The European Journal of General Practice

ISSN: 1381-4788 (Print) 1751-1402 (Online) Journal homepage: https://www.tandfonline.com/loi/igen20

# The insignificance of statistical significance

Sil Aarts, BjöRn Winkens & Marjan van Den Akker

**To cite this article:** Sil Aarts, BjöRn Winkens & Marjan van Den Akker (2012) The insignificance of statistical significance, The European Journal of General Practice, 18:1, 50-52, DOI: 10.3109/13814788.2011.618222

**To link to this article:** https://doi.org/10.3109/13814788.2011.618222

Published online: 05 Dec 2011.

*European Journal of General Practice,* 2012; 18: 50–52

informa
healthcare

**Background Paper**

# The insignificance of statistical significance

*Sil Aarts*[1,2]*, Björn Winkens*[3] *& Marjan van den Akker*[1,4]

[1]Department of General Practice, School for Public Health and Primary Care: CAPHRI, Maastricht University, the Netherlands, [2]Department of Psychiatry and Neuropsychology, School for Mental Health and Neuroscience: MHeNS, Maastricht University, the Netherlands, [3]Department of Statistics and Methodology, School for Public Health and Primary Care: CAPHRI, Maastricht University, the Netherlands, and [4]Department of General Practice, Katholieke Universiteit Leuven, Belgium

KEY MESSAGE(S):

- Medical research will not improve if we simply interpret our results using the clear-cut difference between statistical significance and non-significance
- Medical researchers should be more interested in the size of the observed result than in its statistical significance
- Confidence intervals provide more information than P-values

**Key words:** statistics, significance, clinical relevance, p-value, confidence intervals

Hypothesis testing is often used to provide evidence to support medical research findings. Hypothesis testing allows the researcher to make inferences about an entire population based on information obtained from a relatively small sample of individuals (1). However, hypothesis testing and the corresponding *P*-value present many challenges for interpretation (2).

Many published scientific journal articles include statements such as 'is statistically significant' or 'a statistically significant *P*-value of 0.031 was found' (3). Speakers at medical congresses and workshops also focus on statistical significance. Unfortunately, this focus does not necessarily facilitate decisions about the relevance of the obtained results (3,4). The current paper is not intended to provide an extended and exhaustive overview of statistical analyses and their interpretations; rather, it discusses the interpretation of *P*-values and confidence intervals and the clinical importance of these measures.

## STATISTICAL SIGNIFICANCE

Medical researchers gather sample data to assess the amount of evidence for a specific association or effect. Hypothesis testing has been used for decades to quantify beliefs against a particular hypothesis or in favour of another postulated hypothesis (5). These analyses rely on an arbitrary division of 'significant' or 'non-significant', often employing a threshold of $P = 0.05$ (i.e. an alpha of 5%) to assess the evidence against an a priori postulated null hypothesis (6,7). This clear cut-off implies that on 5% of the cases where the null hypothesis is true, it will incorrectly be rejected (if an alpha of 1% is used, incorrect rejection of the null hypothesis will occur in 1% of the cases). For example, of every 1000 *P*-values that are reported in medical manuscripts, 50 *P*-values will incorrectly accept the alternative hypothesis. Moreover, *P*-values depend on the sample size at hand: a larger sample size will increase the *power* of a study, which is the study's ability to detect a statistically significant difference (i.e. even small associations or effects will be detected). Consequently, clinically irrelevant effects or associations (i.e. a regression coefficient near zero) will appear statistically significant in studies based on a large sample size, whereas possible clinically important differences observed in small studies will be ignored because of their non-significance (8).

Correspondence: S. Aarts Department of General Practice, School for Public Health and Primary Care: CAPHRI, Maastricht University, the Netherlands E-mail: sil.aarts@maastrichtuniversity.nl

*(Received 8 July 2011; accepted 23 August 2011)*

ISSN 1381-4788 print/ISSN 1751-1402 online 
DOI: 10.3109/13814788.2011.618222

Let us consider a hypothetical study in which the 15-item geriatric depression scale (GDS-15) (9) was administered to determine the association between depression and sex. A cut-off score of six (scores ranging from 0 to 15, with a higher score being indicative for more depressive symptoms) was set to differentiate individuals with a clinical depression from non-depressed individuals. We hypothesize that females and males differ regarding their score on the GDS-15. If the mean difference (females' score– males' score) is positive, females report more depressive symptoms than males do, whereas a negative mean difference indicates that males report more depressive problems. The mean score for females is 3.3, and the mean score for males is 3.1. If we use an independent t-test (or an analysis of variance or regression analyses), this mean difference of 0.2 points appears highly statistically significant (e.g. $P < 0.001$). Therefore, we might conclude that females significantly report more depressive symptoms than their male counterparts. The question remains whether this difference has any *clinical significance* or *practical value.* Unfortunately, *P*-values are solely a measure for the evidence for the null hypothesis and give us no indication whatsoever as to the clinical importance of our observed difference (1).

