Paper 139-31

Problems Commonly Associated With Forest Plots Addressed Using High Resolution Graphics in SAS®

Gary A. Foster, St. Joseph's Healthcare, Hamilton, Ontario
Charles H. Goldsmith, St. Joseph's Healthcare & McMaster University, Hamilton, Ontario

ABSTRACT

Forest plots are frequently used to display the results of meta-analyses and have been published in the medical literature since 1988 (Lewis & Clarke, 2001). A review of forest plots reveals that the format chosen can affect the accuracy with which the information is conveyed to the reader. One formatting convention is the use of symbol areas to reflect the relative contributions, or weights, of the studies included in the analysis. Typically a study with a larger weight is represented by a symbol with a proportionately larger area.

The use of symbol areas to convey study weights is not optimal for a number of reasons. First, it is difficult for viewers to accurately decode symbol areas to enable a visual comparison of study weights. Second, as symbol areas increase in size they can obscure other information on the plot. Third, if a logarithmic scale is used, studies with equal weights should, but often do not, have symbols with the same physical dimensions. This issue becomes more apparent as the difference in effect sizes increases.

The macro used to generate the figures in this paper provides SAS® software users with an easy to use program to produce high-resolution forest plots. It will generate forest plots where symbol widths represent study weights, but it will also generate traditional forest plots where symbol areas represent study weights. Regardless of the symbol format or type of scale being used, symbols representing studies of equal weight will have the same dimension anywhere on the plot.

INTRODUCTION

The results of a meta-analysis are commonly published with an accompanying forest plot to visually display the results. A forest plot is a graph that displays information about the studies contributing to the meta-analysis, along with information about the synthesis of those studies. On it, effect sizes and confidence intervals are plotted for each of the studies being evaluated, and with rare exception, it incorporates a method to communicate the relative contribution, or weight, of each study.

By tradition, a square symbol is used on a forest plot to represent a study contributing to a meta-analysis. The area of the symbol represents the weight of the corresponding studies to the meta-analysis, with larger weights having proportionately larger symbol areas. The idea that gave rise to this convention is that researchers sought a method to visually distinguish those studies with a greater contribution from those with a lesser one. By visually decoding the differences in symbol areas it was thought that a viewer would be able to compare the relative contributions of the studies.

Unfortunately, as Cleveland points out, "… area charts do not provide efficient detection of geometric objects that convey information about differences of values" (p. 268, Cleveland, 1994). More generally stated, it becomes increasingly more difficult for a viewer to discern relative differences between two objects as the number of changing dimensions in those objects increase. It is much easier to decode the relative sizes of symbols if they vary in one (i.e., width or height) rather than two (i.e., area) dimensions.

Allowing the symbols to vary in only one dimension will prevent another commonly occurring problem. The large symbol representing a study with a large weight may overlap with or obscure other information on the graph. It could even obscure important information related to the study it represents. If the confidence interval, which is usually narrower for larger studies, was narrower than the width of the symbol and the symbol is filled, then the information

about the confidence interval would be lost. Moreover, if the problem is not pointed out by the author, the reader might erroneously assume that the width of the symbol reflects the width of the confidence interval.

A subtle problem can occur when plotting symbols on a graph that uses a logarithmically scaled x-axis. If the weights of two studies are equal to W, then both symbols should have the same physical dimensions regardless of the effect sizes of the two studies. The symbol for the study with a larger effect size should not be visually smaller than that for the study with a smaller effect size due to the scale being more compressed at the upper end. The visual size of a symbol represents a study's weight and it should not be affected by the type of axis being used.

To overcome the problems associated with these issues, we recommend that several modifications be incorporated into a forest plot. Because it is more difficult to visually decode relative differences in areas than relative differences in objects that vary in one dimension, and because of the problem associated with large symbols, we recommend the use of symbols (rectangles for individual study contributions, and diamonds for summary information) that vary in one dimension: width. All rectangular symbols should have the same height, only their widths should vary relative to the proportional contribution, or weight, of the study. Also, when two or more studies have equal weights, the widths of their rectangles should be the same regardless of the magnitude of the respective effect sizes. Finally, by allowing all symbols to be either solid or empty, the confidence interval will remain visible even if the symbol size exceeds the upper and lower bounds of the confidence interval.

