

Symbol Type and Colour in Graphs

Diederik Pietersma, AstraZeneca, Alderley Park, Macclesfield, United Kingdom

ABSTRACT

Selection of appropriate colours and types for the symbols and lines involved in graphs can be a time consuming process. Factors that influence the choice of symbol type and colour include the number of categories to identify, the number of data points to plot, the amount of overlap among data points and the type of graph involved, such as scatter plot or line plot. A satisfactory choice of symbol type for scatter plots, group average line plots and individual patient line plots might involve texture symbols, filled geometric shapes and upper-case letters, respectively. Limiting the choice of colour to a small set of six colours that are distinct and sufficiently dark for black-and-white printing seems to work best for both symbols and lines. With three treatment groups, the suggested symbol types are “o”, “+” and “<” for scatter plots and “filled circle”, “filled triangle” and “filled square” for group average line plots. If more than five treatment groups are involved it might be better to create a multi-panel display with each treatment group shown in a separate plot on the same page. For individual patient line plots, combining six colours with four distinct line types helps to distinguish the lines of up to 24 patients. Examples of code for symbol type and colour are provided in SAS® and S-PLUS®.

INTRODUCTION

Graphs can be a very useful tool to support the interpretation of clinical safety, pharmacokinetic, and efficacy data. For example, Amit et al (2008) presented several approaches to using graphs for the analysis of safety data from clinical trials. Graphs often involve multiple treatment groups or multiple patients, which need to be identified with appropriate symbols. However, a poor choice of symbol type and colour can hinder rather than facilitate interpretation and may even result in misleading displays (Bessler, 2005; Oates, 2006).

Selection of appropriate colours and types to use for symbols and lines in the various graphs required for the interpretation of clinical study data can be a time consuming process, often involving too many iterations. Factors that influence the choice of type and colour of symbol include the number of categories to identify and the type of graph involved. For plots with only two categories, for example showing both investigational product and placebo, it is relatively easy to select distinct symbol types and colours. However, with many categories, such as dose levels, it can be difficult to find a set of symbol types and colours that result in an effective display. The type of graph also influences proper choice of symbol type and colour. For example, with a scatter plot showing two numeric variables for multiple treatment groups, the symbols should be very distinct to allow viewers to focus on the pattern associated with each treatment group at a time (Cleveland, 1994). However, with a line plot showing laboratory test results in time for all patients in a small study, the lines connecting the data points for each patient help viewers to follow each patient in time, making the choice of symbol type and colour less critical.

The objective of this paper is to provide recommendations for the type and colour of symbols and lines, depending on the type of graph and the number of treatment groups or patients involved. Sample code for scatter plots and line plots will be provided in both SAS and S-PLUS.

SYMBOL TYPE

The choice of symbol types to use with a graph depends to a large extent on the number of data points to plot, the amount of overlapping of data points and the number of categories that need to be identified. In graphs with many partially overlapping data points, as often occurs with scatter plots, using symbols that involve complicated shapes or filled geometric shapes tends to make it difficult to identify the individual observations. For such situations, Cleveland (1994) suggested the use of the following set of symbols, which were chosen based on research in the visual perception of textures:

o + < s w

With two categories to identify, using the first two of these texture symbols is recommended, with three categories the first three, etc. The distinct shapes of the texture symbols makes them very resistant to partial overlap problems in scatter plots. In addition, their unique shapes makes it easier for viewers to focus on the pattern represented by one category at a time and to perceive differences among the patterns of the categories shown in the graph.

PhUSE 2008

With graphs involving not too much data overlap, using circles with different methods of fill can be a very effective coding method (Cleveland, 1994). This approach seems especially useful to identify categories that are ordinal, such as dose level, when the increasing amount of fill can be associated with the increasing dose level. For example, in a study with three dose levels of the investigational drug, the dose levels could be represented with a circle without fill, a circle containing a dot, and a completely filled circle:

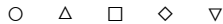


Similarly, treatment and placebo might be represented with a filled and a non-filled circle. However, filled symbols are more prominent than open symbols, which means that the groups they represent will attract more attention. Especially in scatter plots, this can bias the interpretation (Oates, 2006).

