



LUND UNIVERSITY

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](https://doi.org/10.1016/j.joca.2012.04.001)

2012

[Link to publication](#)

Citation for published version (APA):

Ranstam, J. (2012). Why the P-value culture is bad and confidence intervals a better alternative. *Osteoarthritis and Cartilage*, 20(8), 805-808. <https://doi.org/10.1016/j.joca.2012.04.001>

Total number of authors:
1

General rights

Unless other specific re-use rights are stated the following general rights apply:

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: <https://creativecommons.org/licenses/>

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

April 1, 2012

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31

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.

32 **Abstract**

33

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

38

39

40 **Abbreviations**

41

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

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

59

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

64

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

71

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

79

80 The current tradition in medical research of screening variables with hypothesis tests to categorize

81 findings either as statistically significant or insignificant is a simplistic and counterproductive analysis
82 strategy that should be abandoned. This brief editorial attempts to explain why.

83

84 **Statistical precision**

85

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

89

90

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

91

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

94

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

99

100

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

101

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

104

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

107

108

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

109

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

112 have higher. Conversely, the confidence interval

113

$$114 \quad d \pm t_{0.05}SE$$

115

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

117

118 **P-values**

119

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

125

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

130

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

138

139 Third, a statistically insignificant difference between two observed groups (*the sample*) does not
140 indicate that this effect does not exist in the *population* from which the sample is taken, because the p-
141 value is confounded by the number of observations; it is based on the SE, which has \sqrt{n} in the
142 denominator. A statistically insignificant outcome indicates nothing more than that the observed sample

143 is too small to detect a population effect. A statistically insignificant outcome should be interpreted as
144 "absence of evidence, not evidence of absence" (6).

145

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

152

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

161

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

168

169 **Confidence intervals**

170

171 Confidence intervals share some of the p-value's weaknesses, like the multiplicity problem, and
172 analogous with the adjustment of the significance level, the width of confidence intervals can also be
173 adjusted in cases of multiplicity. However, the great advantage with confidence intervals is that they do

174 show what effects are likely to exist in the *population*. Values excluded from the confidence interval are
175 thus not likely to exist in the population. Consequently, a confidence interval excluding a specific effect
176 can be interpreted as providing evidence against the existence (in the unobserved population) of such
177 an effect. The confidence interval limits do thereby allow an easy and direct evaluation of clinical
178 significance, see Figure 1.

179

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

184

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

191

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

196

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

204 Overlapping confidence intervals can therefore not be directly interpreted in terms of statistical

205 significance (7).

206

207 Standard errors are also often used to indicate uncertainty, as error bars in graphical presentations.

208 Using confidence intervals is, however, a better alternative because the uncertainty represented by a

209 standard error is confounded by the number of observations (8). For example, one standard error

210 corresponds to a 58% confidence interval when n is 3 and to a 65% confidence interval when $n=9$.

211

212 When pairwise multiple groups are compared with one and the same reference or control group in

213 terms of relative risk or odds ratios, comparisons of confidence intervals are only valid vis-a-vis the

214 reference group. However, confidence intervals encourage comparing effect sizes, and invalid

215 comparisons are often made between other groups. Assume, for example, that the knee replacement

216 revision risks of a low- (A) and a high (B) -exposed group of smokers are compared with that of a

217 group of non-smokers (C). The three-group comparison leads to two relative risks, A/C and B/C , both

218 having confidence intervals. These cannot be directly compared; they depend on C. An alternative

219 analysis method, floating absolute risks (FAR), have been developed as a solution to this problem (9).

220

221 In conclusion, hypothesis tests and their p-values will probably continue to be important tools for

222 interpreting scientific data. Attempts to ban p-values from scientific journals have not been successful

223 (10), and the aim of this discussion is not to stop authors from using p-values. However, much can be

224 gained by developing the statistical analysis strategy of scientific studies. A better understanding of

225 statistical inference and a more frequent use of confidence intervals are likely to play important roles in

226 such developments. This is not restricted to clinical research. The phenomena discussed here are as

227 important in laboratory science (8, 11). *Osteoarthritis and Cartilage* recommends confidence interval

228 as uncertainty measure in all studies (12). More information on this subject can be found in the guide

229 for authors.

230

231 **Conflict of Interest**

232

233 None.

234 **References**

- 235 1. Rigby AS. Getting past the statistical referee: moving away from P-values and towards interval
236 estimation. *Health Educ Res* 1999;14:713-715.
- 237 2. Nester MR. An applied statistician's creed. *Appl Statist* 1996;45:401-410.
- 238 3. Fidler F, Thomason N, Cumming G, Finch S, Leeman J. Editors can lead researchers to
239 confidence intervals, but can't make them think. *Psychol Sci* 2004;15:119-126.
- 240 4. Ranstam J. Sampling uncertainty in medical research. *Osteoarthritis Cartilage* 2009;17:1416-
241 1419.
- 242 5. Hunter JM. Editorial 1 - Ethics in publishing; are we practising to the highest possible standards?
243 *Br J Anaesth* 2000;85:341-343.
- 244 6. Altman DG, Bland M. Statistics notes: Absence of evidence is not evidence of absence. *BMJ*
245 1995;311:485.
- 246 7. Austin P, Hux J. A brief note on overlapping confidence intervals. *J Vasc Surg* 2002;36:194-195.
- 247 8. Vaux D. Ten rules of thumb for presentation and interpretation of data in scientific publications.
248 *Australian Biochemist* 2008;39:37-39.
- 249 9. Easton DF, Peto J, Babiker AG. Floating absolute risk: an alternative to relative risk in survival
250 and case-control analysis avoiding an arbitrary reference group. *Stat Med* 1991;10:1025-1035.
- 251 10. Editorial. The value of P. *Epidemiology* 2001;12:286.
- 252 11. Cumming G, Fidler F, Vaux D. Error bars in experimental biology. *J Cell Biol* 2007;177:7-11
- 253 12. Ranstam J, Lohmander SL. Ten recommendations for Osteoarthritis and Cartilage (OAC)
254 manuscript preparation, common for all types of studies. *Osteoarthritis Cartilage* 2011;19:1079-
255 1080.
- 256
- 257

