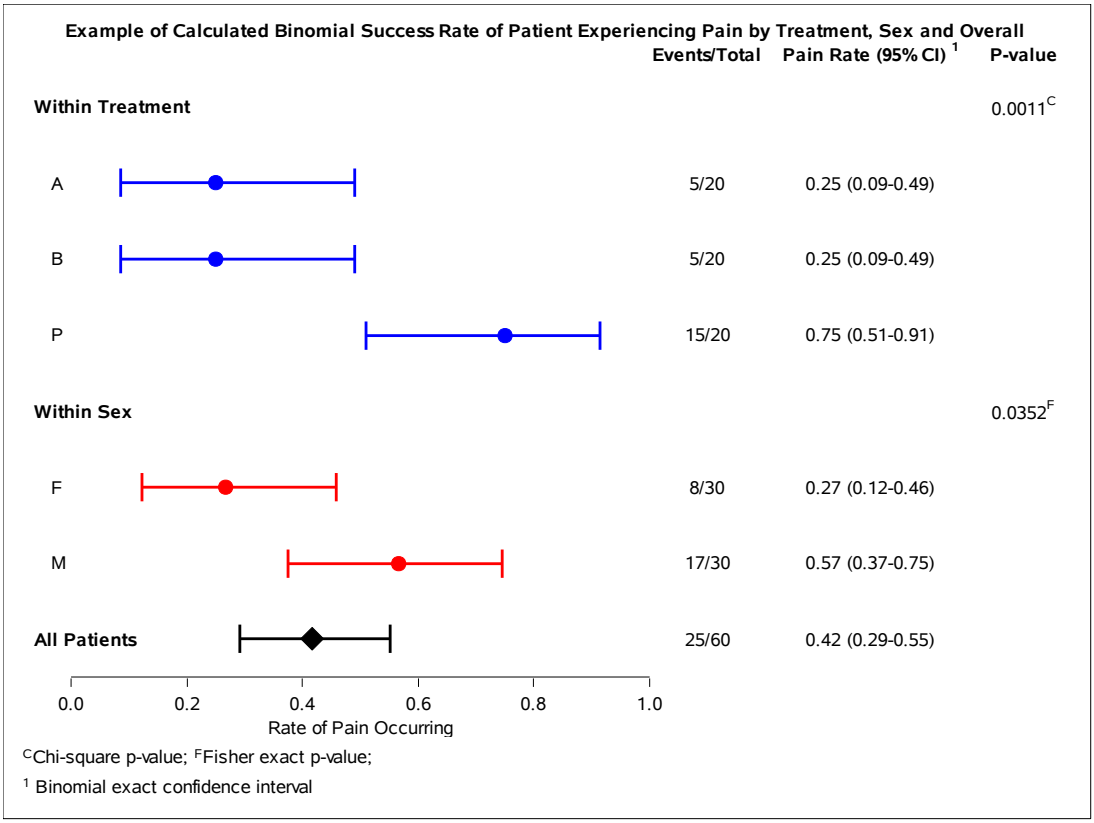

Figure 8. An example of a binomial analysis where the success is the patient experiencing pain. The rate is first calculated across the treatment arms and compared with the Chi-square test. Then the rate is calculated across sex and compared with the Fisher's exact test. Then a pooled rate is calculated and listed with a separate symbol from the others. This image shows how the macro can change the scatterplot color, symbol, and size of different models being run. It also shows that the subtitle text can overlap onto the plot space. Other options shown in this image are the plot walls being turned off, the shading being turned off, and a footnote being added with a footnote marker.

```
%forestplot (DATA=neuralgia,METHOD=binomial,NMODELS=3,EVENTCOV=pain,EVENT=Yes,
TDISPLAY=3,LABEL=Rate of Pain Occurring,BY=TREATMENT|SEX|,
BYORDER=,DISPLAY=ev_t est_range pval,CWEIGHTS=0.05 0.55 0.4,
gpath=~/ibm/,conftype=BE,
title=%str(Example of Calculated Binomial Success Rate of Patient Experiencing Pain by
Treatment, Sex and Overall),
plottype=EMF,modpval=chisq|fisher,shading=0,showwalls=0,
bylabel=Within Treatment|Within Sex|,mtitle=||All Patients,
plotname=binexample,symbolsize=10pt|10pt|12pt,linesize=2pt,
est_range=Pain Rate (95% CI) "{sup "1"}" ,
footnote="{sup "1"}" Binomial exact confidence interval,
symbol=circlefilled|circlefilled|diamondfilled,
symbolcolor=blue|red|black,linecolor=blue|red|black);
```