Adopting these modifications will provide viewers with a graph that is more easily and accurately decoded. If decoding the information contained in the graph is made easier and it results in a more accurate assessment of the relative sizes of the symbols, it follows that a more accurate interpretation of the relative contributions of the studies will be conveyed.

Unfortunately SAS® software does not offer a procedure to generate high-resolution forest plots. The TIMEPLOT and SHEWHART procedures can produce figures that communicate some of the information resulting from a meta-analysis. However each of these procedures has limitations. PROC TIMEPLOT produces dot plots, which are a convenient way to obtain a quick, low-resolution plot of the results. But the PROC TIMEPLOT cannot control the size of the symbols, an important element of any forest plot, nor can it generate publication quality plots. By comparison the PROC SHEWHART can produce high-resolution side-by-side box plots. While box plots show effect size and confidence interval information, they plot other information such as quartiles and outliers that are not usually included in forest plots. Finally, box plots present the data in a vertical rather than a horizontal format, which is atypical for the presentation of meta-analytic results.

Mitchell (2000) provided an excellent SAS® macro to generate traditional forest plots using the GPLOT procedure, but there are several differences between his macro and ours. The most important difference is that Mitchell's macro does not provide a way to generate rectangular symbols whose widths, rather than areas, are proportional to study weight.

The macro used to generate figures in this paper enables SAS® users to quickly and easily produce high-resolution forest plots that incorporate all recommended modifications. For completeness, it provides an option to generate traditional forest plots where symbol areas represent study weights. It does not matter whether symbols are width or area based, or if the x-axis is arithmetic or logarithmic, any two symbols representing identical study weights will have the same dimensions.

USING THE PROGRAM

The program is divided into three components: a data step that captures the results of the meta-analysis to be displayed, the initialization of a set of macro variables that pass settings to the macro that control various aspects of the forest plot, and the macro that produces the plot. Each of these components is commented on below.

READING THE DATA: THE DATA STEP

The first and last few lines of the data step used to enter the results of the meta-analysis displayed in Figure 1 are reproduced below:

```
DATA meta1;
  LENGTH study $30;
  INPUT study 1-35 weight effect l95 u95 symbol bold indent;
  y=_n_;
  CARDS;
    Random                          1500 2.94   2.47   3.50   4 0 1
```

```
      Fixed                          1500 2.86   2.55   3.20   3 0 1
      Overall Pooled Estimate:       .     .      .      .     0 1 1
                                     .     .      .      .     0 0 1
      Random                         1500 4.44   1.87  10.50   4 0 1
      Fixed                          1500 3.90   2.65   5.73   3 0 1
      Sub Pooled Estimate            .     .      .      .     0 1 1
      McMahon (85)                   3000 7.56   3.08  18.50   2 0 1
      James (84)                     750  3.11   2.02   4.80   2 0 1
      Sibutramine at 24 M            .     .      .      .     0 1 0
                                     .     .      .      .     0 0 1
  .
  .
  .
      Kelley (82)                    1000 3.26   2.12   5.02   2 0 1
      Hollander (75)                 1500 3.25   2.01   5.27   2 0 1
      Hauptman (81)                  3000 1.90   1.18   3.07   2 0 1
      Hanefeld (71)                  3000 2.29   1.11   4.71   2 0 1
      Finer (79)                     2000 2.03   1.11   3.71   2 0 1
      Davidson (77)                  4000 2.49   1.83   3.40   2 0 1
      Bakris (70)                    3500 2.88   1.98   4.18   2 0 1
      Orlistat at 12 M               .     .      .      .     0 1 0
                                     .     .      .      .     0 0 1
  ;
  RUN;
```

This data step provides SAS® with the results of the meta-analysis that are required to generate the forest plot, along with associated formatting details. In all, eight variables are read into the data set; STUDY, WEIGHT, EFFECT, L95, U95, SYMBOL, BOLD and INDENT. One variable, Y, is a created variable which controls the vertical placement of the information contained within each observation of the data step. The first observation of the data set corresponds to the lowest entry on the plot and the last to the highest.

The first five variables are associated with the results of the meta-analysis. STUDY is a text variable that captures the label for each study or summary statistic. This information will be displayed for each row on the forest plot. If STUDY is blank it will have the effect of placing a blank space on the forest plot. This is handy as there might be subsections that you wish to keep separate. The numeric value of WEIGHT is the relative contribution of the study identified in STUDY. It is one of two numbers that determine the width of the symbol. EFFECT is the numeric value of the effect size for the item specified in STUDY. Finally, L95 and U95 are the lower and upper 95% confidence limits of the effect size respectively.

