# Why the P-value culture is bad and confidence intervals a better alternative.

Ranstam, Jonas

*Published in:*
Osteoarthritis and Cartilage

*DOI:*
10.1016/j.joca.2012.04.001

2012

April 1, 2012

# Why the p-value culture is bad and confidence intervals a better alternative

Jonas Ranstam

Department of Orthopedics, Clinical Sciences Lund, Lund University,
SE-22185 Lund, Sweden. Email: jonas.ranstam@med.lu.se.

## Abstract

In spite of frequent discussions of misuse and misunderstanding of P-values they still appear in most scientific publications, and the disadvantages of erroneous and simplistic p-value interpretations grow with the number of scientific publications. Osteoarthritis and Cartilage prefer confidence intervals. This is a brief discussion of problems surrounding p-values and confidence intervals.

## Abbreviations

| | | |
|---|---|---|
| P-value or p | = | Probability value |
| d | = | An observed difference, e.g. between exposed and unexposed patient groups |
| t | = | A quantity having a t-distribution |
| df | = | Degrees of freedom |
| n | = | Number of observations |
| SD | = | Standard Deviation |
| SE | = | Standard Error |
| FAR | = | Floating Absolute Risks |

P-values seem to be the solid foundation on which scientific progress relies. They appear in almost every epidemiological, clinical, and pre-clinical research publication, either as precise decimal numbers, inequalities (p>0.05 and p<0.05) or as symbols (***, **, *, and NS). Several scientific arguments criticizing this p-value culture have been published (1). This criticism can, in fact, be traced as far back as to 1933 (2). Attempts to demolish the culture have usually been futile (3), and the problems of the p-value culture are growing with the increasing number of scientific publications. Osteoarthritis and Cartilage recommends presenting sampling uncertainty in the form of confidence intervals. This is a brief presentation of the weaknesses of p-values and strengths of confidence intervals.

First, the aim of a scientific study or experiment is wider than just to observe, because it is required of scientific results that they can be generalized to other patients or cells than only those examined or experimented on. One difference between quantitative scientific research and other forms of investigations is that the research work includes quantification of the uncertainty of the results.

The principle behind the uncertainty evaluation is to consider the studied patients, or cells, as a random sample from an infinite population of patients, or cells. Statistical methods that assess the sampling uncertainty have been the foundation for quantitative medical research (4) since the end of the second world war. The resulting p-values and confidence intervals contain information on the sampling uncertainty of a finding, which influences the generalizability of the results of the individual experiment study.

It is important to understand that these measures of generalization uncertainty have no relevance for the studied sample itself, i.e. the studied groups of patients, animals or cells from which the generalization is made. P-values and confidence intervals guide us in the uncertainty of whether an observed difference is a random phenomenon, appearing just in the studied sample, or if it represents a true difference in the entire (unobserved) population, from which the sample has been drawn and can be expected to be a reproducible finding. The statistical precision section below describes how the uncertainty can be quantified.

The current tradition in medical research of screening variables with hypothesis tests to categorize findings either as statistically significant or insignificant is a simplistic and counterproductive analysis strategy that should be abandoned. This brief editorial attempts to explain why.

## Statistical precision

Statistical precision has two determinants, the number of observations in the sample and the observations' variability. These determinants specify the standard error (SE) of an estimate such as the mean:

$$SE = SD/\sqrt{n}$$

where SD stands for standard deviation, and n is the number of observations. Less variability and more observations reduce the SE and increase the statistical precision.

When comparing the difference between two mean values, for example to estimate the effect of the exposure to a specific agent by comparing exposed with unexposed patient groups, the statistical precision in the mean value difference, *d*, which also is an estimate of the effect from the exposure, can be written:

$$SE = \sqrt{(SD^2/n_1 + SD^2/n_2)}$$

Where SD is the standard deviation common for both groups and $n_1$ and $n_2$ represent the number of independent observations in each group.

Both the p-value and the confidence intervals are based on the SE. When the studied difference, *d*, has a Gaussian distribution it is statistically significant at the 5% level when

$$|d/SE| > t_{0.05}$$

Here $t_{0.05}$ is the value in the Student's t-distribution (introduced in 1908 by William Gosset under the pseudonym Student) that discriminates between the 95% $|d/SE|$ having lower values and the 5% that have higher. Conversely, the confidence interval

$$d \pm t_{0.05}\text{SE}$$

describes a range of plausible values in which the real effect is 95% likely to be included.

