What Is Confidence? Part 1: The Use and Interpretation of Confidence Intervals

From the Department of Emergency Medicine, Harbor-UCLA Medical Center, Torrance, CA; and the UCLA School of Medicine, Los Angeles, CA.

Received for publication September 3, 1996. Revisions received November 7, 1996, and January 10 and April 23, 1997. Accepted for publication April 28, 1997.

Kelly D Young, MD
Roger J Lewis, MD, PhD

Hypothesis testing and the *P* value it generates are overemphasized in statistical analyses published in medical journals. An alternative, the confidence interval (CI), offers significantly more information to readers interpreting results. There have been many authoritative calls for the report of CIs in place of *P* values,[1-7] such as that of the International Committee of Medical Journal Editors, whose guidelines for statistical reporting give the following instructions: "When possible, quantify findings and present them with appropriate indicators of measurement error or uncertainty (such as confidence intervals)," and "Avoid sole reliance on statistical hypothesis testing, such as the use of P values, which fails to convey important quantitative information."[8] In this article, part 1, we provide an overview of CIs for the clinician reading the medical literature. We describe the advantages of CIs and explain and illustrate their proper interpretation. In part 2, which follows this article, we provide added information important for clinical researchers, including a precise definition of CIs, a compact reference to methods for calculating CIs in common situations, and an explanation of the difference between CIs and probability intervals.[9]

[Young KD, Lewis RJ: What is confidence? Part 1: The use and interpretation of confidence intervals. *Ann Emerg Med* September 1997;30:307-310.]

WHY SHOULD WE USE CONFIDENCE INTERVALS?

Suppose we wish to test whether one vasopressor is better than another at raising the mean systolic blood pressure (SBP) in hypotensive patients. Using traditional hypothesis testing, we begin by posing a null hypothesis—for example, that the mean SBPs are the same in groups of hypotensive patients randomized to receive the two vasopressors. The null hypothesis is the hypothesis that no difference exists between the groups. In other words, if our null hypothesis is true, the difference between the two groups' mean SBPs is zero.

Next we conduct the trial. Suppose in our trial we observe a mean SBP for patients given vasopressor A of 70 mm Hg and for patients given vasopressor B of 95 mm Hg. Our observed treatment difference (mean SBP for patients on vasopressor B minus mean SBP for patients on vasopressor A) is 25 mm Hg. We then calculate the probability of observing the difference in mean SBP that we saw, or a greater difference, assuming there is really no treatment difference. This probability is the *P* value.

If the *P* value is less than the traditionally accepted value of .05, we reject the null hypothesis as false and we conclude that our study demonstrates a statistically significant difference in mean SBP between the groups. That is, if the null hypothesis was true, the probability of obtaining the difference in blood pressures actually seen in our trial, or a bigger difference, by chance alone (random error), is less than .05. This 5% probability "cutoff" is considered small enough to justify the conclusion that the observed difference was not due to chance alone. That the *P* value is less than .05 tells us only that the treatment difference that we observed is statistically significantly different from zero. It does not tell us the size of the treatment difference, which determines whether the difference is clinically important, or how precisely our trial was able to estimate the true treatment difference. The observed treatment difference is that seen in our study sample of hypotensive patients. The true treatment difference is the difference that would be observed if all similar hypotensive patients could be included in the study.

Studies comparing two groups often yield a single number, such as the difference in mean SBP. This single number "estimates" the true difference between the groups and is termed a "point estimate." If, instead of using hypothesis testing and reporting a *P* value, we report the point estimate and the corresponding CI surrounding it, we give readers the same information as the *P* value, plus information on the size of the treatment difference (and therefore its clinical importance), the precision of the estimated difference, and information to aid interpretation of a negative result.

The *P* value answers only the question, "Is there a statistically significant difference between the two treatments?" The point estimate and its CI also answer the questions, "What is the size of that treatment difference?", and "How precisely did this trial determine or estimate the true treatment difference?" As clinicians, we should change our practice only if we believe the study has definitively demonstrated a treatment difference and that the treatment difference is large enough to be clinically important.

Additionally, even if a trial does not show a statistically significant difference and we accept the null hypothesis, the CI enables us to distinguish whether there really is no difference between the treatments, or the trial simply did not have enough patients to reliably demonstrate a difference. If there is no difference, we can discard the least cost-effective or least well-tolerated treatment. If the sample population is inadequate in size, further study is warranted to determine whether one treatment has some benefit over the other. Because so much more information can be obtained from CIs and they have greater value to the reader, they should usually be reported in place of *P* values.