The last three variables control the type of symbol and the placement of information on the plot. SYMBOL is a numeric variable that identifies whether a symbol is to be plotted, and if so, whether it is a rectangle or a diamond, and whether it is filled or empty. It has valid values of 0, 1, 2, 3 and 4. When SYMBOL equals 0 (zero) no symbol is plotted. A setting of 0 would allow placing a heading on the graph without plotting a symbol. SYMBOL is set to 1 to request an empty rectangular symbol, and it is set to 2 to request a filled rectangle. SYMBOL is set to 3 or 4 to request a diamond shaped symbol. The diamond will be empty when SYMBOL equals 3 and filled when it equals 4. In Figure 1 the symbol corresponding to the random effect summary is a solid diamond (SYMBOL = 4) and the symbol for the fixed effect summary is an empty diamond (SYMBOL = 3).

BOLD is a numeric indicator variable that can be set to 0 or 1. When BOLD equals 0 the information in STUDY will be displayed on the graph in a regular font. If BOLD equals 1, STUDY will be displayed in bold font.

INDENT is the last numeric indicator variable and it too can be set to 0 or 1. Indent allows headings to be placed in a slightly offset location so that headings and study results can be easily distinguished. When INDENT equals 1 the information displayed in the leftmost column of the plot will begin slightly more to the right than when INDENT equals 0.

THE MACRO VARIABLES

The code that initializes the macro variables for Figure 1 is listed below. Many of these variables must be initialized for the macro to produce the forest plot.

```
TITLE3;
%LET dsn = meta1;
%LET title1 = "Figure 1a:  At least 5% Weight Loss at 12 or 24 Months";
%LET title2 = "(Orlistat or Sibutramine vs Placebo)";

%LET head1 = "STUDY";
%LET head2 = "  OR [95% CI]";

%LET foot1 = "Note:  SUGI 31";

%LET textxl = "Favors Placebo";
%LET textxr = "Favors Treatment";

%LET x1  = 50;
%LET x2  = 65;
%LET vly = 0;
%LET vlx = 1;

%LET c1  = 5;
%LET c2  = 25;
%LET ft  = 5;

%LET wttype = "N";
%LET stype  = "B";

%LET sh  = 0.25;
%LET dh  = 0.45;
%LET ssf = 5;
%LET dsf = 5;

%LET base = 10;
%LET xorder = 0.1 1 10 100;
```

The macro variable name is assigned by the %let statement and it identifies the macro variable that will be passed to the macro. The information to the right of the equal sign is the value of the macro variable. These macro variables identify titles, headings, footnotes, x-axis labels, placement of x-axis labels, placement of a vertical reference line, placement of column headings and footnote, settings for the x-axis scaling, the symbol size factor, and settings for the height of the rectangles and diamonds on the plot. A brief description of the purpose of each of the macro variables follows.

The first macro variable, DSN, holds the name of the data set which contains the results of the meta-analysis to be plotted. The next two macro variables identify two titles. TITLE1 and TITLE2 are the titles that will be shown at the top of the plot. More titles could be added to this list if desired.

At the left of Figure 1 there are two columns of information. The first identifies the STUDY information and the second shows the odds ratio along with its 95% confidence interval for each item plotted. HEAD1 and HEAD2 are the two macro variables that identify what appears as the headings for these columns. A footnote may be placed on the graph underneath the two columns. The macro variable FOOT1 identifies the text to appear in the footnote. C1, C2 and FT are macro variables that control the placement of the two column headers and the footnote respectively. The values assigned to these macro variables correspond to the percentage of procedure output area along the x-axis where the header or footer begins. A larger value will place the item more to the right on the graph.

Two macro variables identify the x-axis labels: TEXTXL and TEXTXR. The placement of these labels is controlled by macro variables X1 and X2 respectively. X1 identifies the percentage of procedure output area along the horizontal domain where TEXTXL ends. X2 identifies the percentage of procedure output area where TEXTXR begins. These four macro variables provide control over the content and placement of two x-axis labels.

