CHEST®

Official publication of the American College of Chest Physicians

Documenting Research in Scientific Articles: Guidelines for Authors: 2. Reporting Hypothesis Tests

Tom Lang

Chest 2007;131;317-319
DOI 10.1378/chest.06-2087

CHEST Postgraduate Education Corner

MEDICAL WRITING TIPS

Documenting Research in Scientific Articles: Guidelines for Authors*

2. Reporting Hypothesis Tests

Tom Lang, MA

(CHEST 2007; 131:317–319)

Proposed by Sir Ronald Fisher in 1920 as a measure of the strength of evidence, p values are part of an area of statistics called the *frequentist* approach to statistics. Also a part of the frequentist approach is a method of choosing between hypotheses, called *hypothesis testing*, which was developed by mathematicians Jerzy Neyman and Egon Pearson in the 1930s. Probability values and hypothesis testing are actually quite different concepts, but they are widely, if mistakenly, seen as parts of a coherent approach to statistical inference.[1] In fact, the frequentist approach is widely used in biomedical research. Although the logic behind it is elegant, it is not intuitively obvious, which is why it is so often misunderstood. The guidelines here should help to make reports of hypothesis testing more complete. The guidelines here have been condensed from those presented in *How To Report Statistics In Medicine*.[2]

Guideline: State the Hypothesis Being Tested

A hypothesis is a testable statement about a proposed relationship between two or more variables. Either the null hypothesis of no difference (to be disproven by the study) or an alternative hypothesis to be supported by the study can be reported.

Guideline: Specify the Minimum Difference Between the Groups That Is Considered To Be Clinically Important

Specifying in advance the minimum clinically important difference between groups keeps the analysis focused on clinical issues and helps to put statistical issues in perspective. The minimum difference is also a component of the statistical power calculation, which helps to determine how large a sample should be.

Guideline: Specify the α-Level, the Probability Below Which Findings Will Be Considered To Be "Statistically Significant"

The α-level is the probability chosen by the researcher to be the threshold of statistical significance. It is actually the probability of committing a type I error or, essentially, of wrongly concluding that the difference between groups was the result of the intervention. The α-level is an arbitrary value but, by tradition, is usually set at 0.05, 0.01, or, less commonly, 0.001. In any event, p values less than the α-level are, by definition, "statistically significant."

Guideline: Identify the Statistical Test Used for Each Comparison

There are many, many statistical tests, and several may be appropriate for the comparison in question. Each test is based on several assumptions, however, so it is important to specify which test was used for each analysis. Cite a reference for complex or uncommon statistical tests.

Guideline: If Appropriate for the Test, Specify Whether the Test Is One-Tailed or Two-Tailed, and Justify the Use of One-Tailed Tests

A two-tailed test (based on a symmetrical distribution of probabilities) divides the α-level, usually

*From Tom Lang Communications and Training, Davis, CA.
The author receives royalties from the sale of *How to Report Statistics in Medicine*, from which this article is taken. He has no other conflicts of interest with the publication of this article.
Manuscript received August 23, 2006; revision accepted August 24, 2006.

Correspondence to: Tom Lang, MA, 1925 Donner Ave, No. 3, Davis, CA 95618; e-mail: tomlangcom@aol.com
DOI: 10.1378/chest.06-2087

0.05 (5%) into the following two parts: 2.5% for the cases in which group A has an end point larger than group B; and 2.5% for the cases in which group A has an end point smaller than group B. That is, if an intervention may make group A either better or worse than group B, a two-tailed test considers both possibilities. A one-tailed test, on the other hand, puts the 5% in only one tail (or direction), if the direction of the result is presumed to be known in advance.

Two-tailed tests require a greater difference to produce the same level of statistical significance (*ie*, the same p value) as one-tailed tests. They are more conservative and are often preferred for this reason. One-tailed tests are used when the direction of the results (not necessarily the magnitude) is known in advance, which is often the case. When using one-tailed tests, researchers should identify the tests as such and give the evidence for knowing the direction of the result.

Guideline: Reference the Statistical Packages or Programs Used To Analyze the Data

Although commercial statistical software packages generally are validated and updated, privately developed programs may not be. In addition, not all statistical software packages use the same algorithms or default options to compute the same statistics. Thus, the results may vary slightly from package to package or from algorithm to algorithm.

Guideline: Report the Results of All Primary Analyses First

The focus of a scientific article should be on the primary comparisons that motivated the work. Statistical analysis can and should be exploratory and interpretive to a point, but these secondary explorations should never overshadow the primary analyses. That is, unsupported (statistically nonsignificant) primary analyses should not be neglected for more intriguing (statistically significant) secondary analyses.

Selective reporting is the practice of presenting only the desirable findings of a study. Such findings are usually those that are statistically significant. The results of all clinically relevant analyses should be reported, whether or not they are statistically significant. It is unethical to suppress contradictory data.

Guideline: Report the Actual Difference and the 95% Confidence Interval

The difference (often, between the means of the groups) associated with the p value should be reported. This difference is an estimate and should therefore be accompanied by a measure of precision, usually the 95% confidence interval. Many authorities now prefer confidence intervals to p values when reporting results because confidence intervals keep the discussion focused on the size of the effect and away from chance as an explanation.