5.2 KAPLAN-MEIER EVENT-FREE RATE EXAMPLE

The second example calculates the survival rate of patients at three different time points. The example shows how multiple time-point estimates can be called and that a scalar multiplier can be used to transform the supplied time variable. The example also shows how the statistics listed in the statistical summary panel are listed in the order they are called in the *DISPLAY* parameter, that the row headers listed in the subtitle panel can flow into the plot panel, and that any text in the graph can be resized.

Figure 9 is an example image showing the survival rate at 500 days, 6 months, and 1 year across the three disease groups of the BMT data set

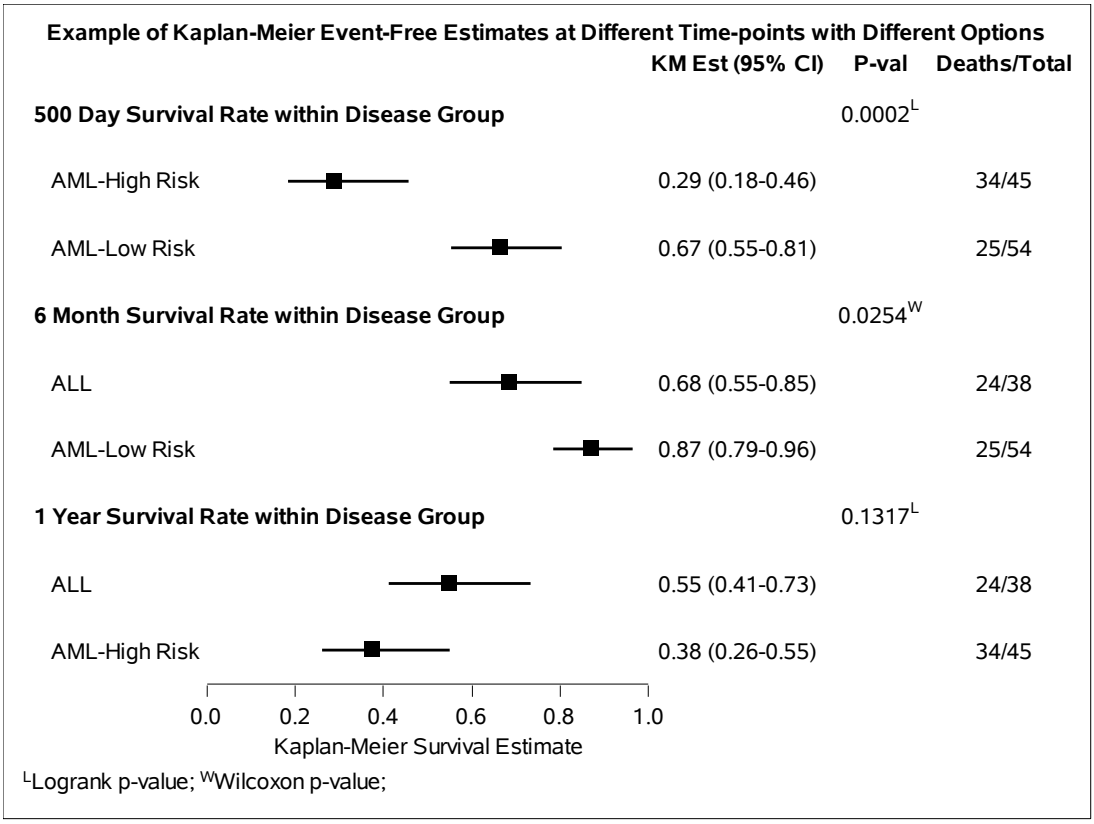

Figure 9. An example to show the flexibility when analyzing Kaplan-Meier event-free rates. There are three separate time-points being analyzed using three different time units. The macro uses parameter *TDIVISOR* to transform time units based on a scalar value, in this case 30.44 to go from days to months and 365.25 to go from days to years. The number of events and the p-values are taken from the total Kaplan-Meier curve and not from the specific time-point. There are separate where clauses subsetting the BMT data set down different for each analysis. The statistics are displayed in the order that they are called in the *DISPLAY* parameter. The font size for any text is adjustable, the header text used in the statistical summary panel is customizable, and the text for each *BY* group's header is customizable.

```
%forestplot(DATA=sashelp.bmt,METHOD=KM,NMODELS=3,
   TIME=t,CENS=status,BY=group,TIMEPOINT=500|6|1,TIMEDX=Days|Months|Years,
   TDIVISOR=|30.44|365.25,
   TDISPLAY=2,XAXISTYPE=linear,LABEL=Kaplan-Meier Survival Estimate,
   T3PVAL=lr,DISPLAY=est_range pval ev_t,CWEIGHTS=0.15 0.42 0.43,
