PhUSE 2016

Paper DV09

Concepts and Strategies for Developing Effective Data Visualizations

Becky Bates, GCE Solutions, Bloomington, IL, USA

ABSTRACT

Data visualization is a hot topic in the pharmaceutical industry today. There are many graphing packages which can create all manner of graphs. Regardless of the sophistication of these data visualization packages, they are still just tools. The results from a tool are only as good as the knowledge and skill of the professional using the tool. Learning how to use a given package is fairly straightforward. Obtaining the skill to develop effective data visualizations is less straight forward and remains a challenge for a majority of programmers and statisticians.

When designed effectively, data visualizations can provide greater insight into complex multifaceted data which potentially lead to more effective and informative decision making.

INTRODUCTION

Data visualizations is the new hot topic in the pharmaceutical industry today. Traditionally, visualizations have been static displays of data which included bar charts, line graphs, and pie charts. The new trend is to interact with the data dynamically that not only incorporates the traditional displays but also allows the users to drill down at various levels and to view subgroups of the data by just pointing and clicking a mouse.

Regardless of the sophistication and capability of data visualization packages today, they are still just tools to be used by programmers and statisticians. The results from a tool are only as good as the knowledge and skill of the professional using the tool. Knowledge how to use a given data visualization package is generally straightforward and easily mastered. However, the skill to be able to develop effective data visualizations is less straight forward and remains a challenge for a majority of programmers and statisticians.

The first step in designing effective visualizations is to understand the key drivers that apply to all visuals. The second step is taking into account various design considerations to optimize the key drivers. The third step is to remember the classic “Do’s” and “Don’ts” of graphing.

KEY DRIVERS

There are 3 key drivers in designing any effective data visualization display,

1. The story: what is the message or point wanting to be conveyed by the visual
2. Information Transfer Rate (Tufte, 1990): the amount of time it takes the viewer to understand the message being visually displayed
3. Venue: where is the visual display going to be present (e.g. poster, presentation, etc.)

It is very important for the designer to keep these drivers in mind throughout the design process for every visual display being created.

THE STORY

Before starting any data visualization display, the designer must know what message or point is going to be conveyed by the data visual. For example, Figures 1 and 2, below, contain the exact same data points. However, Figure 1 connects the data points over time which emphasizes trends over time. Figure 2 only shows error bars at each time point. If trending over time is the more important story, then Figure 1 is the more effective data visual. If comparisons between groups at each time point is the more important story, then Figure 2 is the more effective data visual.

Figure 1

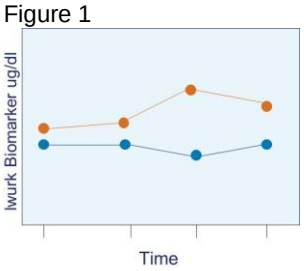
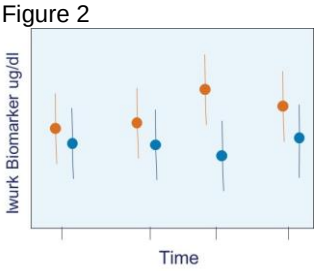


Figure 2



INFORMATION TRANSFER RATE (ITR)

The Information Transfer Rate (ITR) (Tufte, 1990) is the amount of time it takes the viewer to understand the message that is being visually displayed. Viewers typically are not as vested in spending effort into understanding a display as the designer. The harder a viewer has to work at understanding the visual display, the less convincing and less memorable the message becomes. Therefore, the designer must do all they can to minimize the ITR so that the message being displayed is readily understood by the viewer.

