Improving the Quality of the Reporting of Research Results

Jay F. Piccirillo, MD

In *JAMA Otolaryngology-Head & Neck Surgery*, we strive to present the highest quality clinical, translational, and population health research from an array of disciplines aligned with the clinical practice of otolaryngology-head and neck surgery. Many problems exist in the conduct and analysis of clinical research—such as faulty or incorrect study design to answer the chosen research question, bias in the selection of study participants and measurement, improper attention to the role of chance, and incorrect use of statistical tests—to undermine the validity of the published results. In this editorial it is not my intention to describe these problems in detail; instead, I will focus on the problems in results reporting and provide some solutions that we will foster in the journal. I believe results reporting has received too little attention given its enormous importance for evaluating the significance and effect of research results. Indeed, without accurate reporting, the whole research endeavor might be meaningless or even misleading. I will offer solutions to the problems of results reporting, some of which must be implemented during the planning process and well before the conduct of research. Other solutions to the problems in the conduct and analysis of clinical research address the challenges of data analysis, interpretation of results, and results reporting, all of which have been presented previously in different forums. To illustrate my main points, I will use a 2-group randomized trial study design, where average values for the experimental group are compared with the average values for the control group. The points are relevant to almost all other study designs and analytical approaches.

When conducting and reporting the results of clinical research, we are often interested in knowing “how much” of a difference exists between compared groups. The observed results of the study help us infer value in the target population. Unfortunately, this search for how much of a difference is possible in the population given the observed differences in the study sample is not what is reported in much of the published biomedical research. Instead, we commonly see that after completion of data collection, the investigator will calculate the difference between study groups and obtain the probability (P value) of obtaining the observed difference or greater if the null hypothesis is true. In this common practice, if the P value is less than the prestudy-specific α level, then the results are deemed significant. If the P value is greater, the results are deemed nonsignificant. Sometimes important differences are disregarded because those differences did not reach statistical significance. Instead of presenting information relevant to how much of a difference was observed in a study, the practice of reporting results with P values only erroneously focuses on whether the observed difference is significant from a statistical standpoint. How did we get to this common but misleading practice of reporting statistical significance under the null hypothesis?

In his 1925 textbook *Statistical Methods for Research Workers*, Fisher first introduced the word significance to the interpretation of differences between samples: “The value for which \( P = .05 \), or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is considered to be significant or not. Deviations exceeding twice the standard deviation are thus formally regarded to be significant.” This value of 5% was deemed by Fisher as a convenient boundary to judge whether a difference was to be considered significant or not. This boundary of 5% became entrenched in all types of biomedical research when it was adopted by various regulatory agencies (ie, the US Food and Drug Agency), funders and grant reviewers, and journal editors as a statistical boundary for determining significance of the completed research. This rigidity to a single convenient value for the boundary of statistical significance is in contrast to Fisher’s own suggestion for flexibility in interpreting results captured by this quote, “… no scientific worker has a fixed level of significance at which from year to year, in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in light of his evidence and his ideas.”

The traditional research paradigm, referred to as null hypothesis statistical testing (NHST) and generally ascribed to Fisher rivals Neyman and Pearson, requires the establishment of a level of significance, a null and alternative hypothesis, acceptance and rejection of the null hypothesis, and \( \alpha \) (type I) and \( \beta \) (type II) errors. Among its many deficiencies, NHST does not provide the answer to the important questions of how much of a difference truly exists between compared groups and how precise the estimate of difference. In an attempt to improve the results reporting in *JAMA Otolaryngology-Head & Neck Surgery*, I present some of the problems with the current state of results reporting and suggest solutions to these problems. A more detailed description of statistical reporting in the medical literature is available in the fine article by Cummings and Rivara.