With line plots, the lines connecting the symbols help to focus the viewer on all the data points associated with one category at a time. In this situation the symbols do not need to be as distinct as with scatter plots. Using symbols that involve more ink than the texture symbols, such as filled geometric shapes and increasing the line width may actually help to emphasize the mean patterns over secondary chart objects such as error bars and reference lines. A set of five fairly distinct filled geometric shapes consists of a filled circle, triangle, square, diamond, and inverted triangle:



In studies where both scatter plots and line plots with the same categories are produced, for example for treatment groups, the same or a similar set of symbols should be used for all graphs to ensure consistent interpretation. Using the same set of geometric shapes to identify treatment groups, but with the filled variants for group average line plots and non-filled variants, shown below, for scatter plots might work satisfactory in such situations.



Symbols should be large enough to support easy identification of individual data points, but not too large, to limit the amount of partial overlapping in dense data displays. Within a selection of symbols for a graph, the size of some symbols may need to be adjusted to give all symbols similar prominence.

SYMBOL COLOUR

Colour is a very efficient method to visually group different categories in a graph because our visual system is very good at assembling different objects with the same colour (Cleveland, 1994). The use of colour tends to work even better than symbol type in supporting viewers to focus on one category at a time while mentally filtering out the other categories. The use of colour is therefore highly recommended with scatter plots.

A complicating factor is that graphs tend to be produced as a one-size-fits-all solution: the same graphical outputs will be displayed on colour monitors, printed in colour and in black-and-white, and reproduced on black-and-white photo copiers. The black-and-white usage requirement has several consequences. Firstly, yellow and other faint appearing colours should be avoided for use with symbols and lines. Secondly, symbol type should be used in addition to colour to identify the different categories in a graph. This will provide the best possible results when viewed in colour, while still leading to adequate encoding after printing or copying in black-and-white. Finally, in line plots, the width of coloured lines may need to be increased to ensure legibility after printing or copying in black-and-white.

Another issue is colour blindness and the combination of certain colours in the same graph. Using both red and green is discouraged, as the most common form of colour blindness cannot distinguish between red and green.

To specify a colour, different colour coding systems are available (Cleveland, 1994; Bessler, 2005). The HSL system describes hue (colour measured in degrees from 0 to 360 with primary colours red, yellow, green, cyan, blue and magenta), saturation (how pale or deep a colour appears) and lightness (how light or dark a colour appears). The CMYK system is used for printers and specifies colour with percentages of cyan, magenta, yellow and black. For computer monitors the RGB system was developed, specifying red, green and blue with values from 0 to 255.

To identify different categories in scatter plots, Cleveland (1994) recommends using different hues that are easy to distinguish: cyan, magenta, green, and orange. Several web-based colour recommendation tools are also available. For example, the ColorBrewer.org website suggests colours to use with cartographic maps (Brewer, 2008). While recommendations for maps are not directly applicable to symbols, such tools can provide some guidance regarding what colours to choose depending on the number of categories to identify. For example, a list of dark colours to identify seven qualitative classes involves Red, Blue, Green, Purple, Orange, Brown and Grey with their specific description in various colour-coding systems such as CMYK and RGB (Brewer, 2008).

Colour can also be useful to encode numeric or ordinal variables (Cleveland, 1994). For example, with treatment groups representing dose levels, keeping the hue (colour) constant, while decreasing the lightness and increasing the saturation provides a strong sense of order in the identification of the increasing dose levels. However, although varying lightness and saturation may work well when substantial areas of the display are coloured such as bar

charts, the approach seems much less effective with scatter plots involving non-filled symbols that use little ink. In addition, the level of lightness is limited due to proper printing in black-and-white, which reduces the number of levels that can be distinguished within a particular hue.