   TITLE=Example of Kaplan-Meier Event-Free Estimates at Different Time-points with
Different Options,
   LINECAP=0,MODPVAL=logrank|wilcoxon|logrank,SHOWWALLS=0,SHADING=0,
   BYLABEL=500 Day Survival Rate within Disease Group|
           6 Month Survival Rate within Disease Group|
           1 Year Survival Rate within Disease Group,
   EST_RANGE=KM Est (95% CI),PVAL=P-val,EV_T=Deaths/Total,
   PLOTNAME=kmexample,PLOTTYPE=emf,GPATH=~/ibm/,SYMBOLSIZE=12pt,SYMBOL=squarefilled,
   LINESIZE=2pt,SUBSIZE=12pt,SUMSIZE=12pt,TITLESIZE=12pt,FNSIZE=12pt,LSIZE=12pt,
   XTICKVALUESIZE=12pt,
   WHERE=group^='ALL'|group^='AML-High Risk'|group^='AML-Low Risk');
```

5.3 CONCORDANCE INDEX EXAMPLE

The third example of plots the concordance indexes from three different logistic regression models in one graph. The first model is a univariate model with TREATMENT as the covariate. The second model adds SEX as an adjusting factor, and the third adds AGE as a second adjusting factor. The first two model outputs lists CAT1 (the first categorical covariate in the model) in the *MDISPLAY* parameter, which causes the variable levels of TREATMENT to be listed. Because *CATDISPLAY* equals five, all levels including the reference are listed, and because *DISPLAY* includes *OR_EST_RANGE* the odds ratios are still listed in the statistical summary panel (showing Reference for the

reference level). The third model does not list any parameters in the *MDISPLAY* parameter, causing just the concordance value to be listed. The p-value is still listed for the TREATMENT variable due to the *MODCOV* parameter being set to CAT1. The automatic footnotes for p-values are disabled with the *AUTOPFOOT* parameter, and footnotes are manually added. Footnote markers are manually added to text with the "{sup "x"}" format.

Figure 10 is an example image showing multiple logistic regression models plotting the concordance index

