In Praise of Confidence Intervals: Much More Informative Than *P* Values Alone

Michael R. Jiroutek, DrPH, MS;[1] J. Rick Turner, PhD, DSc[2]

From the Campbell University College of Pharmacy & Health Sciences, Buies Creek, NC;[1] and Quintiles, Durham, NC[2]

Modern statistical inference approaches were derived and published in the 1920s and 1930s by visionaries Sir Ronald Fisher, Jerzy Neyman, and Egon Pearson. While Fisher's and Neyman-Pearson's approach and solution to the problem differed, their methods have essentially blended over time into a single methodology that is used ubiquitously across science. As a result, the fundamental underpinnings on which hypothesis testing and confidence interval (CI) construction were built have remained in place for nearly a century. In multiple disciplines, a statistically significant *P* value obtained during hypothesis testing has for decades been deemed a validation of "successful" research. However, cautions in the use of hypothesis testing that rely solely on the interpretation of *P* values have increased in recent years.[1–4] In the specific context of developing new drugs or new drug combination therapies for hypertension (or any other disease or condition of clinical concern), attainment of a statistically significant treatment effect alone is not enough to declare success: compelling evidence of clinical significance is also required. Such evidence is facilitated by the employment of CIs. This Editorial elucidates the connection between hypothesis tests and CIs, and explains why the use of CIs in place of a hypothesis-testing approach alone is preferred and encouraged by the authors.

FUNDAMENTALS OF HYPOTHESIS TESTING

Imagine a phase III (therapeutic confirmatory) clinical trial testing a new antihypertensive drug against a placebo. A formalized hypothesis-testing approach is employed. A research question, research hypothesis, and null hypothesis are created. The research question is: Does the test drug reduce systolic blood pressure (SBP) to a statistically significantly greater degree than the placebo? The research hypothesis is: The test drug reduces mean SBP to a statistically significantly greater degree than the placebo. The null hypothesis is: The test drug does not statistically significantly reduce mean SBP as compared with the placebo.

Trial participants are randomized into one of the two treatment groups. SBP is measured at the beginning of the trial (baseline) and after 12 weeks of receiving one of the two treatments. The treatment effect, defined as mean SBP change among participants in the drug treatment group minus mean SBP change in the placebo treatment group, is calculated and is found to be 8 mm Hg. A formal statistical analysis is conducted, and a *P* value <.05 is obtained. This result means that a statistically significant result has been obtained, and, in accordance with the tenets of hypothesis testing, the null hypothesis is therefore rejected in favor of the research hypothesis (sometimes referred to as the alternate hypothesis, since the null hypothesis is the crux of hypothesis testing). The following statement can then be made: On the basis of this single trial, there is statistically significant evidence that the new drug reduces mean SBP as compared with the placebo.

In addition to compelling evidence that the drug is acceptably safe, such evidence is necessary for regulatory agencies to approve a new drug; however, it is not sufficient. There must also be compelling evidence of clinical significance, a requirement the employment of CIs can help address.

FUNDAMENTALS OF CIs

A single trial provides precise data from the (relatively small) group of participants who took part in the trial. However, this is not actually the main point of interest to us. What we really want to know is how well the result reflects what would likely be seen in the general population of patients with the disease or condition of clinical concern if the drug were to be approved and then prescribed for patients. Employment of CIs allows us to address this.

Consider the hypothetical trial just described. We place a CI around the treatment effect obtained in the trial (8 mm Hg), which is now referred to as the treatment effect point estimate. CIs facilitate quantification of the degree of confidence that is placed in this treatment effect point estimate. In this case, a two-sided CI will be calculated from the data collected in the trial. A lower limit will be placed at a certain distance below the treatment effect point estimate, and an upper limit will be placed at the same distance above it. This CI constitutes a range of values defined by the lower and upper limits of the CI. While various CIs can be chosen, a common one is the 95% CI. This CI is defined as the range of values that is likely to cover the true but unknown population treatment effect with a 95% degree of certainty.

Now consider the first of two hypothetical scenarios. The result obtained is written as follows:

- Treatment effect point estimate and two-sided 95% CI=8.0 (6.5–9.5)

Address for correspondence: J. Rick Turner, PhD, DSc, Cardiovascular Center of Excellence, Quintiles, 4820 Emperor Boulevard, Durham, NC 27703
E-mail: rick.turner@quintiles.com

DOI: 10.1111/jch.12908

This result allows us to make the following statement: The data obtained from this single clinical trial are compatible with a treatment effect in the general population as small as 6.5 mm Hg and as large as 9.5 mm Hg, and our best estimate is 8.0 mm Hg.

Clinical judgment must now be employed, and the focus falls on the lower limit of the CI, which represents the "worst-case scenario." The question of interest is therefore this: Is a decrease of 6.5 mm Hg in SBP clinically significant? Let us presume that the clinical scientists involved in the trial decide that the answer is yes, and that regulatory agencies would agree. In this case, compelling evidence of clinical significance has been provided.

Now consider a second hypothetical scenario. The result obtained is written as follows:

- Treatment effect point estimate and two-sided 95% CI=8.0 (2.0–14.0)

This result allows us to make the following statement: The data obtained from this single clinical trial are compatible with a treatment effect in the general population as small as 2.0 mm Hg and as large as 14.0 mm Hg, and our best estimate is 8.0 mm Hg.