MULTI-PANEL DISPLAYS

With two or three categories within a single scatter plot effective detection and assembly of categories, such as treatment groups, is relatively easy. However, with four or more categories this becomes increasingly difficult. In situations with many categories involved it might be more useful to create multi-panel displays in which each category is shown in a separate plot on the same page. For example, when dealing with a scatter plot and six dose levels, showing a series of six small plots arranged in two rows and three columns tends to be more effective than superimposing the data from the six groups in a single scatter plot. To facilitate comparison of the patterns shown on the different panels, use of a visual reference grid is recommended (Cleveland, 1994). An example of such a reference grid with laboratory value plots include reference lines for the upper and lower limits of normal and a diagonal line with shift plots comparing baseline to last observation on treatment.

Multi-panel displays can be generated in SAS using PROC GREPLAY (Stepien, 2006). In S-PLUS, multi-panel displays can be produced in a very straightforward manner using Trellis graphics.

EXAMPLE CODE – SCATTER PLOTS

Scatter plots often involve many partial or full overlapping data points. In such situations the five texture symbols recommended by Cleveland (1994) are a good choice for symbol type. Five distinct colours are blue, green, magenta, orange and dark grey. If only two categories are involved then a non-filled circle and a non-filled triangle could also be used. When the variable to identify is ordinal, e.g. dose level, and more than two levels are involved, using a multi-panel plot with the increasing levels side-by-side seems preferable over using symbols with increasing amounts of ink in a single scatter plot.

SAS

In SAS, RGB colours can be specified using CXrrggbb, with rr, gg and bb representing the levels of red, green and blue, respectively, using hexadecimal code with range 00-FF. Custom colours can thus be setup as follows:

```
%let myblue   = CX0000CC; * Blue;
%let mygreen  = CX00CC00; * Green;
%let mymagen  = CXCC00CC; * Magenta;
%let myorang  = CXCC6600; * Orange;
%let mydgrey  = CX666666; * Dark grey;
%let myred    = CXCC0000; * Red;
```

With the symbol statement, the color option specifies colour, height can be used to adjust symbol size and the interpolation option (i) is set to "none" for unconnected data points (scatter plot). Height (h) can be used to adjust symbol size. By specifying a text font with the font option, the text string specified with the value option will be used as the symbol type.

```
symbol1 color=&myblue   i=none height=1   font=simplex value=o;
symbol2 color=&mygreen  i=none height=1.1 font=simplex value=+;
symbol3 color=&mymagen   i=none height=1   font=simplex value=<;
symbol4 color=&myorang   i=none height=1   font=simplex value=s;
symbol5 color=&mydgrey   i=none height=1   font=simplex value=w;
```

S-PLUS

In S-PLUS, the symbol colour, type and size can be set with the graphical parameters col (colour), pch (plotting character) and cex (character expansion). Colours can be chosen from a numbered colour scheme that is associated with the graphics device used. For example, the colours black, green and red might be identified with 1, 4, and 8 respectively. These colour schemes can be customized with specific RGB values. The symbol type is assigned with numbers that refer to a list of symbols, similar to colour, or by specifying characters within quotation marks. To adjust the size of individual symbols, the cex parameter can be used. The code below specifies symbol colour, type and size for five different categories.

```
mycol = c(6, 4, 3, 5, 16) # Blue, Green, Purple, Orange, Grey
mypch = c("o", "+", "<", "s", "w") # Texture symbols
mycex = c(1, 1.1, 1, 1, 1) # Symbol size
```

A scatter plot example generated with these settings is shown in Figure 1.