Guideline: Confirm That the Assumptions of the Test Have Been Met

Most statistical tests make assumptions about the data. If these assumptions are suspect, the results of the analyses may also be suspect. A statement that the assumptions were verified is all that need be included.

A common assumption is that the data are approximately normally distributed, a characteristic that permits the use of "parametric" tests. This assumption is often violated. When data are markedly nonnormally distributed, a mathematical "transformation" may be appropriate to make the distribution more normal, or a "nonparametric" test (which does not require data to be normally distributed) may be used instead. If data have been transformed or analyzed with nonparametric tests, this fact should be reported.

Guideline: Give the Actual p Value, to Two Significant Digits, Whether or Not the Value Is Statistically Significant

Probability values less than the α-level (usually 0.05) are considered to be statistically significant; those greater than α are not. However, the p values of 0.051 and 0.049 are close enough that they should be interpreted similarly, despite the fact that the first would be reported as "not significant," and the second as "significant." Providing the actual p value prevents this problem of interpretation. In any event, the smallest p value that needs to be reported is $p < 0.001$.

If the results are not statistically significant, do not use the phrase "showed a trend toward significance" or "approached significance." The result was simply not statistically significant, as defined by the relationship between the p value and the α-level. (Curiously, p values never seem to "trend" away from significance!)

Guideline: Indicate Whether and How Any Adjustments Were Made for Multiple Comparisons

The "multiple comparisons" (or multiple testing) problem is that as more hypotheses are tested on the

same data, the more likely the chance is of making a type I error, or concluding that a difference is the result of an intervention when, in fact, chance is the more likely explanation. For example, assuming that the threshold of statistical significance (α) has been set at 0.05 and 100 p values have been calculated from the same data, 5 of these p values are likely to be less than 0.05 just by chance. In many instances, multiple tests are unavoidable and even desirable, but they must be dealt with carefully to avoid the multiple testing problem.[3]

Multiple testing is often encountered when:

- Establishing group equivalence by testing each of several baseline characteristics or prognostic factors for differences between experimental and control groups (hoping to find none);
- Performing multiple pairwise comparisons, which occurs when three or more groups of data are compared two at a time in separate analyses, as is done in analysis of variance and multiple regression analysis;
- Testing multiple end points that are influenced by the same set of explanatory variables;
- Performing additional, secondary analyses of relations observed after the data have been collected and not identified in the original study design;
- Performing additional, subgroup analyses not planned in the original study;
- Performing interim analyses of accumulating data (*ie*, one end point measured at several different times), which is often done in studies involving potentially toxic or harmful effects to avoid putting study participants at risk unnecessarily; and
- Comparing groups at multiple time points with a series of individual group comparisons.

Of concern with multiple testing is the phenomenon of *data dredging* (the practice of indiscriminately analyzing any and all relationships and reporting those with statistically significant results).[4–6] Historically, great but undue value has been attached to "statistically significant findings" or "positive results." Unfortunately, many authors do seem to engage in a "ruthless search for significance"[7] in an attempt to find statistically significant relationships to report.

Multiple testing can be useful, however. Although the formal experiment is designed to produce answers to specific questions, exploring the data with additional analyses (multiple testing) may help to generate better questions.[8] However, such exploratory analyses must also be interpreted wisely: "Hypothesis-generating studies (sometimes referred to somewhat contemptuously as 'fishing expeditions') should be identified as such. If the 'fishing expedition' catches a boot, the fishermen should throw it back, not claim that they were fishing for boots."[9]

Guideline: Distinguish Between Clinical Importance and Statistical Significance

The most common reporting error in biomedical research is confusing statistical significance with clinical importance. A p value has no clinical interpretation. The clinical importance of the finding should incorporate the overall quality of the study, the size of the difference or the strength of the relationship found, and the biological implications of the findings, in addition to the p value.

ACKNOWLEDGMENT: This article draws heavily from *How To Report Statistics in Medicine*, by Tom Lang and Michelle Secic.[2]

References

1 Goodman SN. Toward evidence-based medical statistics: 1. The P value fallacy. Ann Intern Med 1999; 130:995–1004
2 Lang T, Secic M. How to report statistics in medicine. 2nd ed. Philadelphia, PA: American College of Physicians, 2006
3 Chalmers TC, Smith H Jr, Blackburn B, et al. A method for assessing the quality of a randomized control trial. Control Clin Trials 1981; 2:31–49
4 Bailar JC III, Mosteller F. Guidelines for statistical reporting in articles for medical journals: amplification and explanations. Ann Intern Med 1988; 108:266–273
5 Haines SJ. Six statistical suggestions for surgeons. Neurosurgery 1981; 9:414–418
6 Smith DG, Clemens J, Crede W, et al. Impact of multiple comparisons in randomized clinical trials. Am J Med 1987; 83:545–550
7 Morgan PP. Confidence intervals: from statistical significance to clinical significance [editorial]. Can Med Assoc J 1989; 141:881–883
8 Schoolman HM, Becktel JM, Best WR, et al. Statistics in medical research: principles versus practices. J Lab Clin Med 1968; 71:357–367
9 Mills JL. Data torturing [letter]. N Engl J Med 1993; 329:1196–1199

Documenting Research in Scientific Articles: Guidelines for Authors: 2. Reporting Hypothesis Tests

Tom Lang

Chest 2007;131;317-319
DOI 10.1378/chest.06-2087