258

259

260 **Legend**

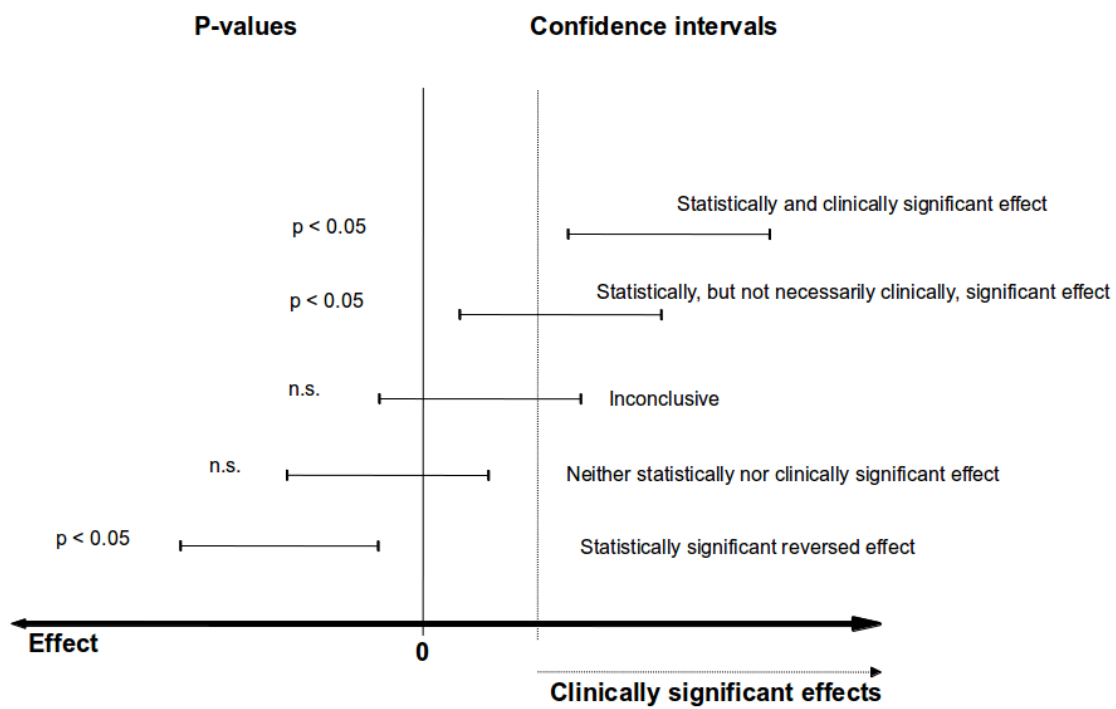
261

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

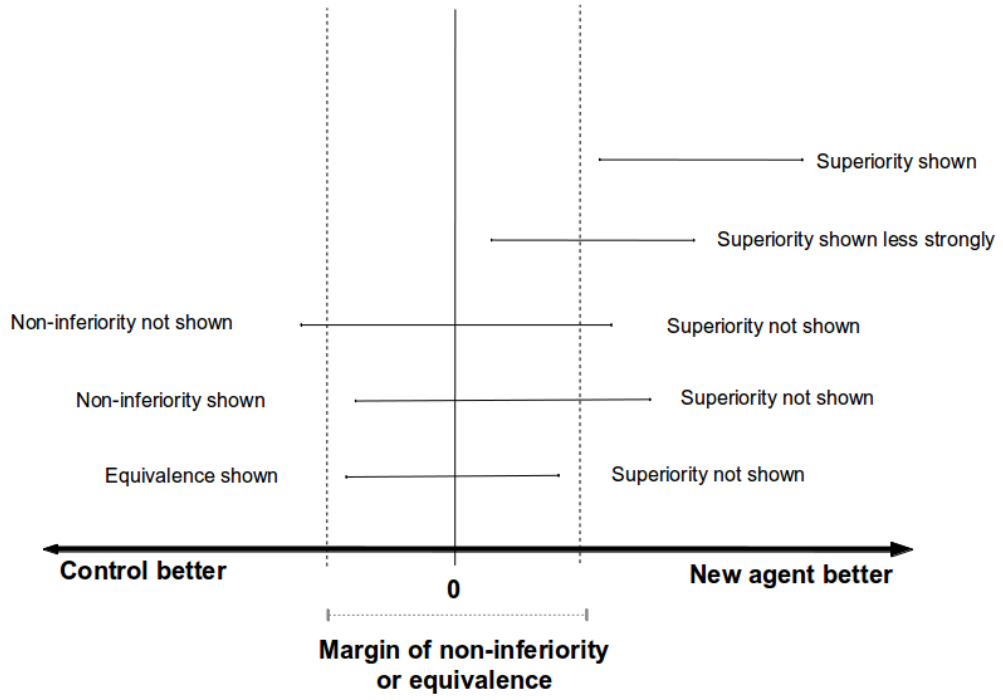
264

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

267 Figure 1.



269 Figure 2.
270



271