VENUE

The venue basically defines the bounds on what graphing elements (e.g. color, level of detail) are available to the designer. The following table lists venues and describes what the bounds are for each venue

Venue	Bounds
Presentations	Can use color but augmentation with numeric values should be minimized due to the limited amount of time spent on any slide. Additionally, these numeric augmentations are typically not readable and can detract from the display
Posters	Can use color and may want to increase numeric augmentation. However, viewers may still will only spend 10 mins or less on any given poster
Journals	Cannot use color but augmentation with numeric values and other details can be increased as the viewer may spend a considerable amount of time reviewing the results of a journal article
Regulatory Agencies	Cannot use color and may be stuck with more traditional graphs especially if they are expected from previous submissions
Internal meetings	Can use color and can be detailed if the viewers are very invested in the data being displayed

DESIGN CONSIDERATIONS

The design considerations incorporate the following:

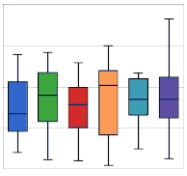
- Type of graph – pie chart, bar chart, heat map, etc.
- Graphing elements – line, color, fonts, movement, ordering, juxta positioning, details
- Visual comparisons – length vs area vs volume

It is through the combined use of all of these design considerations that can greatly improve or terribly impede the ITR.

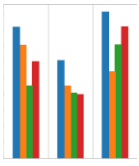
TYPE OF GRAPH

The same data can be displayed using different types of graphs. Each graph type enhances the different aspects of the data. For example, box plots are great for succinctly displaying the distribution of the data as well as the summary statistics such as mean and quartiles. Bar charts are great for visualizing comparisons (e.g. frequencies, means, counts) across groups. Heat maps are good for showing special densities. Line plots are good for showing trends over time. Therefore, the designer needs to understanding what aspects of the data (e.g. distribution, trends over time) best conveys the message and then choose the graph type that displays that aspect well.

Box Plot

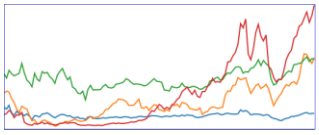


Bar Chart



Heat Map

Line Graph

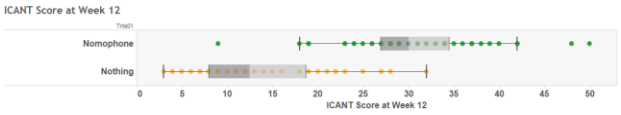


GRAPHING ELEMENTS

No graphing element is insignificant. All elements must be carefully chosen as they can either greatly increase the ITR or greatly impede it. For example, too much color can draw the eye toward the color and overpower the message being displayed (see Figure 3). Not enough color can make it hard to focus in on the message being displayed (Figure 4). This is true for all of the elements. Too many dark reference lines can obscure the data but no reference lines may make comparisons more difficult. Too large of font size can be distracting while too small of font size may make labeling hard to read.

Figure 3

Figure 4

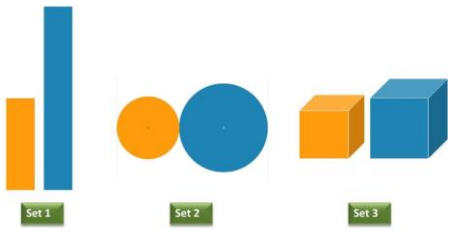


A simple test for the designer to determine if the graphing elements are done well or not is by closing their eyes and then noticing what are the first things their eyes notice upon opening them. If the data in the display is noticed first, that is a good sign. If any of the graphing elements are observed prior to the data, then the graphing element is overdone. If the eye is not drawn to anything, the graphing elements may be understated.

VISUAL COMPARISONS

Most data visualizations usually involve some kind of comparison either between treatment groups, subgroups, sites, time points, etc. The basic premise of visual comparisons is to either compare differing lengths, areas, or volumes. The human eye is very good at comparing differences when expressed as a length (e.g. lines or bars). That is to say, a person can readily determine if one line is a half or quarter the size of another line. However, areas are harder to detect such differences and volumes are even worse. This point is illustrated in Figure 5 where the blue bar, cirle, and cube are twice the size of the yellow bar, circle, and cube, respectively.