Macro variable VLY identifies the location on the y-axis where the bottom of the vertical reference line is placed. In Figure 1 it is set to 0. The reference line is drawn to the location on the y-axis of the uppermost citation. VLX

identifies the location on the x-axis where the vertical reference line is located. Since the objective is to plot information about odds ratios in Example 1, the vertical reference line is placed at 1 on the x-axis as this is the point where treatment and control are equally effective. If this line intersects the confidence interval of a study it would imply that a statistically significant effect was not found in that study.

The next variable WTTYPE identifies whether WEIGHT is the weight derived from the meta-analysis or if it is the study sample size. If the weight is known then WTTYPE should be set to W, if the weight is not known but the sample size is known then WTTYPE is set to N. When WTTYPE is set to W the symbol sizes are proportional to W, when it equals N the symbol sizes are proportional to the square root of N. STYPE allows the user to choose between two symbol types. It is set to R for rectangles, where the width of the symbol represents study weight, or B for boxes, where the area represents study weight.

The last six macro variables in the list are the most important because they control the type and size of the symbols on the plot and identify the scale of the x-axis. SH and DH, control the heights of the rectangle and diamond symbols respectively. Increasing the values of these variables increases the heights of the symbols. SH is only relevant when STYPE is set to R. SSF and DSF are scaling factors that manipulate the size of rectangles and diamonds respectively. A smaller value of SSF will result in wider rectangles while a larger value will result in narrower rectangles, but the relative differences in widths among symbols will remain constant. The widths of the diamond symbols are controlled independently in a similar manner with DSF.

As a starting point, the following guidelines will produce a maximum symbol width that is approximately equal to ten percent of the width of the graph:

WTTYPE	STYPE	Set SSF and DSF to
W	R	(maximum value of WEIGHT/4)
W	B	square root (maximum value of WEIGHT/4)
N	R	square root (maximum value of WEIGHT/4)
N	B	square root (square root (maximum value of WEIGHT/4))

The macro variable XORDER identifies the major tick marks along the x-axis. In Figure 1, major tick marks are placed at 0.1, 1.0, 10, and 100 because a logarithmic scale was requested. The lower and upper extreme values in this set of numbers must contain all possible values of L95 and U95 in your data set. If this is not the case, the graph will not be properly displayed.

Finally, BASE is used to identify the scale of the x-axis. When BASE equals 0 an arithmetic axis is generated. Any other positive integer for BASE will result in a logarithmically scaled x-axis, with base being equal to the value assigned to BASE.