## P-values

A p-value is the outcome from a hypothesis test of the null hypothesis, $H_0$: $d = 0$. A low p-value indicates that observed data do not match the null hypothesis, and when the p-value is lower than the specified significance level (usually 5%) the null hypothesis is rejected, and the finding is considered statistically significant. The p-value has many weaknesses that need to be recognized in a successful analysis strategy.

First, the tested hypothesis should be defined before inspecting data. The p-value is not easily interpretable when the tested hypothesis is defined after data dredging, when a statistically significant outcome has been observed. If undisclosed to the reader of a scientific report, such post-hoc testing is considered scientific misconduct (5).

Second, when multiple independent hypotheses are tested, which usually is the case in a study or experiment, the risk that at least one of these tests will be false positive increases, above the nominal significance level, with the number of hypotheses tested. This multiplicity effect reduces the value of a statistically significant finding. Methods to adjust the overall significance level (like Bonferroni adjustment) exist, but the cost of such adjustments is high. Either the number of observations has to be increased to compensate for the adjustment, or the significance level is maintained at the expense of the statistical power to detect an existing effect or difference.

Third, a statistically insignificant difference between two observed groups (*the sample)* does not indicate that this effect does not exist in the *population* from which the sample is taken, because the p-value is confounded by the number of observations; it is based on the SE, which has $\sqrt{n}$ in the denominator. A statistically insignificant outcome indicates nothing more than that the observed sample is too small to detect a population effect. A statistically insignificant outcome should be interpreted as "absence of evidence, not evidence of absence" (6).

Fourth, for the same reason a statistically significant effect in a large sample can represent a real, but minute, clinically insignificant, effect. For example, with sufficiently large sample size even a painkiller reducing pain with as little as an average of 1 mm VAS on a 100 mm scale will eventually demonstrate a highly statistically significant pain reduction. Any consideration of what constitutes the lowest clinically significant effect on pain would be independent of sample size, perhaps depend on cost, and possibly be related to the risk of side effects and availability of alternative therapies.

Fifth, a p-value provides only uncertainty information vis-a-vis a specific null hypothesis, no information on the statistical precision of an estimate. This means that comparisons with a lowest clinically significant effect (which may not be definable in laboratory experiments) cannot be based on p-values from conventional hypothesis test. For example, a statistically significant relative risk of 2.1 observed in a sample can correspond to a relative risk of 1.1, as well as to one of 10.0, in the population. The statistical significance comes from the comparison with the null hypothesis relative risk of 1.0. That one risk factor in the sample has lower p-value than another one says nothing about their relative effect.

Sixth, when the tested null hypothesis is meaningless the p-value will not be meaningful. For example, inter-observer reliability is often presented with a p-value, but the null hypothesis in this hypothesis test is that no inter-observer reliability exists. However, why should two observers observing the same object come to completely independent results? This is not a meaningful hypothesis to test using p-values. Showing the range of plausible values of the inter-observer reliability in the population is much more relevant.

## Confidence intervals

Confidence intervals share some of the p-value's weaknesses, like the multiplicity problem, and analogous with the adjustment of the significance level, the width of confidence intervals can also be adjusted in cases of multiplicity. However, the great advantage with confidence intervals is that they do show what effects are likely to exist in the *population*. Values excluded from the confidence interval are thus not likely to exist in the population. Consequently, a confidence interval excluding a specific effect can be interpreted as providing evidence against the existence (in the unobserved population) of such an effect. The confidence interval limits do thereby allow an easy and direct evaluation of clinical significance, see Figure 1.

Confidence interval limits are important criteria in the evaluation of relative treatment effects in equivalence and non-inferiority clinical trials, the trial designs used for testing if a new drug at least is as good as an old one. The reasons for preferring the new drug could be fewer side effects, lower cost, etc.

The margin of non-inferiority or equivalence introduces here the notion of clinical significance into randomized trial comparisons of treatment effect. By defining what is a clinically significant difference in treatment effect it becomes possible to evaluate non-inferiority, see Figure 2. It is thus not sufficient to show statistical insignificance (again this indicates "absence of evidence, not evidence of absence"), it is necessary to show clinical insignificance with a confidence interval narrow enough to exclude clinically significant effects (as this shows evidence of absence).

The advantages of using confidence intervals instead of p-values has been frequently discussed in the literature (1). In spite of this, confidence intervals are often misunderstood as representing variability of observations instead of uncertainty of the sample estimate. Some further common misunderstandings should be mentioned.