PhUSE 2008

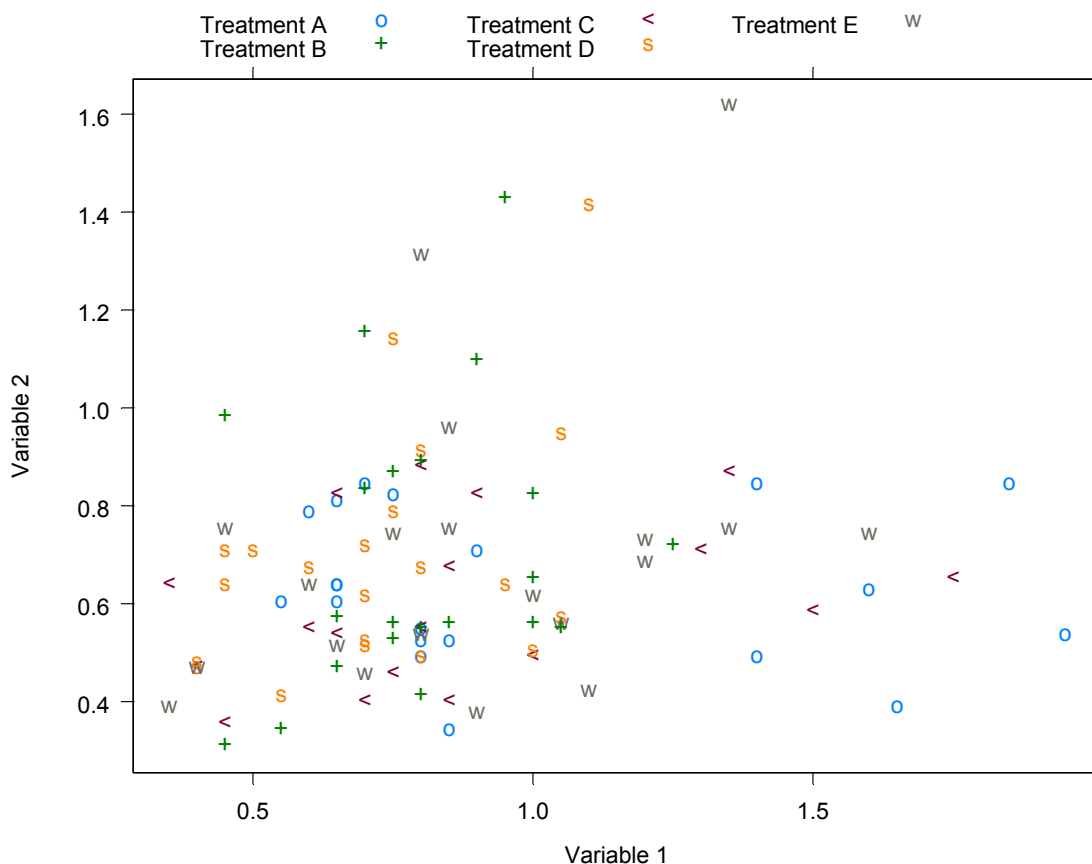


Figure 1 Scatter plot with five treatment groups identified with texture symbols.

EXAMPLE CODE – GROUP AVERAGE LINE PLOTS

Line plots are often used to show treatment group means for time dependent data and to show profiles of time dependent data for individual patients. For group average line plots, filled geometric shapes seems to work quite well as the increased use of ink helps to emphasize the mean patterns shown.

Varying line type may not be necessary for group average line plots, given the limited number of lines involved. Using a solid and a dashed line in the same plot would actually put more emphasis on the treatment represented with the solid line at the expense of the treatment represented with the dashed line.

SAS

With the symbol statement, the interpolation option (i) is set to "join" to connect the data points, the line option sets the type of line and width (w) can be used to adjust line width. Filled geometric shapes as symbol types can be achieved by setting the font option to 'marker' and setting the value option to the letter code associated with the desired geometric shape.

```
symbol1 c=&myblue i=join line=1 w=2 h=1 font=marker, value=Z; * Dot;
symbol2 c=&mygreen i=join line=1 w=2 h=1 font=marker, value=C; * Triangle;
symbol3 c=&mymagen i=join line=1 w=2 h=1 font=marker, value=U; * Square;
symbol4 c=&myorang i=join line=1 w=2 h=1 font=marker, value=P; * Diamond;
symbol5 c=&mydgrey i=join line=1 w=2 h=1 font=marker, value=D; * Inverted triangle;
```

Figure 2 shows an example of a group average line plot generated with these settings.

