The Illusion of Certainty and the Certainty of Illusion: A Caution When Reading Scientific Articles

TA Lang

Abstract

Critical thinking is necessary to read the scientific literature. However, in addition to questions about the science, often one must also question the meaning of the text. This article provides an example of the analyses needed to understand a single sentence. In so doing, it raises several interesting issues of meaning, measurement, statistical analyses, and the form in which results are presented and interpreted.

Keywords: Statistics; Literature; Reporting standards

Science is based on writing. Only writing allows science to be recorded, evaluated, reproduced, systematic, cumulative, and public; the characteristics that distinguish it from authority, intuition, and tradition as a way of establishing “truth.” Publication—the final stage of research—also depends on writing, as does evidence-based medicine, which is literature-based medicine.

Given the importance of scientific writing, technical writing skills should be a larger part of the biomedical curricula than they are. However, at least in clinical medicine, physicians know they need help searching the literature and so consult librarians. They know they need help designing studies and analyzing data and so consult statisticians. They know they need help preparing visuals for their manuscripts and presentations and so consult medical illustrators and photographers. But they don’t know (or won’t acknowledge) that they need help reporting and publishing their research and so do not often consult medical writers and editors. Thus, a large portion of the scientific literature is not readily understandable.

As a medical writer-editor for 35 years, I have edited thousands of scientific articles. Not merely read them, but dissected, studied, rewritten, revised, expanded, shortened, and otherwise reconstituted them until they reported the research accurately, completely, clearly, and economically. One of the most important lessons I have learned is that a simple, straightforward sentence often presents the illusion of certainty. The illusion is so strong that there seems to be no reason to question its meaning. Only on closer examination does the illusion become apparent. Further, such sentences are found in the vast majority of scientific manuscripts, which is to say, these illusions are also a certainty in the scientific literature.

An example is the sentence, “She is a woman with many convictions.” The most common first interpretation is that the woman holds strong beliefs. When asked to provide another meaning, many people realize that the sentence may also...
mean that the woman has been convicted of many crimes. However, few people identify a third possibility: that she was a successful prosecutor who had convicted many criminals. Such ambiguity is the worst enemy of scientific writing.

Some time ago, I found a more interesting example of a sentence with meaning far beyond its apparent one. In this article, I pose the questions that need to be answered to understand the sentence and its implications. I didn’t save the original manuscript, so I’ve had to make up some numbers, but the concepts are clear.

The Example

This sentence was in the Results section of a poorly-written abstract: “One group of patients was significantly more depressed than the other.” The sentence seemed straightforward, but the more I thought about it, the more questions I had about it.

The Questions

Q1: What is the context in which the sentence appears?

The sentence was the second one in the results section of the abstract. However, it was not immediately apparent whether the sentence was a description of the patients at baseline, an incidental finding that might confound the results, or a result itself. Eventually, from the rest of the article, I identified it as a result of the study. Meaning is a product of message and context. Change the context, and the same message can be interpreted differently. “The wall was built to scale” means something different to an architect than it does to a climber. For this reason, the context of every scientific article needs to be clear to rule out other interpretations made possible by different contexts.

The textual factors of message and context are similar to the visual factors of image and background. The image (or message) must be shown against a background (context) to produce meaning. The interaction is illustrated by images in which the meaning changes depending on what is seen as image and what is seen as background (Figure 1).

Q2: Who and what were studied?

I then tried to identify the two groups. It turns out that the groups were the treatment and control arms. Again, context made this identification easier. The sentence could have read, “Patients in the control group were significantly more depressed than patients in the treatment group.” We can’t stop here, however. The sample size at the beginning and end of the study need to be reported to determine how many patients complete or withdrew from the study or were lost to follow-up. These numbers become the denominators for intent-to-treat analysis (the number of patients assigned to each study group) and for per-protocol analysis (the number
of patients who completed the protocol as planned).

**Q3:** How was depression measured?

Science is based on measurement, so all study variables must be defined in objective, measurable terms. In this case, depression was measured with the Beck Depression Inventory, a common, validated instrument for measuring depression. This finding was encouraging because many authors do not say how they measured their variables. For example, it is not enough to refer to “hypertensive” patients; a measurable definition of hypertension must be given, such as a systolic blood pressure of 140 mm Hg or greater.

Studies of adolescents often define “current smoker” as any respondent who has smoked at least one cigarette in the previous 30 days. We can disagree with this definition, but at least we know how “current smoker” was defined.

**Q4:** How large was the difference?

Here, the authors reported that “The mean depression score of the treatment group was 38% lower than that of the control group.” All well and good, but results expressed as percentages are always suspect. Numerators and denominators should always be available for all percentages.

One of my favorite examples of the inappropriate use of percentages is the old laboratory joke about how 33% of the rats lived, 33% died, and the last one got away. It is also usually true that a 50% reduction from 2 to 1 is not the same as a reduction from 2000 to 1000. Be careful when interpreting percentages!

In a well-known example, the results of a study testing a drug to prevent myocardial infarction were reported in two ways. The absolute risk reduction is the difference in the event rates between treatment and control groups (In this study, the event rate was the proportion of patients experiencing a heart attack.). The relative risk reduction is the difference in the event rates between treatment and control groups expressed as a percentage of the event rate in the control group. In the example, the event rate in the control group was 4.1% (4.1% had heart attacks) and that of the treatment group was 2.7%. Thus, the absolute risk reduction was 1.4% (4.1% – 2.7% = 1.4%), and the relative risk reduction was 34% (1.4% / 4.1% = 34%). Guess which number appears in the ad for this drug ...

Back to the example; the difference between means was 3 points.