Problems With P Values

Recently, the American Statistical Association (ASA) published a statement on statistical significance and P values. In that statement, the ASA presented a few key principles illustrating the problems with relying on \( P \) values when conducting research, as well as the proper use and interpretation of \( P \) values. A selection of the problems and principles most relevant to results reporting with \( P \) values and informed by the ASA statement are presented below.
1. P Values Are Inherently Unstable
One of the fundamental principles of science is replication. Unfortunately, the nature of NHST and P values is such that when experiments are repeated, the actual P values can vary to a surprising degree—even when the statistical power is high—for each experiment sampling from the same population with the same effect size.17 This instability in P values has been referred to as the “dance of the P values.”10

2. The P Value, or Statistical Significance, Does Not Measure the Size of an Effect or the Importance of a Result
P values do not present information regarding how much of a difference is compatible with the observed results, and statistical significance is not equivalent to clinical significance. As expressed in the ASA statement, “Smaller P values do not necessarily imply the presence of larger or more important clinical effects, and larger P values do not imply a lack of importance or even lack of clinical effect. Any effect, no matter how tiny, can produce a small P value if the sample size is large enough or measurement precision high enough, and large effects may produce unimpressive P values if the sample size is small or measurements are imprecise. Similarly, identical estimated effects will have different P values if the precision of the estimates differs.”

3. Scientific Conclusions Should Not Be Based Only on Whether a P Value Passes a Specific Threshold
Practices that reduce data analysis or scientific inference to mechanical bright-line rules—such as P < .05—for justifying scientific claims or conclusions can lead to erroneous beliefs and poor decision making. As stated in the ASA statement, “A conclusion does not immediately become “true” on one side of the alpha divide and “false” on the other. Researchers should bring many contextual factors into play to derive scientific inferences, including the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis.” The widespread use of statistical significance, interpreted as P < .05, as a license for making a claim of a scientific finding or implied truth leads to considerable distortion of the scientific process. It is incorrect to think that the probability of a conclusion being in error can be calculated from the data in a single experiment without reference to prior research or the plausibility of the underlying mechanism.

4. Proper Inference Requires Full Reporting and Transparency
As asserted in the ASA statement, “P values and related analyses should not be reported selectively. Conducting multiple analyses of the data and reporting only those with certain P values (typically those passing a significance threshold) renders the reported P values essentially uninterpretable. Cherry picking promising findings, ...leads to a spurious excess of statistically significant results in the published literature and should be vigorously avoided.” The statement goes on to recommend, “Researchers should disclose the number of hypotheses explored during the study, all data collection decisions, all statistical analyses conducted, and all P values computed. Valid scientific conclusions based on P values and related statistics cannot be drawn without at least knowing how many and which analyses were conducted, and how those analyses (including P values) were selected for reporting.” Kirkham and Weaver recently conducted an excellent review of the frequency of multiple hypothesis testing in the otorhinolaryngology literature and strategies for adjusting for multiple testing.18

5. By Itself, a P Value Does Not Provide a Good Measure of Evidence Regarding a Model or Hypothesis
As stated in the ASA statement, “Researchers should recognize that a P value without context or other evidence provides limited information. For example, a P value near .05 taken by itself offers only weak evidence against the null hypothesis. Likewise, a relatively large P value does not imply evidence in favor of the null hypothesis; many other hypotheses may be equally or more consistent with the observed data. For these reasons, data analysis should not end with the calculation of a P value when other approaches are appropriate and feasible.”

Solutions to the P Value Problems
There are several solutions to the P value problems in results reporting.10,19 These solutions are presented here (and in the JAMA Otolaryngology—Head & Neck Surgery Instructions to Authors20) with the expectation that authors will adhere to them when submitting manuscripts for consideration for publication. Often manuscripts can be immediately improved by reporting the results more broadly as outlined below and avoiding reducing the results simply to meeting or not meeting statistical significance alone. Additional solutions to the P value problem, such as Bayesian analysis and bootstrapping, will not be presented here but are equally satisfactory ways to avoid the P value problems.

