Displaying data

Key points:

- Organize figures to set up a narrative
- Choose the format that best communicates what's important
- Make figures easy to interpret
    - Minimize non-data elements
    - Show result as directly as possible
    - Keep figures consistent
- Represent data honestly
- Compose figure legends to emphasize the result
    - Title states result
    - Just enough information to interpret figure

That a writing course would include material on figures might seem strange, but visual displays of information are a means of communicating, just like writing. The types of data only partially determine how they'll be displayed, and as you transform the raw data into figures, you make choices. For example, should you plot all the data as a scatter plot, or should you show the means and standard errors? Should you represent different treatment groups on a bar graph with different colors, or different shades of gray? The choices you make as you generate figures are not only a matter of taste, but also affect how the reader interprets the figure, so figure organization and design is an important part of composing a paper or proposal.

Organizing figures

The first decisions to make regarding data are which pieces of data belong in which figure, and in which order the figures and tables appear. (This section concerns papers and proposals about a series of experiments; the order of figures in a paper about one complex experiment is usually determined by relative importance.) These are possibly the most important decisions in this process, as the organization of figures determines the organization of the results section. Or at least it should—referring to a figure already explained interrupts the logical progression of the paper and makes it harder for the reader to determine what each result is. Therefore, *each figure should contain all of the data that support a single result,* so that each section of the results explains a single figure. All of the data obtained from the same method and displayed the same way may not belong together in a figure if that method was used to address multiple questions.

Once you have decided which pieces of data belong in each figure, you can place them in order. This order won't necessarily be the order in which you performed the experiments; instead, it should be *the order that tells the most interesting story possible*. For the figures to tell a story, they must be in logical order, so that you can explain each figure in relation to the previous one. To make that story as interesting as possible, imagine what question each figure or table would bring up and place the figure that addresses that question immediately after. For example, if one of your figures shows that your treatment alters a behavior, that brings up the question of how it does that—what signaling pathway does the treatment affect, and in which parts of the brain? Thus, placing this figure before those that explore the mechanism yields an interesting story. If no figure addresses that question, the figure either belongs at the end of the paper or in another paper entirely.

Choosing a format

Another early decision is how to display the data—table or graph? What type of graph? Should you include quantification or not? Sometimes your choices are dictated by the type of data you have. Most obviously, discontinuous data (where the values on the x axis are categories or where the distance between them does not reflect their numerical values) cannot be connected with lines. Also, non-parametric data (where the values in a group do not follow a normal or Gaussian distribution) should be indicated by medians rather than means, as the statistical tests for such data compare medians.

→ *Type of display reflects type of data*

The answers to questions of how to display data more often depend on what you're trying to say with the data. If the values themselves matter, use a table. For example, if you determined the k_m and k_{cat} of two isoforms of an enzyme using two substrates, these values should be listed in a table since determining their values would be more difficult if they were plotted on a graph than if they were simply listed. Another case in which tables are appropriate is when a few values of a variety of types of data must be presented, as in clinical studies that must describe the patient population by their demographic and health characteristics. Displaying multiple characteristics on the same graph would not make sense in this case, and a table is simpler than ten or twenty graphs with two bars each. Similarly, whether to plot your data as points with bars indicating error or as box-and-whisker graphs depends on whether the distribution of the data matters. If the point you make is simply that the means or medians differ, there's no need to show more than that. Determining whether to include quantification follows the same principle: if the magnitude of the difference in fluorescence intensity or band density matters, you must include quantification, but if the point is simply that one is greater than the other, you may not need to.

→ *Show only the aspect of the data that matters*

Designing figures for easy interpretation

The overall principle you may have deduced from the previous paragraph is that figures should only show as much as is necessary to make the point you intend and no more, as extraneous elements make identifying the point more difficult. A similar idea applies to the design of each panel of a figure: show the data and minimize everything else. This suggestion is best explained in Edward Tufte's *The Visual Display of Quantitative Information*, which uses hundreds of illustrations to support the idea and states it in a way that's easy to apply. Tufte advises designers of statistical graphics to *maximize the ratio of data-ink to total ink* used for a graphic. Data-ink refers to portions of a figure that directly represent data, for example, on a scatter plot, the dots are data-ink, while the axes, axis labels, legends/keys (the box on the graph, not the text below), and gridlines are non-data-ink. One way to maximize the data-ink ratio is to minimize the amount of non-data-ink by eliminating all nonessential elements, such as gridlines, and reducing the amount of ink used for other non-data elements as much as possible. Axis labels can be reduced by making them correspond to features of the data, such as minima and maxima, and the total amount of text may be reduced by eliminating redundant labels, such as titles that repeat information in the y axis label.

The purpose of following the data-ink guideline is to make interpretation of a figure easier. One way to do this is to minimize the amount of information the reader must keep in his or her mind while processing the figure. For example, if the point of a graph is to show that a drug causes a greater decrease in uptake in one genotype than another, the graph should show the change directly so the reader doesn't have to mentally calculate it. The graph should include one bar per genotype representing the uptake in the presence of the drug as a percent of uptake without the drug rather than two bars per genotype that show uptake in each condition. The reader also has to hold multiple pieces of information when understanding the figure requires shifting attention between the graphics and the legend. This can be avoided by *labeling as much of the graph as possible*, for example by using colored text in the upper corner of a fluorescence micrograph to indicate what each color represents.

