Clinical Chemistry 56:8
1229–1233 (2010)

Put Your Best Figure Forward: Line Graphs and Scattergrams

Thomas M. Annesley*

There is an old saying that "a picture is worth a thousand words." In truth, only a well-prepared, self-explanatory picture is worth a thousand words. The same holds true for research studies, for which 1 of the main methods we use to communicate our message is in figures and graphs. Figures and graphs tell much of the story by giving readers a visual anchor to help them see, understand, and remember information. Think about a report that you recently read and found useful. You likely do not remember the text used to state the results, or even the actual numbers, but you can recall much about the trends, relationships, outcomes, categories, or general experimental parameters shown in a graph. Despite the fact that you no longer have much recollection of the text, you can draw a reasonable representation of a graph from the published report and tell what you remember from it.

In this educational article I discuss line graphs and scattergrams and use examples to illustrate how to put your best *figure* forward so readers will remember you and your message.

Basics of a Good Graph

The components of a graph include axes, labels, scales, an origin, tick or reference marks, symbols, and a legend. Beyond these basics, however, a *good* graph has several attributes:

1. It draws attention to the data and not the graph itself.
2. The data points (symbols) and connecting lines are easy to read and distinguish.
3. Both the numbers and labels for the axes are readable and their meaning is clear.
4. The lengths of the 2 axes are visually balanced (ratio of x axis to y axis = 1.0 to 1.3).
5. The scales used on each axis match the range of the data.
6. Tick marks are used appropriately.
7. The legend is clear and concise.
8. The reader can understand the message without referring back and forth to the main text.
9. The data deserve to be graphed.

Line graphs and scattergrams make use of a horizontal and a vertical axis, typically called the x and y axis, respectively, to illustrate the relationship between 2 or more variables. By convention, the variable plotted on the x axis is referred to as the *independent variable*. The independent variable is the variable that is manipulated or changed by the investigator. The variable plotted on the y axis is the *dependent variable*. This variable is called the dependent variable because its value responds to (depends on) the value of the independent variable. It changes when the independent variable changes.

For example, one may study serum phenytoin concentration versus prescribed dose. The dose is the independent variable and the resulting serum concentration is the dependent variable because it depends on (or is caused by) a change in the independent variable. Think of it as asking a question: Does changing the dose (cause) result in a change in the circulating phenytoin concentration (effect)? This way of identifying a cause and effect relationship may often help you to determine whether the study involves independent and dependent variables and how you should design a figure to show the experimental results.

Another example is a study of serum prostate-specific antigen (PSA) as a noninvasive predictor of tumor staging. In this case the known (independent) variable is the tumor stage (a predefined variable or reference point), and the unknown (dependent) variable is the concentration of PSA in a patient's serum. It is possible to plot more than 1 dependent variable in a graph (e.g., total and free PSA), but there should be only 1 independent variable, in this example the tumor stage, plotted in a graph.

Although many studies have an independent variable, its presence is not a strict requirement. In some cases the study involves looking at an association of 2 variables, without any underlying proof of causation. A comparison of 2 analytical methods for quantifying troponin is a good example. In this case neither method has an effect on the other, and therefore no independent variable exists. The data for either method could

University of Michigan Health System, Ann Arbor, MI.
* Address correspondence to the author at: University of Michigan Health System, Room UH2G332, 1500 East Medical Center Drive, Ann Arbor, MI 48109-5054. E-mail annesley@umich.edu.

Received April 30, 2010, accepted May 6, 2010.
Previously published online at DOI: 10.1373/clinchem.2010.150060

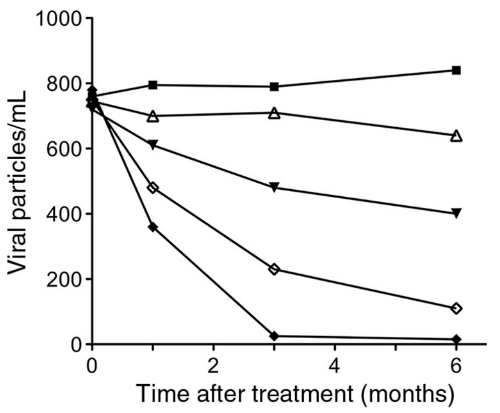

Fig. 1. Change in blood viral load during daily oral treatment with albenovir; ■, 0 mg/kg; △, 2 mg/kg; ▼, 5 mg/kg; ◇, 10 mg/kg; ♦, 20 mg/kg.

be plotted on the *y* axis. This having been said, however, it is still important in any study to determine if the data you are analyzing and graphing has an independent variable.