A consequence of the dominant p-value culture is that confidence intervals are often not appreciated by themselves, but the information they convey are transformed into simplistic terms of statistical significance. For example, it is common to check if the confidence intervals of two mean values overlap. When this happens, the difference of the mean values is often considered statistically insignificant. However, Student's t-test has a different definition of the mean difference standard error than what is used in the calculation of the overlapping confidence intervals. Two means may well be statistically significantly different and still have somewhat overlapping confidence intervals. Overlapping confidence intervals can therefore not be directly interpreted in terms of statistical significance (7).

Standard errors are also often used to indicate uncertainty, as error bars in graphical presentations. Using confidence intervals is, however, a better alternative because the uncertainty represented by a standard error is confounded by the number of observations (8). For example, one standard error corresponds to a 58% confidence interval when n is 3 and to a 65% confidence interval when n=9.

When pairwise multiple groups are compared with one and the same reference or control group in terms of relative risk or odds ratios, comparisons of confidence intervals are only valid vis-a-vis the reference group. However, confidence intervals encourage comparing effect sizes, and invalid comparisons are often made between other groups. Assume, for example, that the knee replacement revision risks of a low- (A) and a high (B) -exposed group of smokers are compared with that of a group of non-smokers (C). The three-group comparison leads to two relative risks, A/C and B/C, both having confidence intervals. These cannot be directly compared; they depend on C. An alternative analysis method, floating absolute risks (FAR), have been developed as a solution to this problem (9).

In conclusion, hypothesis tests and their p-values will probably continue to be important tools for interpreting scientific data. Attempts to ban p-values from scientific journals have not been successful (10), and the aim of this discussion is not to stop authors from using p-values. However, much can be gained by developing the statistical analysis strategy of scientific studies. A better understanding of statistical inference and a more frequent use of confidence intervals are likely to play important roles in such developments.  This is not restricted to clinical research. The phenomena discussed here are as important in laboratory science (8, 11). *Osteoarthritis and Cartilage* recommends  confidence interval as uncertainty measure in all studies (12). More information on this subject can be found in the guide for authors.

**Conflict of Interest**

None.

## References

1. RigbyAS. Getting past the statistical referee: moving away from P-values and towards interval estimation.  Health Educ Res 1999;14:713-715.

2. Nester MR. An applied statistician's creed. Appl Statist 1996;45:401-410.

3. Fidler F, Thomason N, Cumming G, Finch S, Leeman J. Editors can lead researchers to confidence intervals, but can't make them think. Psychol Sci 2004;15:119-126.

4. Ranstam J. Sampling uncertainty in medical research. Osteoarthritis Cartilage 2009;17:1416-1419.

5. Hunter JM. Editorial 1 - Ethics in publishing; are we practising to the highest possible standards? Br J Anaesth 2000;85:341-343.

6. Altman DG, Bland M. Statistics notes: Absence of evidence is not evidence of absence. BMJ 1995;311:485.

7. Austin P, Hux J. A brief note on overlapping confidence intervals. J Vasc Surg 2002;36:194-195.

8. Vaux D. Ten rules of thumb for presentation and interpretation of data in scientific publications. Australian Biochemist 2008;39:37-39.

9. Easton DF, Peto J, Babiker AG. Floating absolute risk: an alternative to relative risk in survival and case-control analysis avoiding an arbitrary reference group. Stat Med 1991;10:1025-1035.

10. Editorial. The value of P. Epidemiology 2001;12:286.

11. Cumming G, Fidler F, Vaux D. Error bars in experimental biology. J Cell Biol 2007;177:7–11

12. Ranstam J, Lohmander SL. Ten recommendations for Osteoarthritis and Cartilage (OAC) manuscript preparation, common for all types of studies. Osteoarthritis Cartilage 2011;19:1079-1080.

**Legend**

Figure 1. Statistically and clinically significant effects, measured in arbitrary units on an absolute scale, as evaluated by p-values and confidence intervals.

Figure 2. The use of confidence intervals in superiority, non-inferiority and equivalence trials, measured in arbitrary units on an absolute scale.

Figure 1.

