CHEST®

Official publication of the American College of Chest Physicians

# Do CIs Give You Confidence?

Geoffrey R. Norman and David L. Streiner

*Chest* 2012;141;17-19
DOI 10.1378/chest.11-2193


ISSN:0012-3692

CHEST

Commentary

STATISTICAL QUANDARIES

# Do CIs Give You Confidence?

*Geoffrey R. Norman, PhD; and David L. Streiner, PhD*

**This article describes the conceptual basis for the *P* value and the CI. We show that both are derived from the same underlying concepts and provide useful, but similar information.**

***CHEST 2012; 141(1):17–19***

One of the more confusing issues about the use of statistics is the use of CIs. Although almost everyone has seen lots of examples of "$P < .05$," and some may even be able to accurately describe what it means, the contribution of the CI to enlightenment is much less clear.

Strangely, if you believed some of the rhetoric offered by the advocates of CIs, they appear to be the answer to everyone's dreams. It would seem that anyone who does not quickly undergo a religious conversion to CIs and move to banish "$P < .05$" from his home and work is clearly a presenile proto-Victorian. Sixteen years ago, Stephen Walter[1] wrote a wonderful summary of the issues, pointing out that editors of many mainstream journals, including *BMJ*, *The Lancet*, *Annals of Internal Medicine*, and *CMAJ*, had all endorsed CIs as a viable alternative to *P* values. However, perusal of recent issues of any of these journals reveals that the projected demise of *P* values was a bit premature. In any case, the proponents of CIs appear to be balanced by as many naysayers who remain in the *P* value camp. Indeed, Walter's article was accompanied by a brief note from the editor of *American Journal of Epidemiology* stating that "the diversity of opinion among editorial board members makes it difficult, if not impossible, to arrive at a consensus which could be translated into a *Journal* policy."[1] Peace in the Middle East looks simple by comparison.

Manuscript received August 30, 2011; revision accepted September 1, 2011.

**Affiliations:** From the Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON, Canada.

**Correspondence to:** Geoffrey R. Norman, PhD, McMaster University, 1280 Main St W, MDCL 3519, Hamilton, ON, L8N3Z5, Canada; e-mail: norman@mcmaster.ca



**DOI: 10.1378/chest.11-2193**

So what is the fuss about? What does a CI give us that is different from a *P* value? Are there circumstances where we should use one or the other? To address these questions, regrettably, we must go back to the basics and be very clear on definitions of these terms.

## The Ubiquitous *P* Value

The use of *P* values has a long history, dating back to seminal work by Fisher and Pearson in the 1920s. In its simplest form comparing two groups, the *P* value is "the probability of observing a result at least as extreme as the observed result if in truth there is no difference between the groups."[2] A small *P* value ($P < .0001$) says that it is unlikely that a difference this large could arise by chance (< one chance in a thousand), so it is likely a real effect.

It is critical to keep in mind that, in this simple example, the *P* value depends on three quantities: the difference between the means (the bigger the difference, the smaller the *P* value), the SD of the individual observations (the smaller the SD, the smaller the *P* value), and the sample size (the bigger the sample size, the smaller the *P* value). In short, big differences are "more significant" (ie, a smaller *P* value) than little differences (and "more significant" is placed in quotation marks for a very good reason, stay tuned), and small differences arising from large samples are "more significant" than small differences from small samples. The *P* value, then, is a way to separate real effects from effects due to random fluctuations in the data and sampling error. As such, that's not a bad thing. After all, it is the nature of the world that when you divide people into two groups, using a coin flip or any other strategy, the groups will never come out exactly the same on any measure (height, weight, BMI, or anything else), so without some help from our statistician friends, we would be unable to tell

differences arising from random variation from "real" differences.

But that is all we get from it. In particular, it does not tell us *anything* directly about how big the effect is. Of course, big effects will likely have smaller *P* values than small effects, but the whole thing is confounded with sample size. As Dave Sackett once said in a faculty meeting, "Too small a sample and you can prove nothing. Too large and you can prove anything." The trouble is that the *P* value has become deified over the years, so far too often we see abstracts with statements like, "The difference between the groups was highly significant ($P < .0001$)." So it is not zero. You are really, really sure it is not zero. But that is all a "highly significant" *P* value is telling you. It is also not telling you that a study that yielded a *P* level of .001 has more meaningful results or a bigger effect size than a study with a *P* level of .05. Again, it is dependent in part on the sample size.

Worst of all, relying on *P* levels to reveal truth leads to dichotomous thinking. If $P = .051$, then we dismiss the result as nonsignificant, but if $P = .050$, then it suddenly emerges as "real" (and, more important, as publishable). But probability is a continuum, and to dichotomize it at the magical .050 is artificial. Is there really a difference between .051 and .050? We do not think so either.

The problem with *P* values in part comes down to words. Somehow "highly significant" sounds a lot like "highly important." Perhaps we should change the terminology to be more descriptive, along the lines of:

$P = .05$: Quite unlikely to be zero.
$P = .01$: Really unlikely to be zero.
$P = .0001$: Really, *really*, ***really*** unlikely to be zero.

But the other problem is that a significant *P* value tells you what the difference is not (zero), but does not tell you what it *is*. And that is where CIs come in. But to understand what they do and do not do, we need a bit more theory.

The *P* value, in the simple case we have been considering, comes about from imagining a normal distribution centered on a difference of zero, with an SD equal to the SEM, which is $\sigma/\sqrt{n}$, where $\sigma$ is the SD and $n$ is the sample size. We superimpose the observed difference on this distribution, and if it is far enough away from zero, we say it is significant. Specifically, if it is greater than $1.96 \times$ SEM, then the area in the tail is 0.05, so any difference $> 1.96$ SEM is significant. All this is shown in Figure 1.