Figure 5



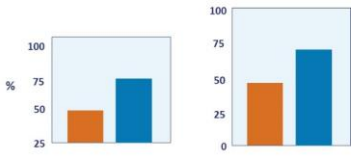
Therefore, when wanting to compare values between two or more groups, the designer must determine how fine a difference they want the viewer to discern and how important discerning the magnitude of the difference is to the message. If discerning the difference is key to the message, then line/bars may be more appropriate. If the differences are very large or discerning smaller differences are not key to the message, then areas may be an appropriate choice. If at all possible, volumes should be avoided.

DO'S AND DON'TS

The purpose for data visualizations is to represent numerical results and relationships in a visual manner. Therefore, it is imperative that visual displays maintain the mathematical relationship they are representing.

Mathematical relationship distortions occur when the scales of one or more of the axis are not appropriately constructed. A classic mistake is made when the scale for bar charts do not start at zero. This results in bars being artificially shortened which distorts the relative differences between bars (see Figure 6).

Figure 6



Don't Do

Another distortion occurs when the axis range differ across graphs (e. g. 0 – 100% vs 0-25%) and the information displayed within each graph is still being compared to other graphs (see Figure 7 below).

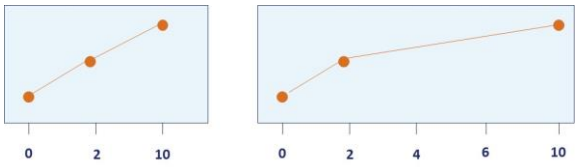
Figure 7

Don't

Do

One more very common mistake occurs when scalar values are displayed as class variables. For example, displaying the time point values 0,2 and 10 with equal spacing on the x-axis when the distance between 0 and 2 is not the same as the distance between 2 and 10. This results in the distortion of the slope between time points 3 and 6. Slope of the line is the visual display for rate of change (see Figure 8).

Figure 8



Don't

Do

To avoid the mathematical relationship distortions above, the designer must pay close attention the range of the scales and the consistency of the ranges across visual displays. Almost all graphing software packages provide a default scale which many times is not the most appropriate especially when multiple displays need to be taken into consideration.

One last final don't. Avoid chart junk at all costs Chart junk is considered to be anything that does not contribute to or distracts the viewer from understanding the message being conveyed by the visual display. Cutesy symbols (seeFigure 9), 3-D graphs with the z-axis not used for any measure (ss Figure 10), and overbearing graphing elements (see Figure 11) are just a few examples of chart junk.

Figure 9

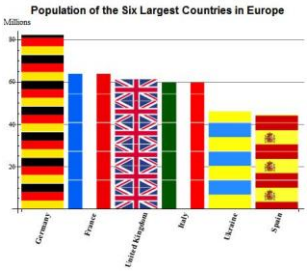


Figure 10

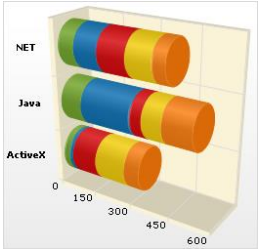


Figure11

CONCLUSION

With all of the various types of data wanting to be visualized within the pharmaceutical industry, today confounded by all the various software packages and graphing types within each package, it is understandable why programmers and statisticians may feel overwhelmed on how to design data visualizations let alone effective data visualizations. By focusing of the message needing to be conveyed and minimizing the IRT, this cuts a path through the myriad of options and distractions which can get in the way of designing effective data visualizations.

REFERENCES

Tufte, E. (1990). *Envisioning Information*. Cheshire, CT: Graphics Press LLC

CONTACT INFORMATION

Author Name: Becky Bates
Company: GCE Solutions

Address; 1408 E Empire St
City / Postcode: Bloomington, IL, 61701
Work Phone: +01 309-807-5879
Fax: 706-243-6415
Email: Rebecca.bates@gcesolutions.com
Web: gcesolutions.com