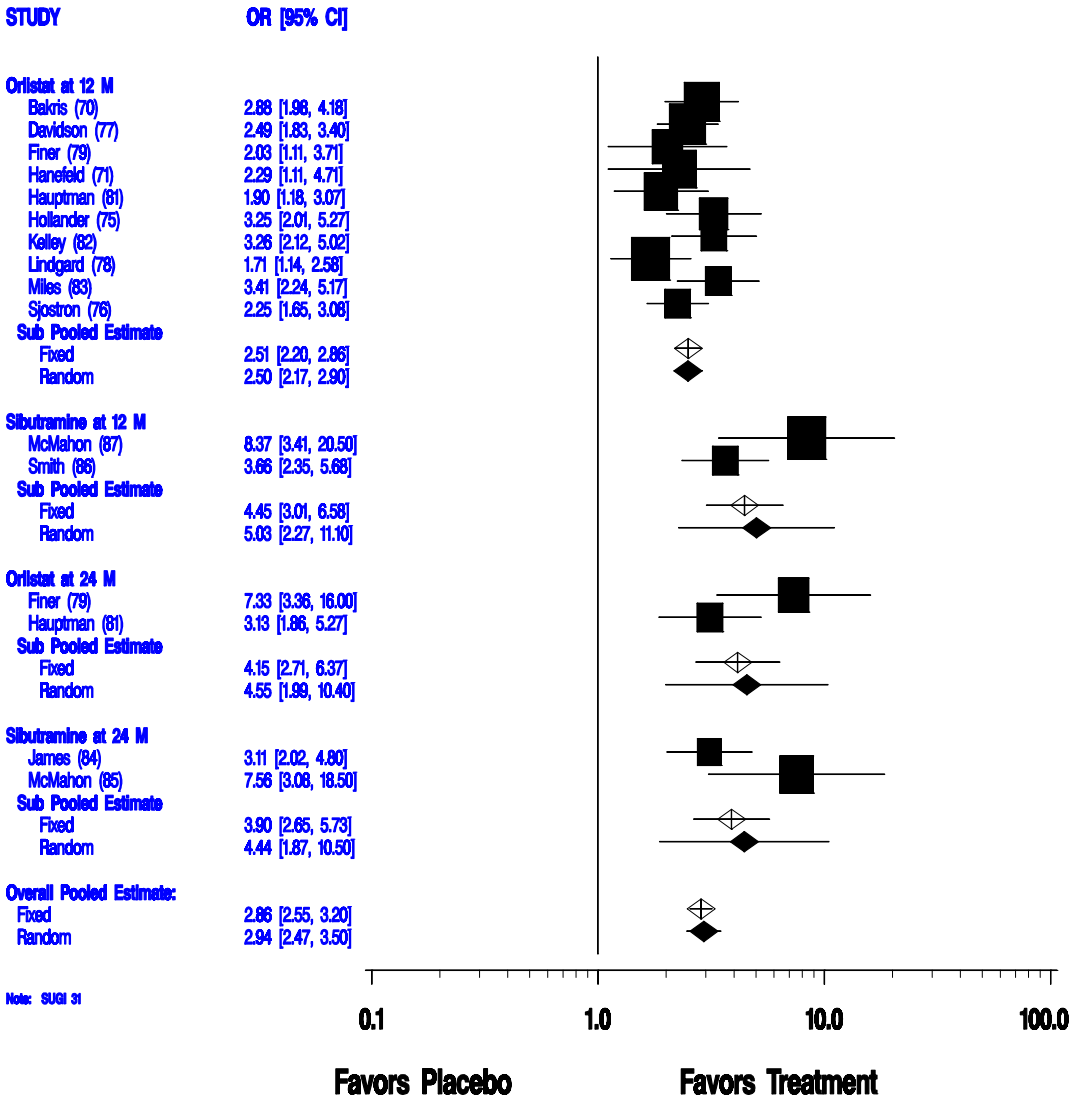
THE MACRO

After the data have been entered into a data set and all the macro variables have been set, the FOREST macro is invoked to generate the forest plot.

EXAMPLES

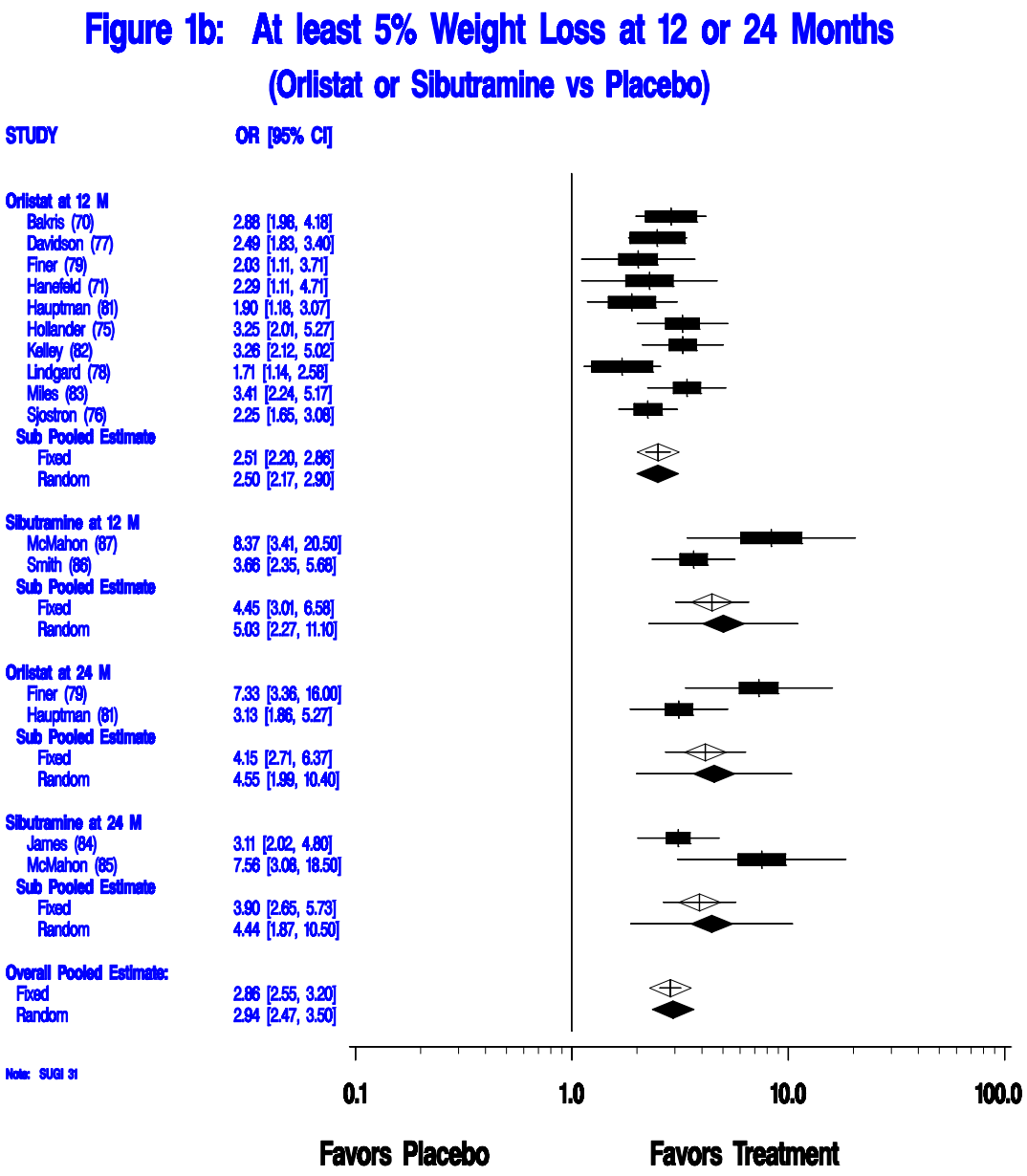
Figure 1a displays a traditional forest plot where symbol areas represent study weights. Figure 1b shows a modified forest plot where rectangle widths represent study weights. All macro variable settings for Figure 1a are shown above.