Another way to make figure interpretation simpler is to *keep figures consistent* with the conventions of the field and with one another. For example, in representations of genes, exons are generally shown as boxes and introns as lines connecting them. Following the convention means that the reader doesn't have to spend time figuring out your scheme. Similarly, if you use colors to distinguish experimental groups in one panel or figure, using the same colors to identify the same groups in other figures reduces the effort required to understand later figures. Also, if multiple graphs are to be compared, use the same range for all of those panels so that the sizes of bars or positions of points have the same relationship to their value. This avoids possible misinterpretation, as readers may assume that the ranges match rather than reading the axis labels.

Designing figures for integrity

Possible misinterpretation may result from other figure design choices as well, and should obviously be avoided. Such choices include starting an axis at a value other than zero without indicating it by a line break or using a non-linear scale without indicating it with appropriate tick marks. These would both result in distances representing a data value that is not proportional to the value itself, and readers assume that distance is proportional to value. Similarly, readers may assume that the range of an axis represents the range of the data, so using a narrower axis range might also misrepresent the data by implying that it's not as widely distributed.

→ *Distances should be proportional to values*

Not only should you avoid figure design that misrepresents data (however unintentionally), you should also avoid design choices that may imply that you're intentionally misrepresenting it. For example, photos of blots or gels should have contrast that does not eliminate the background, as such high contrast might hide light bands. Also, when regions of different micrographs or photos of blots make up a single panel, they should be separated by white space. A line within the panel might suggest that one portion of a single photo was manipulated differently from the rest or that one part of the image was clone-stamped from another region. For further discussion of what is acceptable and questionable in image manipulation, see the Office of Research Integrity's website, which includes Photoshop videos: http://ori.dhhs.gov/education/products/RIandImages/default.html.

→ *Consider how your image manipulations will affect reviewer trust*

→ *Explain adjustments that selectively affect some regions or colors of the image*

Writing figure legends

A figure consists not only of labeled graphics, but also includes the legend, or text describing it. This text enables the figure to stand alone, so that the reader may understand it without reference to the text of the paper or proposal. The legend should include enough information about how the figure was generated for the reader to evaluate whether the data in it support the result the author says they do. For example, you might think that it's obvious that your image of a square with a bunch of bands on it is a Western blot, but the reader might be unsure about whether the bands are protein or RNA. On the other hand, interpreting a figure doesn't require every step of your protocol—you don't have to say which secondary antibody you used for a reader to be able to decide whether your Western shows a decrease in expression of a protein in response to your treatment. Including minor details makes it harder for the reader to find the information she really needs; if she questions the data, she can consult the methods section. A lot of information may be necessary to understand the figure if it results from an unusual, complex method; in this case, consider explaining your approach graphically by adding a panel to the figure.