Again, clinical judgment is now brought to bear, and the following question is asked: Is a decrease of 2.0 mm Hg in SBP clinically significant? Answering this question is likely to be a tougher call. Any degree of SBP reduction is theoretically desirable. However, there are already other drugs on the market that lower SBP to a greater extent, and therefore even if the clinical scientists involved in the trial decided that the answer is yes, a second question arises: Would regulators agree and hence be likely to approve this drug if its safety profile is acceptable? Also, if it were approved and prescribed, it may be difficult to detect such a decrease in clinical practice. So, let's presume that the clinical scientists decide that this lower estimate of efficacy is not clinically significant. In that case, the company developing the drug may decide to run a trial using a higher dose of the drug, or to terminate the drug's development program and focus resources elsewhere.

RELATIONSHIP BETWEEN CIs AND STATISTICAL SIGNIFICANCE

A salient attribute of CIs, therefore, is that they facilitate judgments of clinical significance. It is absolutely possible to obtain a result that is statistically significant but not clinically significant. Gardner and Altman[5] commented that "presenting *p*-values alone can lead to them being given more merit than they deserve. In particular, there is a tendency to equate statistical significance with medical importance or biological relevance." Biological relevance and therefore medical importance are much better represented in terms of clinical significance.[6]

That said, attainment of statistical significance remains a necessary component of the statistical approach needed to obtain approval of a new drug, a statement that brings us to a second salient attribute of CIs. CIs are intimately related to probability levels since statistical significance can be deduced from the values of the limits of two-sided CIs. In the present context, if the lower limit of a 95% CI is greater than zero, by definition the upper limit will also be (even) greater than zero. Therefore, the CI excludes zero, and hence the treatment effect is deemed statistically significant at the 0.05 alpha level, ie, $P<.05$. (For the sake of completeness, we should also state explicitly that, when the limits of a CI lie on either side of zero, ie, the interval contains the value zero, statistical significance is not attained.)

Now consider our first hypothetical scenario again. It was represented by this result:

- Treatment effect point estimate and two-sided 95% CI=8.0 (6.5–9.5)

Since 6.5 lies above zero, as does 9.5, the result is statistically significant at the 0.05 alpha level. That is, based on this single trial, there is statistically significant evidence that the new drug lowers average SBP more than placebo. Now consider our second hypothetical scenario again. It was represented by this result:

- Treatment effect point estimate and two-sided 95% CI=8.0 (2.0–14.0)

Using slightly different but equivalent wording, since both of the limits lie above zero, ie, zero is not included in the interval, the result is statistically significant at the 0.05 alpha level. That is, based on this single trial, there is statistically significant evidence that the new drug lowers average SBP more than placebo.

IMPLICATIONS AND IMPORTANCE OF THESE EXAMPLES

Both scenarios just presented are therefore representative of results attaining statistical significance. However, and of considerable importance, the first result also attained clinical significance: this was not the case for the second result. This highlights an immediate advantage of CIs as compared with presenting results purely describing the degree of statistical significance attained. As noted previously, the biological relevance of a clinical trial's results, and therefore medical importance, are much better represented in terms of clinical significance: the use of CIs addresses both statistical and clinical significance.

As a final comment, it should be acknowledged that while the lower and upper limits of CIs represent the attainment or not of statistical significance, ie, in our examples they show whether or not the result is significant at the 0.05 alpha level, they do not provide the precise *P* value. It could be 0.04 or 0.03, for example. That said, the further the lower limit lies above zero, the smaller the *P* value will be, and hence the greater the degree of statistical significance attained. It can reasonably be argued that the magnitude of the *P* value itself provides useful information. Satisfying this argument, however, is very easily achieved: the actual *P* value obtained from the hypothesis-testing component of the overall statistical approach can be presented along with the CIs. Such a hypothetical result could be captured as follows:

- Treatment effect point estimate and two-sided 95% CI=8.0 (6.5–9.5), *P*=.02.

We recommend such complete reporting for all clinical trials.

Disclosures: The authors report no specific funding in relation to the preparation of this paper. No editorial support was used.

References

1. Hayat MJ. Understanding statistical significance. *Nurs Res*. 2009;59: 219–223.
2. Sullivan GM, Feinn R. Using effect size—or why the P value is not enough. *J Grad Med Educ*. 2012;4:279–282.
3. Chavalarias D, Wallach JD, Li AH, Ioannidis JP. Evolution of reporting P values in the biomedical literature, 1990-2015. *JAMA*. 2016;315:1141–1148.
4. Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. *The American Statistician*. 2016;70:129–133.
5. Gardner MF, Altman DG. Estimation rather than hypothesis testing: confidence intervals rather than *p*-values. In: Gardner MF, Altman DG, eds. *Statistics with Confidence*. London: British Medical Association; 1986.
6. Turner JR, Karnad DK, Kothari S. *Cardiovascular Safety in New Drug Development and Therapeutic Use: New Methodologies and Evolving Regulatory Landscapes*. Springer International Publishing: Gewerbestrasse, Switzerland; 2016.