Fig. 1 shows an example of a line graph with the desired attributes mentioned earlier. This graph represents data from a hypothetical study of the efficacy of a new antiviral drug, albenovir. In this study albenovir was given orally daily to randomized groups of patients at 5 doses (0, 2, 5, 10, and 20 mg/kg). Blood samples were collected from patients at selected time points the beginning of treatment, and the samples were analyzed for circulating viral particles. The change in viral load vs time is plotted in the graph. In this figure the symbols representing the different doses are large and easily differentiated from one another, which allows them to be easily understood. The connecting lines are also clear and wide enough to draw attention to the data. A general rule is that the symbols and any lines or curves inside of the 2 axes are the most prominent features, the wording in the axes labels somewhat less prominent, and the axes and tick marks the least prominent. In this graph the 2 axis lines are proportional in length and narrow enough that they do not draw attention away from the data. The font size for the wording of the axis labels, which again highlights more important information, is larger than that of the numbers and tick marks on the axes. The tick marks are on the outside of the axes because they are associated with the numbers on the axes and not the plotted data points inside the axes. The scales are also proportional to the range of values and there is minimal wasted space throughout the graph.

The legend for this figure is concise, and the message can be understood even without having access to the main text. The graph in Fig. 1 does not include a title (often included in PowerPoint slides) because there is a legend that conveys the important information.

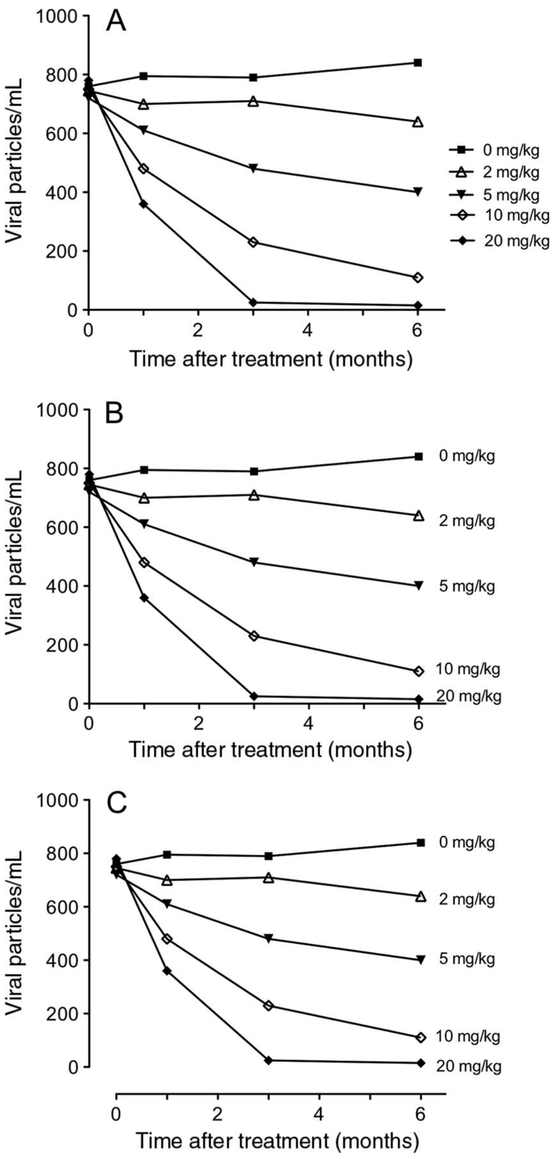

Fig. 2. Change in blood viral load during daily oral treatment with albenovir.

(A–C), different representations of the same data.

Fig. 2 shows several style options for improving the layout of the data plotted in Fig. 1. Although Fig. 1 is simple and clean, the reader must refer back and forth to the legend to associate the symbols and lines with the different treatment protocols. If there is extra space within the graph, or to the right, then one can consider adding a key to the symbols, as illustrated in Fig. 2A. If a key is added, however, it is important that the order of the symbols (top to bottom or left to right) in the key be the same as the order in which the symbols and lines are plotted in the actual graph, as was done here. A benefit of this approach is that it can simplify the message in the legend. If space allows, it is possible to consider an even a more effective design and place individual labels next to each line or set of data (Fig. 2B).