**Q5:** What does the author mean by “significantly?”

In medical writing, “significant” is reserved for its statistical meaning, but the term is still often used to mean “markedly” or “substantially.” An accompanying p value or a 95% confidence interval usually mean the term is used for its statistical meaning, but one still has to be careful. In the present example, “significant” was used in its statistical sense.

Probably the most common reporting error in medical articles is assuming that a statistically significant result is also clinically important. Because p values provide a simple way to interpret data, they are often the rationale for determining whether a result is important. However, even if interpreted correctly, p values themselves must be reported correctly. In particular, we need to know the alpha level (usually 0.05), the value defining statistical significance; the statistical test used to calculate it; whether the assumptions of the test (e.g., normal distribution of data) have been met by the data, whether the test was one- or two-tailed; and the statistical software program used in the analysis (to establish that it is a known and validated program).
Q6: Is the difference clinically important?
Most authors are quick to report statistically significant differences, but many never bother to say whether the difference is clinically important. In fact, the effect size (say, the differences between means) is usually more important than the p value (As someone once said, “Group means do not present for treatment.”). The effect size can be interpreted clinically, whereas a p value is a measure of chance.

Especially in large datasets, small and clinically unimportant differences may nevertheless be statistically significant, simply because of the amount of data analyzed. I once edited a study of the useful life of pacemaker leads. The author had analyzed data from thousands of leads, most of which were one of four brands. He was very excited because he found that one brand had a statistically significantly shorter useful life than that of the other three. However, the mean useful life of the three other brands was about 60 months, and that of the fourth was ... 59 months, in this case, not really an important difference, despite a significant p value.

Eventually, the authors of our example revealed that the difference between the means of the treatment and control groups was 3 points.

Q7: How precise is this estimate?
The results of most biomedical studies are in fact estimates, and estimates require a measure of precision. In clinical medicine, this measure is usually the 95% confidence interval. I don’t know what it means to be 95% confident, so I think of the interval as being the range in which the mean difference would occur in 95 of 100 similar studies.

Confidence intervals are useful because they keep the interpretation focused on the effect size and therefore on the medicine or biology, not the p value. Confidence intervals that contain both clinically important and clinically unimportant values, however, suggest that, even if the difference in means is statistically significant for the current trial, the estimate is not precise enough to conclude that the treatment will likely be effective in 95 of 100 similar trials; that is, the result is clinically inconclusive.

Eventually, the authors of our example reported that “The difference between means was 3 points (95% Confidence Interval: 1.5 to 4.5 points).”

Q8: What is the measurement scale for the estimate and confidence interval?
A little research revealed that the Beck Depression Inventory is a scale that runs from 0 to 63. For people with a history of depression, scores of 0 to 9 indicate no or minimal depression; from 10 to 18, mild depression; from 19 to 29, moderate depression, and 30 to 63, severe depression. So, the 3-point difference between means, and its 95% confidence interval, has to be interpreted in the context of a scale that ranges from 0 to 63.

Q9: What is the smallest clinically meaningful difference?
Even if the difference between means is statistically significant, the clinical implications of a difference should be described. In the current example, what are we to make of a 3-point difference? Or a 1- or a 5-point difference? Does it matter whether the difference crosses one of the threshold scores that define a different severity of depression? Does it matter whether the difference occurs at the low end or high end of the scale?

For example, there is some evidence that pain measured on a 10-cm visual analog scale is nonlinear; that is, the drop from 9 to 8 is more pronounced than a drop from 4 to 3. Pain is notoriously diffi-
cult to measure, but we have no assurance that many scales reflect a linear relationship among scores.

I couldn’t determine the smallest clinically meaningful difference on the Beck Depression Inventory, but perhaps my research was not thorough enough. The clinical meaning of a 3-point difference is thus unclear. As a percentage of the full range of scores (0 to 63), a 3-point difference is a 5% change, although I don’t know whether this calculation means anything.

**Q10:** What were the actual mean values?

This is where things got interesting. One of the tables in the article showed that the mean score was 5 in the treatment group and 8 in the control group. These values are consistent with the 3-point difference between means and with the 38% lower score of the treatment group. However, both means (5 and 8) are in the range of no or minimal depression (scores 0 to 9), so the interpretation of the finding that “one group is more depressed than the other” is incorrect. My guess is that the authors based their interpretation of the result solely on a significant p value.

**Q11:** What was the proportion of patients in each group who were depressed after treatment?

The example compared the mean values of two groups on a continuous scale. However, a common error in reporting the results of clinical research is to focus on changes or differences in means rather than indicating how many patients got better or worse (Figure 2). In the manuscript, I had hoped for something like: “At the end of the study, 90% of patients in the treatment group and 63% of those in the control group had scores below 10, indicating no or mild depression,” but these data were not reported. If depression was the main outcome, one hopes that a score of 10 or higher, indicating depression, was an eligibility criterion for entry into the study.

The issue here is the “unit of observation.” I once edited a manuscript describing a study of 25 eyes, but I never determined whether there were 25 patients or only 13. The unit of observation was eyes, but often the number of patients involved needs to be given. The number of heart attacks, for instance, may not correspond to the number of patients if some of the patients have had two or more heart attacks.

**Conclusions**

Not all sentences are this involved, but many are and require analysis as detailed as the example presented here. This analysis can take time—and skill and training and experience—sometimes even when the meaning of the sentence is really cor-
rect. What makes good writing and good editing valuable is that they reduce readers’ time and effort in analyzing a text in this way. The problem is that many scientific articles are poorly written and poorly edited. Most authors have never been taught how to communicate technical information in writing, and most journals do not have the time to edit a paper thoroughly. This situation pretty much assures that readers of the scientific literature will regularly encounter the “illusion of certainty” and therefore must be prepared for the “certainty of illusion.”

**Conflicts of Interest:** None declared.

**References**


---

**Visit Us on the Web**

www.theijoem.com  
www.theijoem.org