WHAT IS A CONFIDENCE INTERVAL?

A CI may be viewed as the range of possible values for the true treatment difference that are statistically likely, given the results of a particular trial. "Statistically likely" is defined similarly to the 5% probability cutoff used in hypothesis testing.

In our hypothetical trial comparing the ability of two vasopressors to increase SBP in hypotensive patients, we obtained a point estimate for the true treatment difference of 25 mm Hg. Now suppose we calculate the 95% CI around this point estimate as 5 to 44 mm Hg. What does this mean? The 95% CI of 5 to 44 mm Hg implies that, if the true difference is 5 mm Hg, the probability of seeing a treatment difference of 25 mm Hg or greater is .025 and that, if the true difference is 44 mm Hg, the probability of seeing a treatment difference of 25 mm Hg or less is .025. Analogous to the traditionally accepted *P* value of .05, this probability (2×.025) is considered small enough to conclude that the true treatment difference is unlikely to be less than 5 mm Hg or greater than 44 mm Hg. We therefore interpret the results of the trial as showing that the true treatment difference is most likely between 5 and 44 mm Hg.

The true treatment difference is the difference that would be observed if all similar hypotensive patients could be included in the study. The CI is the range of possible values for the true treatment difference that are statistically consistent with the point estimate obtained from the trial. The point estimate of 25 mm Hg we observed in our trial is our "best estimate" for the true treatment difference based on this trial. The CI of 5 to 44 mm Hg tells us our results are statistically consistent with a true difference anywhere in the range of 5 to 44 mm Hg. That is, our data suggest that vasopressor B increases the mean SBP of hypotensive patients anywhere from 5 to 44 mm Hg more than vasopressor A.

In the past it was commonly taught that the 95% CI spans the values 2 SDs above and below the observed normally distributed mean. This approach to visualizing the CI is applicable only to continuous data and accurate only for large samples that approximate normal distribution.

HOW TO INTERPRET A CONFIDENCE INTERVAL

Suppose in another trial it is found that a new nebulized medication decreases the rate of admission to the hospital for asthmatic patients by 15%, with a 95% CI of 2% to 28%. This decrease in the rate of admission would be statistically significant ($P<.05$). If the medication truly decreased admission rate by 15%, we would be likely to change our practice and give the medication to our patients. However, the CI tells us that the data from the trial are also consistent with the possibility that admission rates are decreased by only 2%. If the medicine is costly, we would probably not change our practice in this case. On the other hand, the trial results are also consistent with a decrease in admission rate of as much as 28%. Definitive conclusions regarding the appropriate change in practice cannot be made, and further study of the medication is needed. It is important to remember that given the data observed in a trial, any value along the entire range of the CI is plausible.

Returning to our first clinical example, a treatment difference of 0 is equivalent to the null hypothesis that there is no difference in mean SBP between patients given vasopressor A and patients given vasopressor B. In our trial, the CI of 5 to 44 mm Hg does not include 0; therefore a true treatment difference of zero is not statistically consistent with our data. We conclude that the null hypothesis that there is no difference is not statistically consistent with our observed data, and we reject the null hypothesis. The results are therefore statistically significant, equivalent to a *P* value less than .05. When a 95% CI does not include a zero treatment difference, this demonstrates that the results are statistically significant, equivalent to a *P* value less than .05. Therefore the presence or absence of a zero treatment difference in a 95% CI gives the same information as a statement that *P* is greater or less than .05.

Our point estimate of 25 mm Hg gives an estimate for the size of the treatment difference. However, our results are also statistically consistent with any value within the range of the CI of 5 to 44 mm Hg. In other words, the true treatment difference may be as little as 5 mm Hg, or as much as 44 mm Hg. If vasopressor B has many more severe side effects than vasopressor A, a reader may conclude that even an increase of SBP as much as 44 mm Hg does not warrant the use of vasopressor B, although the treatment difference is statistically significant. Another reader may believe that even an increase in mean SBP of 5 mm Hg would be beneficial, despite the side effects. With *P* values, authors report results as statistically significant or not, leaving little basis on which to draw conclusions relevant to clinical practice. With CIs we may decide what treatment difference we consider clinically important and reach conclusions appropriate for our practice.