Sometimes data points have numerical values that fall directly on (or very close to) the *x* or *y* axis, as they do for these albenovir data. When this is the case, data points may be visually distorted or obscured by the axis line, especially when a graph is reduced to print size. In this situation (and only in this situation), one or both axes can be offset to allow a clearer visualization of the data (Fig. 2C). As an exercise, compare the data presentation and legends for Fig. 1 and 2 and evaluate how each influences what you see and read.

Common Mistakes

The next 3 examples highlight common mistakes authors make when preparing graphs. Fig. 3 shows the relationship between plasma and serum sodium for paired specimens from 150 patients. Sodium concentrations, even in critically ill patients, fall within a fairly narrow range from 125 to 165 mmol/L. Because many computer programs automatically default to *x*- and *y*-axis intercepts of 0, the graph may look like the one shown in Fig. 3A. There are 3 problems with this type of data presentation. First, the data points are compressed close together, making it difficult to see any scatter, or abnormal high and low values. Second, even with a correlation line, it can be difficult to see whether 1 or 2 outliers may have exerted undue influence on the overall correlation data. Third, because such a plot fails to properly convey the information in the data, it wastes space, which editors dislike for both economic and aesthetic reasons. These same data can be presented more clearly by narrowing the ranges of the axis scales to fit the true range of the data (Fig. 3B). An even better representation of the data can be obtained by creating a Bland–Altman plot (Fig. 3C), in which differences outside of the 95% limits of agreement are easily seen.

Similarly, editors often see results graphed as shown in Fig. 4. In this hypothetical example, an investigator developed a new HPLC assay for plasma alanine to support a collaborative study of rats undergoing high-stress experiments. To validate the stability of alanine in blood during courier transport across the university, specimens were collected into 4 different anticoagulant-containing tubes and stored at room temperature for selected time periods before centrifugation and freezing of the plasma. An obvious time-dependent loss of alanine in heparin specimens can be seen in Fig. 4A, even though the figure wastes a lot of usable space and compresses the data. Expanding the *y*-axis scale, as in Fig. 4B, not only makes much better use of space, it also shows a time-dependent increase in plasma alanine of almost 10% in oxalate tubes over an 18-hour period.

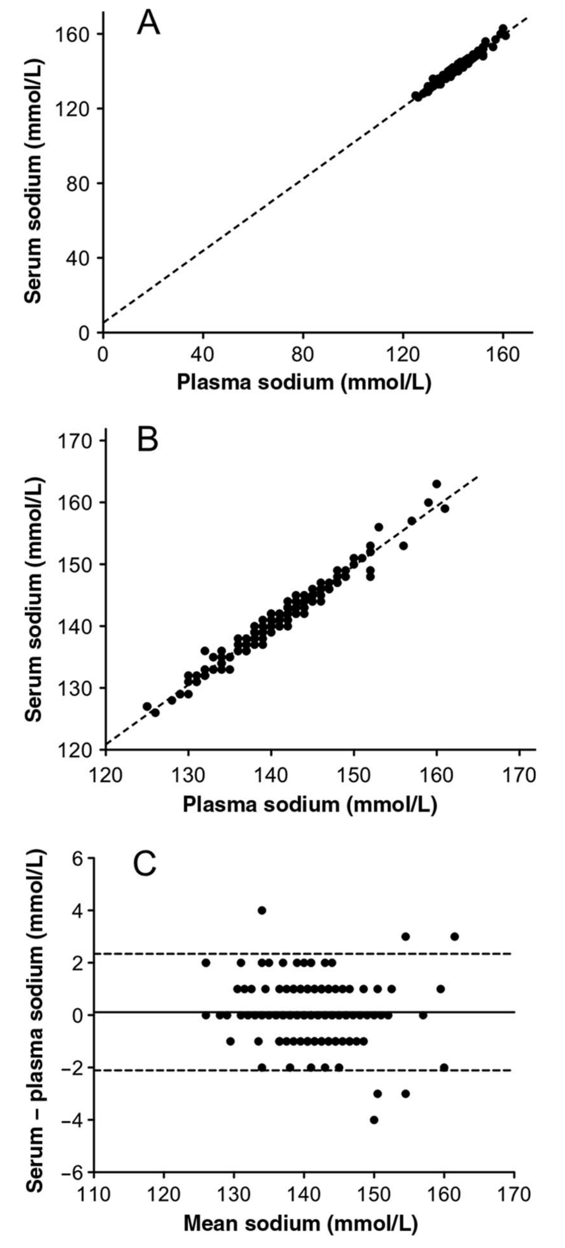

Fig. 3. Plasma vs serum sodium for paired specimens from 150 patients.