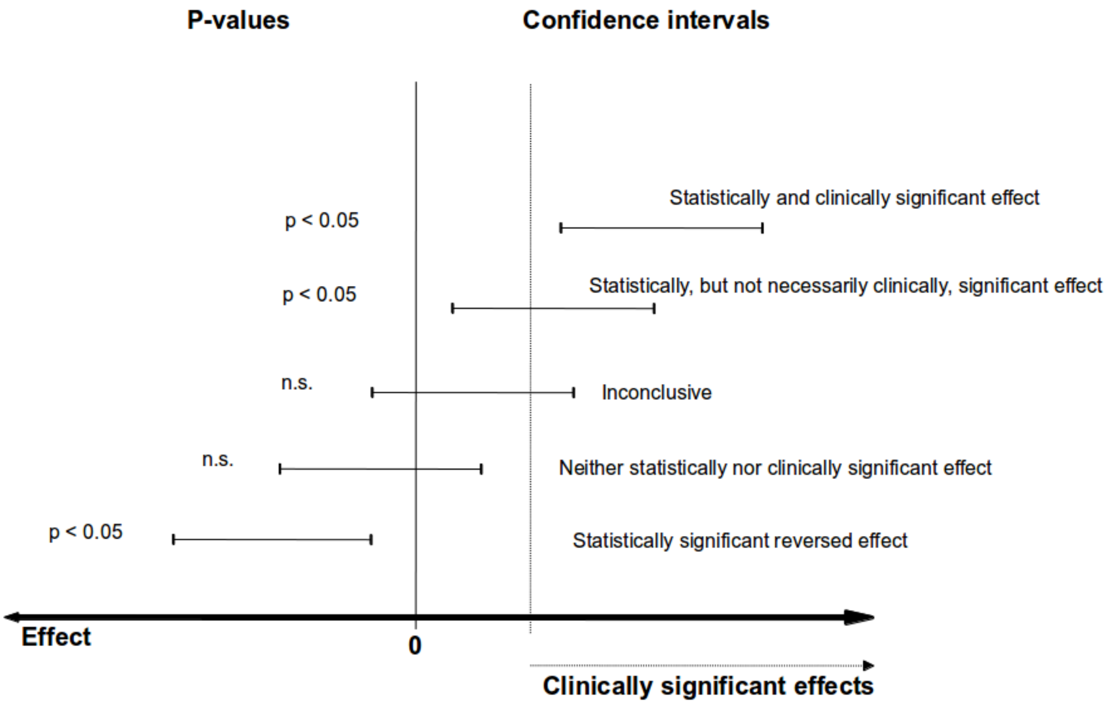
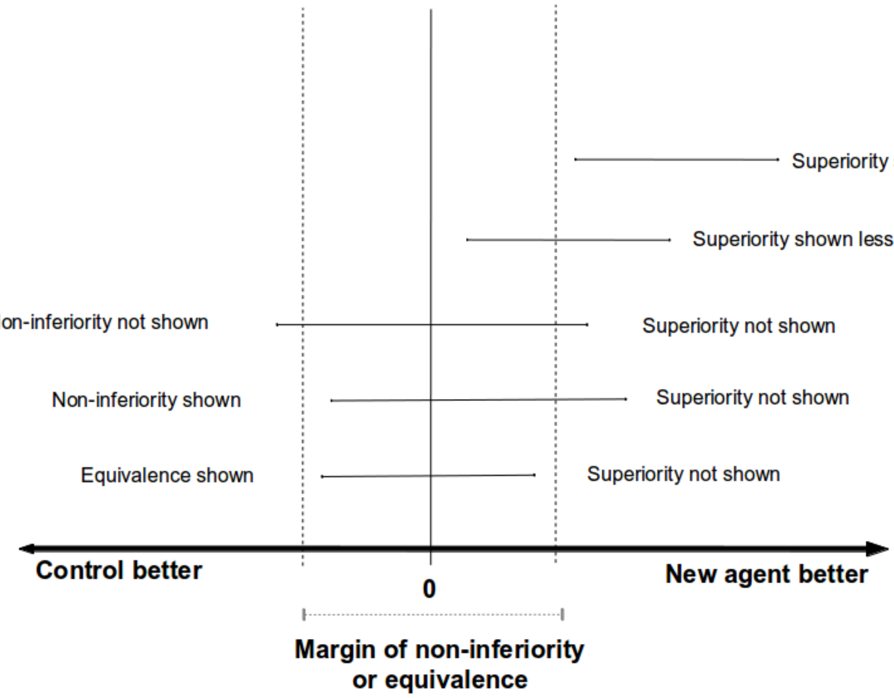

P-values
Confidence intervals
p < 0.05
Statistically and clinically significant effect
p < 0.05
Statistically, but not necessarily clinically, significant effect
n.s.
Inconclusive
n.s.
Neither statistically nor clinically significant effect
p < 0.05
Statistically significant reversed effect
Effect
0
Clinically significant effects

Figure 2.


Superiority shown
Superiority shown less strongly
Non-inferiority not shown
Superiority not shown
Non-inferiority shown
Superiority not shown
Equivalence shown
Superiority not shown
Control better
0
New agent better
Margin of non-inferiority
or equivalence