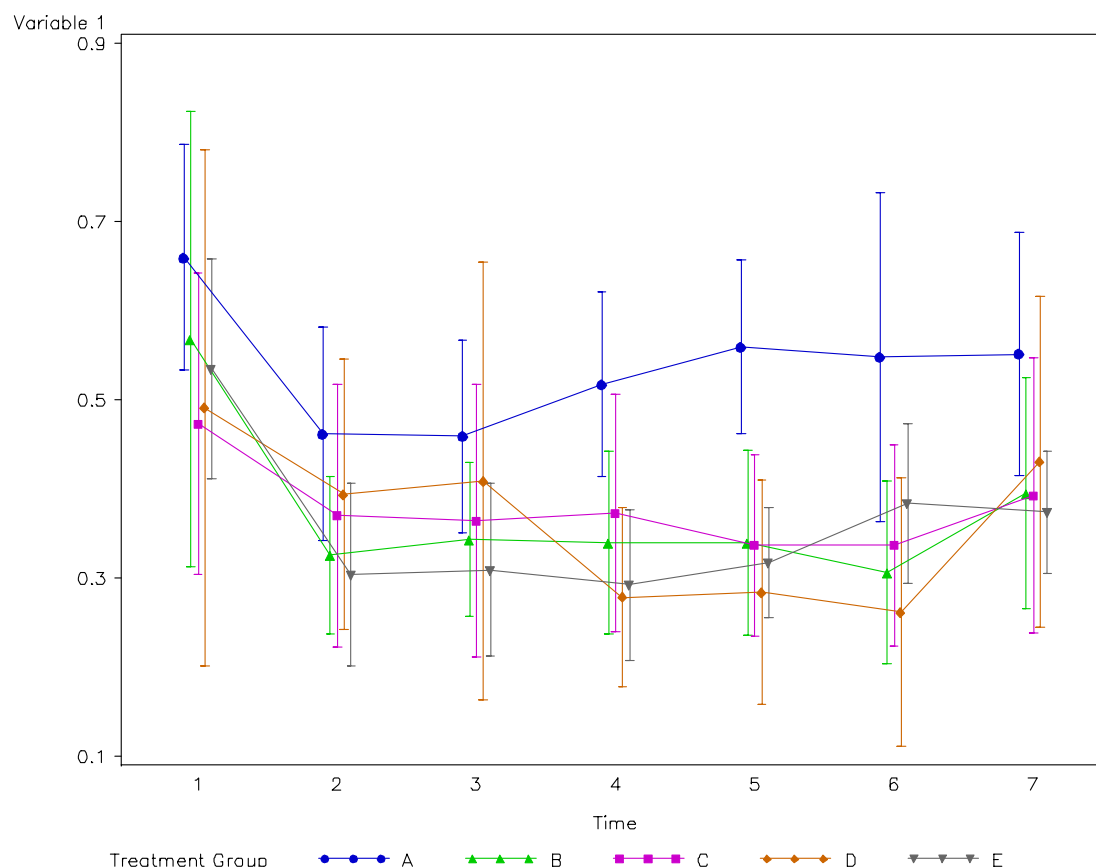


Figure 2 Line plot showing means for five treatment groups.