(A), *x*- and *y*-axis scales of 0–165 mmol/L; (B), *x*- and *y*-axis scales of 120–170 mmol/L; (C), Bland–Altman plot.

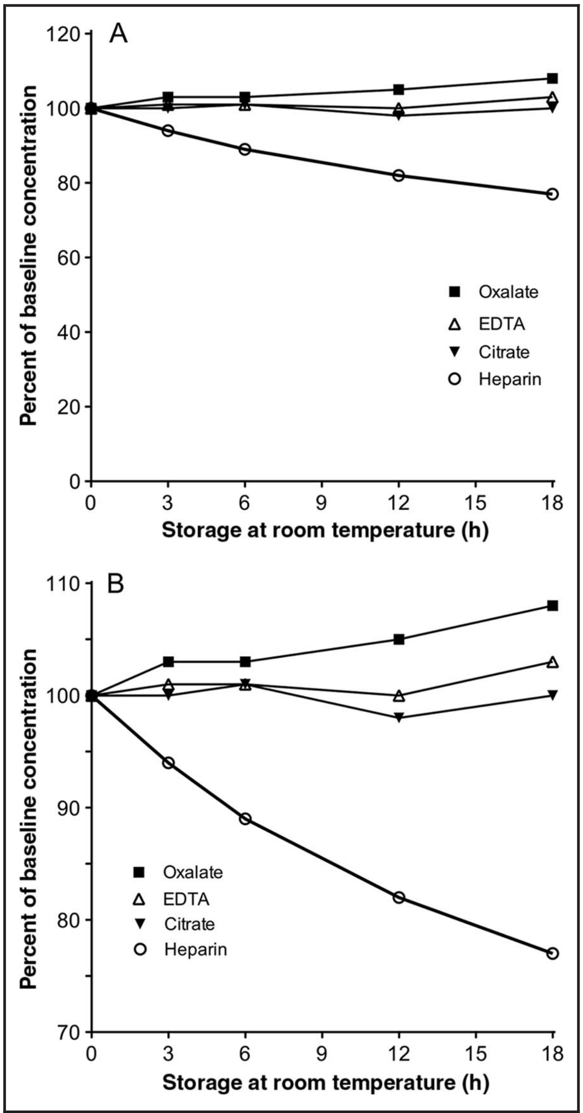

Fig. 4. Percent change in plasma alanine concentration after storage of whole blood at room temperature.

(A), *y*-axis scale of 0–110 μg/L; (B), *y*-axis scale of 70–110 μg/L.

The technique shown in Fig 4B is sometimes referred to as a *suppressed zero*. Although not necessarily bad if it increases the clarity of data presentation, one can see how a suppressed-zero scale can be misleading if used to exaggerate what would otherwise be small differences. The news media are often criticized for making graphs of economic data that have a restricted *y*-axis range, thus artificially magnifying the significance of any changes. So if you decide to use an expanded scale, bring this to the attention of the reader by stating directly in the figure legend that the scale has been expanded or does not start at 0.

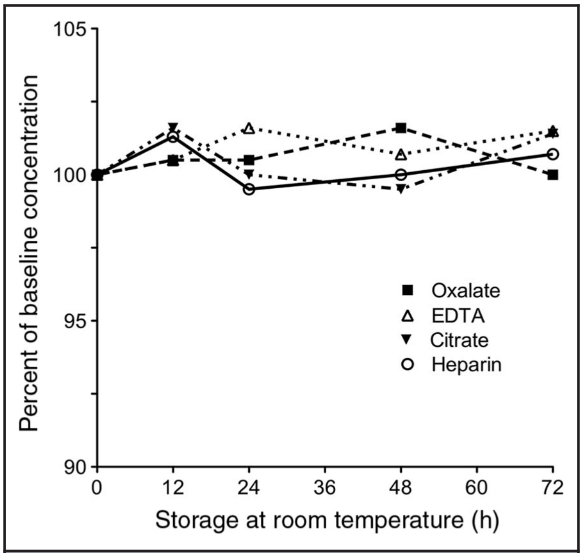

Fig. 5. Percent change in whole-blood tacrolimus concentration after storage at room temperature.