1. Estimation of the Size of the Effect
Investigators should avoid formulating their research question and reporting the results as a dichotomous expression of testing the null hypothesis of “no difference” and instead answer the relevant questions, “How large is the effect?” or “To what extent are the data compatible with an effect as large as ...?”.

When planning the study, the investigator needs to identify the effect size index that will best answer the question. For example, if she is interested in comparing groups on a dichotomous outcome then risk difference, relative risk or odds ratio are examples of appropriate indices. If she is interested in comparing groups on a continuous outcome, then Cohen d (difference in means divided by the standard deviation of the sample) or one of several other similar standardized effect sizes metrics is appropriate. And finally, if she is interested in measuring the strength of a relationship or association between 2 or more variables, then correlation coefficient (or a variation) would be appropriate.

2. Precision of the Estimate (Confidence Intervals)
To fully interpret the results from a study, investigators need to provide confidence intervals (CIs) to quantify the precision of the estimate. The CI provides a range of plausible values for the effect size index. Investigators may select the desired level of confidence (eg, 95%, 99%), which is then used to interpret the likelihood that the true value of the effect is contained within the CI. For instance, there is a 5% risk that the 95% CI

Downloaded From: https://jamanetwork.com/ on 10/09/2023
around the observed effect will exclude the true value. The wider the CI, the less precise the effect size estimate and the less confident we can be of the results.

3. Inference About the Target Population
Confidence intervals are relevant and necessary whenever an investigator wishes to make an inference about the target population from the study sample.17 The correct interpretation of CIs when making an inference about the target population is the proportion of CIs from individual studies that would contain the true effect. For instance, it could be concluded that 95 CIs out of 100 would include the true parameter estimate. A more common interpretation is that with reasonable certainty (ie, 95% chance) the true effect lies within the upper and lower bounds of the 95% CI.

4. Determination of Whether the Results Are Compatible With a Clinically Meaningful Effect
When planning the study and interpreting the results, one of the requirements is a determination of how big of a difference is compatible with a clinically meaningful difference. The CIs determine whether the results are compatible with a clinically meaningful difference. For example, consider a patient education program developed to reduce hospital readmission after major head and neck cancer surgery. Before conducting the study, the investigators determined that a 10% reduction in readmission after introduction of the education program, as compared with before the program, would be a clinically meaningful reduction. After the study was completed, the investigators found that education program was associated with a 7% reduction in hospital readmission (P = .08) and the value of the upper bound of the 95% CI was 15% and the lower bound was −1%. The complete interpretation of the study finding is that the results of the education program are not only compatible with a clinically meaningful effect of 10% but could be as great as 15% based on the value of 15% for the upper bound of the CI. Furthermore, it is also true that the results are compatible with no reduction (CI includes 0) or even a 1% increase (lower bound of the 95% CI was −1%) in readmission among those participants who were exposed to the education program. Thus, the results of this study are inconclusive, but given the fact that the results are compatible with a clinically meaningful effect, further investigation of the education program is warranted. To conclude, based on the P value of .08 that the education program is ineffective, is incorrect.

5. Replication and Steady Accumulation of Knowledge
The results from one study must be interpreted with knowledge of similar research from the past and cannot be interpreted without context. Thus, results must be interpreted according to what others have found. Reporting results according to effect sizes and not P values allows for easy comparison of results across studies in the discussion section of the published article. Reporting effect sizes and CIs allows for meta-analytic thinking,10 which is the steady accumulation of knowledge from multiple studies. In JAMA Otolaryngology–Head & Neck Surgery, we look to publish original investigations where the investigators planned the study with sufficient sample size to have adequate power to detect a clinically meaningful effect and report the results with effect sizes and CIs. Authors should interpret the effect sizes in relation to previous research and use CIs to help determine whether the results are compatible with clinically meaningful effects. And finally, we acknowledge that no single study can define truth and that the advancement of medical knowledge and patient care depends on the steady accumulation of reliable clinical information.