Figure 1a: At least 5% Weight Loss at 12 or 24 Months
(Orlistat or Sibutramine vs Placebo)



MACRO VARIABLE SETTINGS FOR FIGURE 1B THAT DIFFER FROM THOSE USED FOR FIGURE 1A

```
%LET title1 = "Figure 1b:  At least 5% Weight Loss at 12 or 24 Months";
%LET stype = "R";
%LET dh = 0.40;
%LET ssf = 25;
%LET dsf = 20;
```



COMMENTS ON EXAMPLES

When a meta-analysis includes many studies, as is the case in the example shown above, the resulting traditional forest plot may contain overlapping symbols (e.g., Figure 1a). The option to increase the space between lines may not be available when a large number of studies are included. It would be possible to make all symbols smaller so there would be no overlap, but a visual comparison among them would become increasingly difficult as the symbols decreased in size. By comparison, the rectangular symbols shown in Figures 1b clearly convey study weight information without having any overlapping symbols. With no overlapping symbols, and with symbols differing in one rather than two dimensions, it is much easier to visually compare different study weights. A clear advantage of using rectangular symbols is that their widths can be adjusted to facilitate the comparison among studies without any risk of overlap.

CONCLUSION

Several potential problems associated with traditional forest plots have been identified in this paper. If symbol areas are used to convey study weights, study symbols can overlap or obscure confidence interval information. If one study has a far larger weighting than others it might be difficult to plot the symbols to accurately reflect those differences. The use of a logarithmic scale raises plotting difficulties due to the non-interval nature of the scale. All of these problems have been addressed with the relatively easy to use macro.

REFERENCES

Cleveland WS, The Elements of Graphing Data (Revised Edition), p. 228. Summit NJ: Hobart Press, 1994.

Lewis S, Clarke M. Forest plots: trying to see the wood and the trees. BMJ 2001; 322; 1479-1480.

Mitchell RM, Forcing SAS/GRAPH® software to meet my statistical needs: A graphical presentation of odds ratios. Proceedings of the 25th Annual SAS® Users Group International (SUGI) Conference, pp. 882-887.

ACKNOWLEDGMENTS

Thanks to Susan Tomlinson for her careful review of several drafts of this paper.

CONTACT INFORMATION

Gary Foster
St. Joseph's Healthcare
105 Main Street East, Level P1
Hamilton, ON L8N 1G6

E-mail: fosterg@mcmaster.ca