Fig. 5 shows a graph that meets all of the criteria for a good graph except one. Can you guess what it is? It is a good example of results that do not need to be presented as a graph. They are useful results and should be reported, but the message can be conveyed just as easily in the main text: "When whole blood specimens were collected into oxalate, EDTA, citrate, or heparin-containing tubes, and stored at room temperature for up to 72 hours, no statistically significant change in the tacrolimus concentration was observed for any of the tube types."

Learning Exercise

Using the information presented here about the characteristics of good and bad graphs, you should be able to identify features that add to or detract from the visual impact of a graph. The example shown in Fig. 6 has at least 12 problems. Can you identify these? Answers are provided in a box after the list of selected additional reading materials.

Final Thoughts

People are visual and expressive by nature, and authors (including this one) want to *show* what they have done. A picture can be worth a thousand words, but a few well-chosen words also can replace a picture. The key is to know when to use one or the other to most effectively state your message.

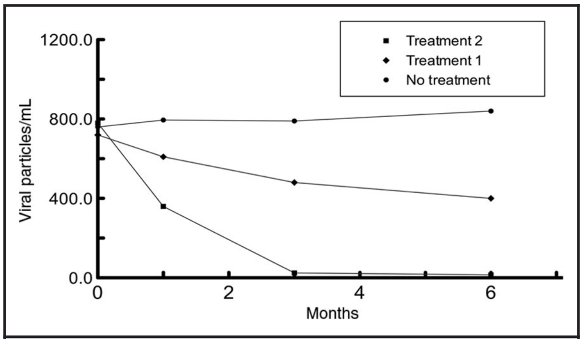

Fig. 6. Change in viral load during treatment; ●, 0 mg/kg; ◆, 5 mg/kg; ■, 20 mg/kg.

Resources and Additional Reading

Day RA, Gastel B. How to write and publish a scientific paper. Westport (CT): Greenwood Press; 2006.
Freeman JV, Walters SJ, Campbell MJ. How to display data. Malden (MA): Blackwell Publishing; 2008.
Gustavii B. How to write and illustrate a scientific paper. New York: Cambridge University Press; 2008.
Lang TA. How to write, publish, and present in the health sciences. Philadelphia (PA): ACP Press; 2010.
Zeiger M. Essentials of writing biomedical research papers. New York: McGraw Hill; 2000.

Author Contributions: *All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.*

Authors' Disclosures of Potential Conflicts of Interest: *Upon manuscript submission, all authors completed the Disclosures of Potential Conflict of Interest form. Potential conflicts of interest:*

Employment or Leadership: T.M. Annesley, AACC.
Consultant or Advisory Role: None declared.
Stock Ownership: None declared.
Honoraria: None declared.
Research Funding: None declared.
Expert Testimony: None declared.

Role of Sponsor: The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, or preparation or approval of manuscript.

Answer to Learning Exercise (Problems with Fig. 6)

The symbols are too small.

The symbols are too similar (solid box, solid circle, solid diamond) and are difficult to distinguish.

The data-connecting lines are narrow and do not draw attention to the data.

The text in the labels is small.

The *x* and *y* axes are too wide and draw the focus away from the data.

The numbers on the axes are proportionately too large.

The numbers on the axes are 2 different font sizes.

The *y*-axis numbers have an unnecessary decimal point.

The scale for the *y* axis is too large and creates wasted space.

The *x* axis says "months" and a fuller description may alleviate the need for the reader to refer to the main text.

The tick marks are on the inside of the axes and hide the symbols.

The ratio of the *x* axis to the *y* axis is too large (ideally 1.0 to 1.3)

The symbol legend within the graph identifies different treatments, whereas the figure legend identifies milligram per kilogram doses.

The symbol order (top to bottom) in the legend within the graph is different from the order (top to bottom) of the actual symbols in the figure.