S-PLUS

In S-PLUS, the line type and width are set with graphical parameters `lty` and `lwd` using a vector of numbers as shown in the example code below.

```
mycol = c(6, 4, 3, 5, 16) # Blue, Green, Purple, Orange, Grey
mypch = c(1, 2, 28, 5, 6) # Dot, Triangle, Square, Diamond, Inverted triangle
mycex = c(1)                # Symbol size
mylty = c(1)                 # Solid line
mylwd = c(1.5)               # Line width
```

EXAMPLE CODE – INDIVIDUAL PATIENT LINE PLOTS

With line plots showing individual patient data, such as laboratory parameter results in time, often a large number of categories need to be identified. To reduce problems with viewing overlapping data points, filled symbols, such as a dot, are best avoided. Finding a set of unique symbols that use little ink can be challenging. A solution is to use the letters of the alphabet to identify the individual patients. In addition, letters are very easy to remember for a reviewer. In small studies, consistent mapping between letter and patient identifier allows reviewers to follow individual patients across different patient line plots.

With more than a few lines in a plot, many of which are crossing each other, using line types and colours that are easy to identify can help viewers to focus on one individual patient at a time. A limited set of distinct line colours can be combined with different line types to create uniquely identifiable lines. Dotted lines are not recommended as they are much less pronounced than dashed or solid lines and may become difficult to perceive with black-and-white printing. A list of four distinct example line types involves: solid, dashed, dash-dotted, and long dashed. Combining six distinct colours with four distinct line types helps to distinguish the lines of up to 24 patients identified with upper-case letters. With more than 24 patients, line plots tend to become too cluttered to allow for the use of symbols to identify each individual patient. In such cases, the use of solid black lines without symbols may work best to visualize patterns in time.

PhUSE 2008

SAS

With the symbol statement the interpolation option (i) is set to "join" to connect the data points and the line option sets the type of line (1=solid, 3=dashed, 8=dash-dotted, 4=long-dashed). Upper-case letter symbols can be achieved by specifying a text font with the font option and setting the desired letter with the value option.

```
symbol1 c=&myblue i=join line=1 w=1 h=1 font=simplex value=A; * Solid lines;
symbol2 c=&mygreen i=join line=1 w=1 h=1 font=simplex value=B;
symbol3 c=&mymagen i=join line=1 w=1 h=1 font=simplex value=C;
symbol4 c=&myorang i=join line=1 w=1 h=1 font=simplex value=D;
symbol5 c=&mydgrey i=join line=1 w=1 h=1 font=simplex value=E;
symbol6 c=&myred i=join line=1 w=1 h=1 font=simplex value=F;
symbol7 c=&myblue i=join line=3 w=1 h=1 font=simplex value=G; * Dashed lines;
etc...
```

S-PLUS

In S-PLUS, the line type and width are set with graphical parameters lty and lwd using a vector of numbers as shown in the example code below.

```
mycol = c(6, 4, 3, 5, 16, 8) # Blue, Green, Purple, Orange, Grey, Red
mypch = LETTERS[1:24] # Upper-case letters of the alphabet
mycex = c(0.7) # Symbol size
mylty = rep(x=c(1, 8, 3, 4), each=6) # Solid, Dashed, Dash-dotted, Long-dashed
mylwd = c(1.5) # Line width
```

An example plot based on these settings is shown in Figure 3.