→ *Include only information in a legend necessary to interpret the figure*

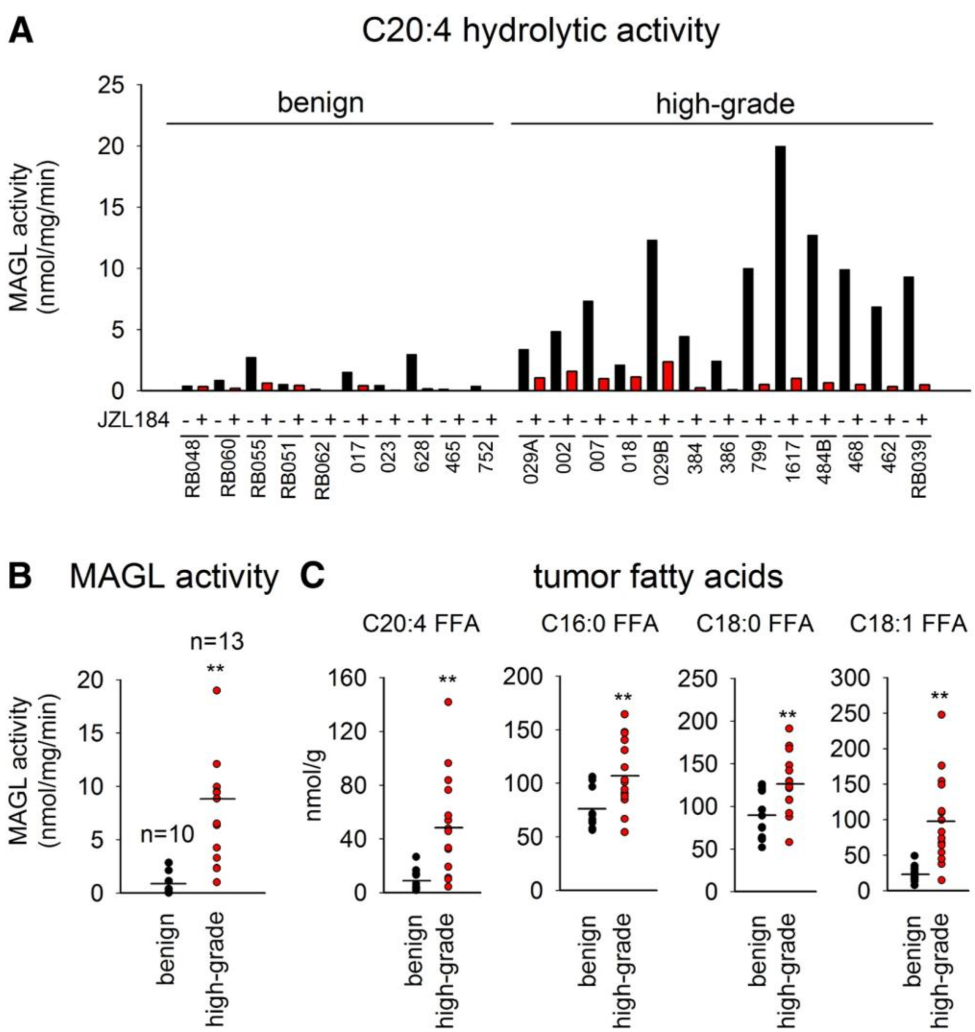
The purpose of figure legends stated above might differ from the one you know—it's not just to help the reader interpret the figure, but to compare that interpretation to the author's. In order to do that, the reader must know what your interpretation is: what result they support. The result is, of course, in the results section, but making the reader search through the text expects a lot of effort, and some (especially experts in the field) may not look at anything but the figures. Thus, you should include it in the legend; making it the title is an effective way to emphasize the point of the figure. You might also want to state the result each panel shows, but this requires caution, since interpreting the figure in the legend makes it difficult for the reader to draw her own conclusions. Further, since interpretation (such as comparisons between groups or other observations about the data) belongs in the results section, including it in the legend confounds the reader's expectations (and may make it seem that you don't know the conventions of scientific writing or that you copied and pasted from the results section). Nonetheless, if each panel's result is best stated differently from the overall result of the figure, you can state them in the panel titles, followed by a basic explanation of how you got the data.

→ *Interpret the figure only in the title(s)*

The guidelines laid out above may make more sense when used to evaluate a figure. Here is a fairly well-designed one:

Example: Unified figure with internal inconsistencies

Figure 3. High-Grade Primary Human Ovarian Tumors Possess Elevated MAGL Activity and FFAs Compared to Benign Tumors. (A) C20:4 MAG hydrolytic activity measurements for individual tumor specimens. Pretreatment with JZL184 (1 μM, 30 min) confirmed that the majority of the observed hydrolytic activity is due to MAGL. (B) Summary graph of MAGL activity in benign versus high-grade tumors, where each value is expressed as the JZL184-sensitive portion of total C20:4 MAG hydrolytic activity shown in part (A). (C) Levels of FFAs in benign versus high-grade tumors. **p < 0.01 for high-grade versus benign tumor groups. Data are presented as means ± SEM; n = 10–13/group



From Daniel K. Nomura et al., Monoacylglycerol Lipase Regulates a Fatty Acid Network that Promotes Cancer Pathogenesis. *Cell*, Volume 140, Issue 1, 49-61, 8 January 2010. 10.1016/j.cell.2009.11.027

Analysis of example

This figure is unified: all panels support the same result, as the title is a single statement. The title helps the figure stand alone by making clear what result it's supposed to show. However, the legend for A goes a bit beyond helping the figure stand alone by interpreting the results of JZL184 treatment. Perhaps the authors intended to explain the purpose of the drug, in which case they could have said what JZL184 is—an MAGL inhibitor. Otherwise, the scope of the legend is appropriate: both methods are common enough that the reader can infer how the authors did their experiments. The legend does appear to include a minor error in its mention of SEM, as no error bars are included on any of the graphs.

Each graph, when considered individually, is designed well. Each is as simple as that type of graph can be, and the dots in B and C are large enough to see but small enough to distinguish from one another. Further, they're labeled thoroughly—almost no reference to the legend is required.

When the panels are considered together, however, we notice some inconsistencies. First, the data in A are displayed differently from those in B despite the fact that they're both showing hydrolytic activity. It's unclear why a bar graph is necessary to show the effect of JZL184; a scatter plot would show equally well that the drug decreases activity and use much less space. Possible explanations for this choice are that the authors wanted to show the activity of each tumor specimen or that JZL184 decreases activity in every tumor (one +JZL data point matches the –JZL activity of another specimen with higher overall activity). Another difference that may cause even more confusion is the different uses of the same color: in A, red indicates treatment with JZL, while in B and C it indicates high-grade tumors. The reader might first think that the panels disagree, since the red data points are lower than the black ones in A and higher in B and C. Using a different color for high-grade tumors would avoid this confusion. The most problematic inconsistency is the different scales for similar panels, as the reader may incorrectly conclude that tumors contain similar levels of C16:0 FFA as C18:0 FFA, when levels of C18:0 FFA are actually somewhat higher.