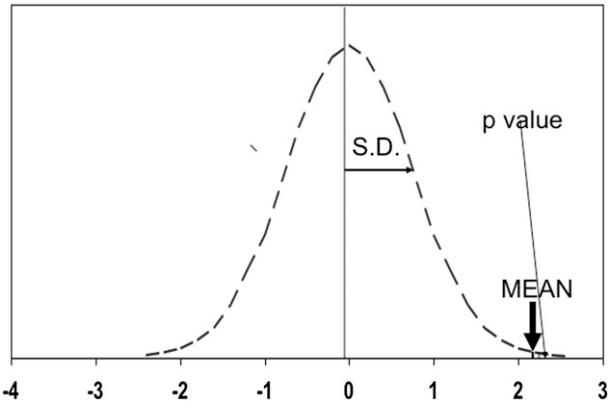


FIGURE 1. *P* value as the area of the Ho distribution to the right of the sample mean.

## The Less Ubiquitous CI

CIs turn this whole thing on its head. We begin with the observed difference, call it "*d*." We superimpose a normal distribution centered on *d*, then take the same 1.96 SEM and add and subtract it from *d*. The upper and lower limits define the 95% CI, where we can be confident that the true difference lies within these boundaries 95% of the time. So a narrow CI indicates that the difference has been estimated with a high degree of precision, and a wide CI says that there is much more uncertainty in the actual magnitude of the difference. This is shown in Figure 2.

Expressing it this way has the added advantage that you can infer statistical significance. If the 95% CI overlaps zero, then we know that the difference is not significant, since this means there is a $>5\%$ probability that the difference is zero or negative. If it doesn't overlap zero, then the result is significant at the .05 level. The big advantage of the CI, then, is that it focuses on the magnitude of the difference and gets away from the silliness of .05, .001, and .00001, although you can still determine whether the difference is significant.

It is well to keep in mind that, while the CI does give you a better picture of how large the difference

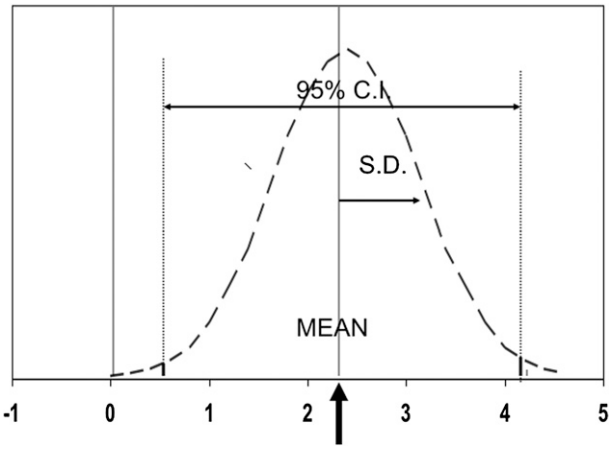


FIGURE 2. CI as the symmetrical interval around the sample mean containing 95% of possible mean values.

is, it is still inappropriate to assume that smaller CIs mean bigger effects. Like the *P* value, it depends on sample size. To properly look at the clinical importance of effects, you need to go to something like an effect size or an OR.

Are there any other downsides? Of course, there is no free ride. The CI works well if you are talking about the difference between two means. But as you get to more complex designs, it gets increasingly clumsy. If you go to a simple $2 \times 2$ table and compute an OR or relative risk, the CI can be calculated, but it is no longer symmetrical. But when you get to multiple comparisons, such as a one-way analysis of variance with four groups (eg, placebo, drug A, drug B, drug C, and mean $FEV_1$), an overall F test and *P* value is simple to calculate. But we would need a CI for $(4 \times 3)/2 = 6$ pairwise comparisons, and then we have to worry about multiple tests.[3]

## Putting It Into Practice

It should be evident by now that despite the rhetoric, it really is not the case that one is holy and the other is evil. They really are alternate routes to expressing very similar ideas, and as a consequence, you cannot really say that the CI should be used under some circumstances and the *P* value under others. Indeed, one article we came across in *CHEST* (and we are quite sure it is not unique) had the best of both worlds:

> Independent factors associated with an accelerated decline of lung function were chronic colonization with Pseudomonas aeruginosa (PA) [odds ratio (OR), 30.4; 95% confidence interval (CI), 3.8 to 39.4; p = 0.005], more frequent severe exacerbations (OR, 6.9; 95% CI, 2.3 to 10.5; p = 0.014), and more systemic inflammation (OR, 3.1; 95% CI, 1.9 to 8.9; p = 0.023).[4]

This makes it clear that they are variations on a theme: One can be derived from the other. As you can see, a relatively large *P* value like .023 corresponds to a lower limit of the CI of 1.9, fairly close to the null hypothesis (which is 1.0 since these are ORs); conversely, the highly significant *P* value of .005 has a CI with a lower limit of 3.8. Nevertheless, they provide different perspectives on the data, so both are useful.


## Acknowledgments

**Financial/nonfinancial disclosures:** The authors have reported to *CHEST* that no potential conflicts of interest exist with any companies/organizations whose products or services may be discussed in this article.


## References


1. Walter SD. Methods of reporting statistical results from medical research studies. *Am J Epidemiol*. 1995;141(10):896-906.
2. Altman DG. Why we need confidence intervals. *World J Surg*. 2005;29(5):554-556.
3. Streiner DL, Norman GR. Correction for multiple testing: is there a resolution? *Chest*. 2011;140(1):16-18.
4. Martínez-García MA, Soler-Cataluña JJ, Perpiñá-Tordera M, Román-Sánchez P, Soriano J. Factors associated with lung function decline in adult patients with stable non-cystic fibrosis bronchiectasis. *Chest*. 2007;132(5):1565-1572.