We may also use CIs to obtain important information from trials with results that were not statistically significant (so-called negative trials). Suppose we found the 95% CI for the difference in mean SBP to be –5 mm Hg to 55 mm Hg, with the same point estimate of 25 mm Hg. Now our results are consistent with vasopressor B increasing mean SBP as much as 55 mm Hg more than vasopressor A, or as much as 5 mm Hg less. Because the CI includes 0 (a zero treatment difference), equivalent to the null hypothesis that no treatment difference exists, the results are not statistically significant, and *P* is greater than .05. Because *P* is greater than .05, we may be tempted to conclude that there is no advantage to using vasopressor A or B in clinical practice. However, our data are also consistent with vasopressor B increasing SBP as much as 55 mm Hg more than vasopressor A. Although *P* is greater than .05, there remains the possibility that a clinically important difference exists in the two vasopressors' effects on mean SBP. Negative trials with results consistent with a clinically important difference usually occur when sample size is too small, resulting in low power to detect an important treatment difference.[10,11]

It is important to know how precisely the point estimate represents the true difference between the groups. The width of the CI gives us information on the precision of the point estimate. The larger the sample size, the more precise the point estimate, and the CI will be narrower. As mentioned earlier, negative trials with too small a sample often do not show a statistically significant result but still do not rule out a clinically important treatment difference.[10,11] In this case, the CI is wide and imprecise and includes zero or no treatment difference (or both) and clinically important treatment differences. Conversely, positive trials with large samples may show a statistically significant treatment difference that is not clinically important—for example, an increase in mean SBP from 70 to 72 mm Hg.

If a CI includes both zero and clinically important treatment differences, we can make no definitive conclusions about clinical practice, despite the lack of statistical significance. Similarly, if a CI excludes zero but includes clinically trivial differences, we cannot draw definitive conclusions about clinical practice, despite a significant *P* value. It is important to remember that the data are statistically consistent with the entire range of the CI from a trial.

SUMMARY

In most cases, CIs should be reported in place of *P* values. A point estimate and the CI surrounding it give information on the size of the treatment difference observed, its statistical significance, and the likely range of possible true treatment differences and permit the determination of clinical importance. If the CI includes clinically important values, it cannot be concluded that a potential benefit has been unequivocally ruled out. Conversely, if the CI includes clinically unimportant values, it cannot be concluded that a beneficial effect has been unequivocally established.

REFERENCES

1. Gardner MJ, Altman DG: Confidence intervals rather than P values: Estimation rather than hypothesis testing. *BMJ* 1986;292:746-750.

2. Simon R: Confidence intervals for reporting results of clinical trials. *Ann Intern Med* 1986;105:429-435.

3. Berry G: Statistical significance and confidence intervals. *Med J Aust* 1986;144:618-619.

4. Braitman LE: Confidence intervals extract clinically useful information from data. *Ann Intern Med* 1988;108:296-298.

5. Pocock SJ, Hughes MD: Estimation issues in clinical trials and overviews. *Stat Med* 1990;9:657-671.

6. Braitman LE: Confidence intervals assess both clinical significance and statistical significance. *Ann Intern Med* 1991;114:515-517.

7. Borenstein M: The case for confidence intervals in controlled clinical trials. *Control Clin Trials* 1994;15:411-428.

8. Bailar JC, Mosteller F: Guidelines for statistical reporting in articles for medical journals. *Ann Intern Med* 1988;108:266- 273.

9. Young KD, Lewis RJ: What is confidence? Part II: Detailed definition and determination of confidence intervals. *Ann Emerg Med* 1997;30:311-318.

10. Freiman JA, Chalmers TC, Smith H, et al: The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial: Survey of 71 "negative" trials. *N Engl J Med* 1978;299:690-694.

11. Brown CG, Kelen GD, Ashton JJ, et al: The beta error and sample size determination in clinical trials in emergency medicine. *Ann Emerg Med* 1987;16:183-187.

Reprint no. 47/1/84040

Address for reprints:

Roger J Lewis, MD, PhD
Harbor-UCLA Medical Center
Department of Emergency Medicine
1000 West Carson Street, Box 21
Torrance, CA 90509
310-222-6741
Fax 310-782-1763
E-mail rlewis@humc.edu