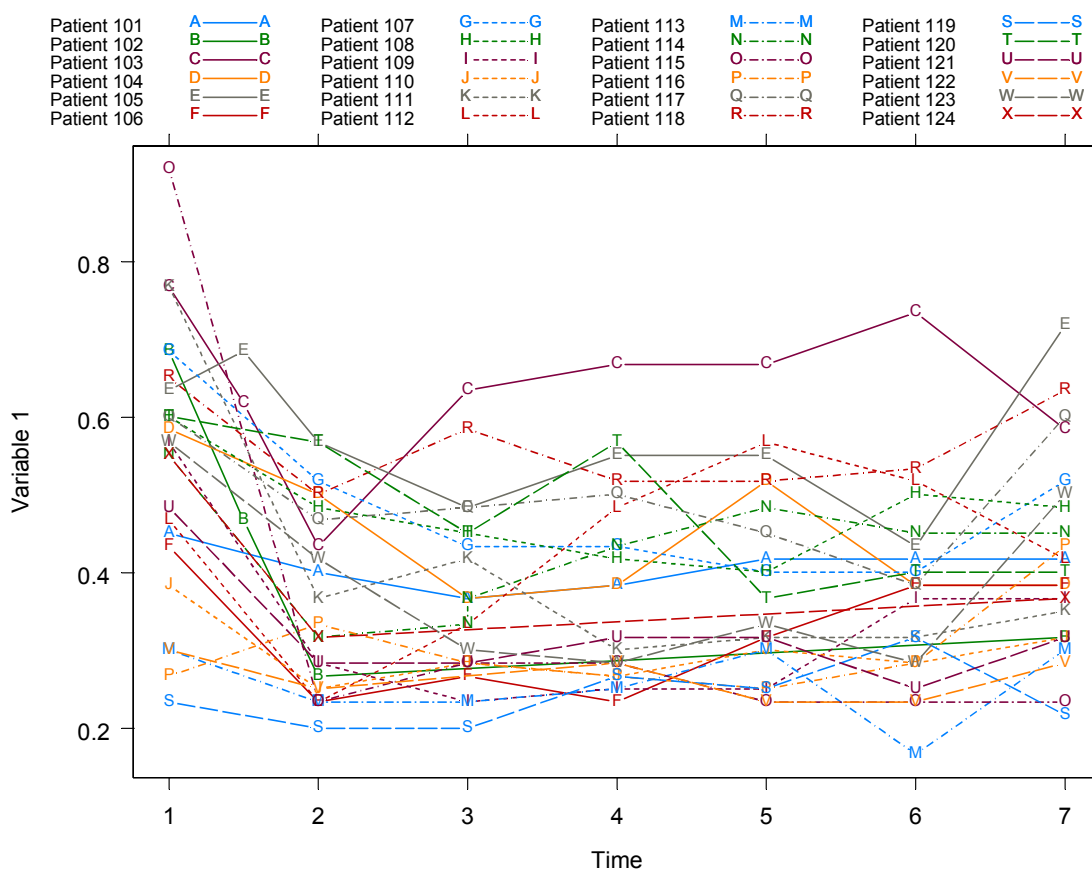


Figure 3 Line plot with 24 patient profiles identified with upper-case letters.

CONCLUSION

Recommendations for the use of symbol type and colour depend on the type of graph involved and the number of categories to identify. A satisfactory choice of symbol type for scatter plots, group average line plots and individual patient line plots might involve texture symbols, filled geometric shapes and upper-case letters, respectively. Limiting the choice of colour to a small set that is distinct and sufficiently dark for black-and-white printing seems to work best for both symbols and lines. Using guidelines for symbol type and colour is expected to reduce the number of iterations required to generate graphs that facilitate the interpretation of clinical trial data.

REFERENCES

- Amit O, Heiberger RM, Lane PW. Graphical Approaches to the Analysis of Safety Data from Clinical Trials. *Pharmaceutical Statistics* 2008; 7: 20-35.
- Bessler L. Effective Communication with Colour. Paper PS18 PhUSE 2005.
- Brewer CA. <http://www.ColorBrewer.org>, accessed 2008-05-09.
- Cleveland WS. *The Elements of Graphing Data*. Revised Edition. Hobart Press, Summit, NJ, 1994.
- Oates JA. How to Prevent Producing Misleading Graphical Displays. Paper P004 PhUSE 2006.
- Stepien S. Presenting Graphs on One Page. Paper CC07 PhUSE 2006.

ACKNOWLEDGMENTS

I would like to thank my AstraZeneca colleagues who provided useful feedback regarding this paper.

CONTACT INFORMATION

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

Diederik Pietersma
AstraZeneca
Mereside, Alderley Park,
Macclesfield SK10 4TF
United Kingdom
Email: diederik.pietersma@astrazeneca.com

Brand and product names are trademarks of their respective companies.