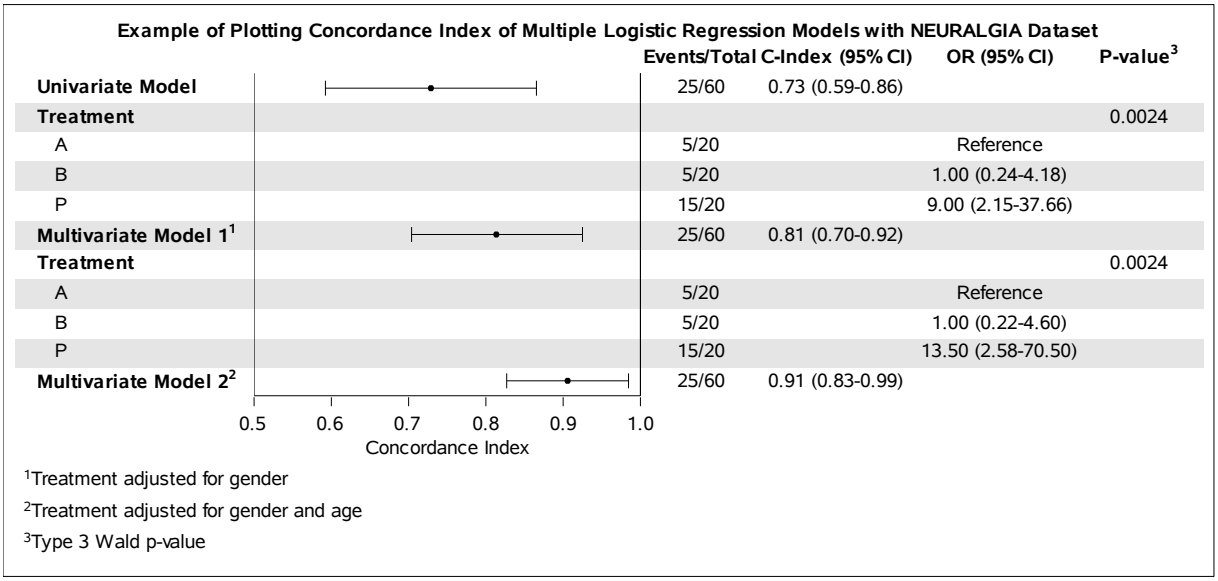

Figure 10. This is an example of plotting the concordance indexes from three different logistic regression models in one graph. The first model is a univariate model with TREATMENT, the second model is a multivariate model with TREATMENT and SEX, and the third is a multivariate model with TREATMENT, SEX, and AGE.

```
%forestplot(DATA=neuralgia,METHOD=LOGISTIC,NMODELS=3,LOGPROC=logistic,
   EVENTCOV=pain,EVENT=Yes,CATCOV=treatment|treatment sex|treatment sex,CONTCOV=||age,
   MDISPLAY=cat1|cat1|,GPATH=~/ibm/,PLOTTYPE=emf,EST_TYPE=C,
   MTITLE=Univariate Model|Multivariate Model 1"{sup "1"}"|
          Multivariate Model 2"{sup "2"}",
   DISPLAY=ev_t est_range or_est_range pval,AUTOPFOOT=0,
   T3PVAL=wald,REFLINE=1,HEIGHT=4in,WIDTH=8.5in,CWEIGHTS=0.2 0.33 0.47,
   TITLE=Example of Plotting Concordance Index of Multiple Logistic Regression Models
         with NEURALGIA Dataset,
   LABEL=Concordance Index,CATDISPLAY=5,MIN=0.5,PVAL=P-value"{sup "3"}",
   FOOTNOTE="{sup "1"}"Treatment adjusted for gender
           `"{sup "2"}"Treatment adjusted for gender and age
           `"{sup "3"}"Type 3 Wald p-value);
```

CONCLUSION

The forest plot is a very powerful tool for quickly and visually analyzing the output from one or multiple models. The process to create one is generally difficult and very tedious, but the macro *FORESTPLOT* makes the process easy and straightforward. There are a multitude of formats and analysis methods available to the macro with more constantly being added. As forest plots become easier to create they become even more useful for analysis, often being able to replace tables as they are easier to read and understand, and thus a more powerful data presentation. Withing a capability of automatically conducting repeated analyses and streamline the extracting and ploting procedures, this macro can become a useful tool in reproducible research. The macro *FORESTPLOT* is constantly improving, with one primary goal of being able to have multiple panels of graphs to compare different models or outcomes side-by-side.

RECOMMENDED READING

- SAS 9.3 Graph Template Language Reference, Third Edition
- SAS 9 PROC TEMPLATE Styles Tip Sheet
- SAS/STAT® 9.2 User's Guide, Second Edition. Specifically the LIFETEST procedure, FREQ procedure, Logistic procedure, GENMOD procedure, and PHREG procedure.

REFERENCES

[1] Therneau T (2014). _A Package for Survival Analysis in S_. R package version 2.37-7, <URL: http://CRAN.R-project.org/package=survival>.

[2] cNeil BJ: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29-36, 1982.

CONTACT INFORMATION

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

Name: Jeffrey Meyers
Enterprise: Mayo Clinic
Address: 200 First Street SW
City, State ZIP: Rochester, MN 55905
Work Phone: 507-266-2711
E-mail: Meyers.jeffrey@mayo.edu / jpmeyers.spa@gmail.com