## CONFIDENCE INTERVALS

A confidence interval (CI) provides more information than a *P*-value and can help us with the interpretation of our research findings. Most medical studies are based on a 95% CI. This implies that with 95% confidence, the true population value is likely to fall within the confidence interval, bounded by a lower and upper extremity (5). Put another way, if 100 random samples are drawn from the same population, 95% of the confidence intervals obtained in these samples will include the true population value. Confidence intervals not only provide evidence for statistical significance (i.e. an effect or association will be significant when the 95% CI does not include the value that is postulated as the null hypothesis), (1) but also provide information on the magnitude and direction of the obtained results.

Let us reconsider the above-mentioned hypothetical study. The null hypothesis states that the mean difference between females and males on the GDS-15 (scale ranging from 0 to 15) is zero. Hence, if zero is detected in the 95% CI, the null hypothesis is not rejected. Examples of possible study results, using an $\alpha$ of 5%, are displayed in Table I. Although example 1 is statistically significant (even when an $\alpha$ of 1% is used), this result is unlikely to have any clinical relevance because the difference between females and males is extremely small. Although this result is highly significant, the claim that females significantly report more depressive problems than their male counterparts do is unlikely to be of any real importance. Example 2 is not only statistically significant but also clinically relevant; the difference between females and males on the GDS-15 is approximately two whole points. Moreover, the confidence interval is quite narrow, which indicates that the sample size is large enough to make a proper judgement. Example 3 is not statistically significant. The confidence interval in this example is very large (almost six points), which makes it difficult to draw any firm conclusions. Since the confidence interval in this example includes both negative and positive values, it is not yet clear if there is a difference between these two groups (if females report more depressive symptoms than males or vice versa). Consequently, this study should be repeated using a larger sample size, which will decrease the width of the confidence interval.

Table I. Three examples of possible results observed in a hypothetical study.

| | Mean difference* | *P*-value | 95% CI | *N* |
|---|---|---|---|---|
| Example 1 | 0.15 | 0.001 | 0.05–0.25 | 2000 |
| Example 2 | 2.10 | 0.005 | 1.25–2.95 | 1200 |
| Example 3 | 1.30 | 0.089 | 21.10–3.70 | 400 |

*Score ranging from 1 to 15 with a higher score being indicative for more depressive symptoms.

As shown in these three examples, the CI provides additional and more useful information than *P*-values.

## SUGGESTED STRATEGY

To date, statistical significance has, unfortunately, often been thought to be equivalent to clinical importance (10). However, because a *P*-value is simply a dichotomous measure for the amount of evidence against a null hypothesis (i.e. a *P*-value near zero provides more evidence against the null hypothesis), it does not provide any information on the clinical importance of a research finding. Consequently, medical research will not improve if we simply interpret our results using the clear-cut difference between significance or non-significance (11). It is extremely important not to trust the clear-cut, arbitrary level of a *P*-value or to perceive it as a golden standard; more careful considerations are required, such as confidence intervals. Moreover, the interpretation of the results should be evaluated in light of other available statistical measures of evidence, such as relevant difference, explained amount of variance (i.e. $R^2$), odds ratios, survival rates, correlation coefficients and regression coefficients. Because critical appraisal is a rather subjective matter, the observed results should be interpreted in terms of context and type of the study and should be compared with the available medical literature.

In conclusion, medical researchers should be more interested in the size of the observed result than whether the result is statistically significant. Since the conclusions reached in medical studies provide input for further medical research and lead to medical decisions, the

medical researcher should use his or her medical knowledge and clinical expertise to evaluate the strength of the observed results, whether they are significant or not.

***Declaration of interest:*** The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

## REFERENCES


1. Akobeng AK. Confidence intervals and p-values in clinical decision making. Acta Paediatr. 2008;97:1004–7.
2. Greenfield ML, Kuhn JE, Wojtys EM. A statistics primer. Validity and reliability. Am J Sports Med. 1998;26:483–5.
3. Fethney J. Statistical and clinical significance, and how to use confidence intervals to help interpret both. Aust Crit Care 2010;23:93–7.
4. Stang A, Rothman KJ. That confounded P-value revisited. J Clin Epidemiol. 2011;64:1047–8.
5. Sterling TD. Publication decisions and their possible effects on inferences drawn from test of significance-or vice versa. J Am Stat Assoc. 1959;54:304.
6. Howell D. Statistical methods for psychology. 5th ed. Pacific Grove, CA: Duxbury Press; 2002.
7. Moore DS, McCabe GP. Introduction to the practice of statistics. 4th ed. New York: W.H. Freeman; 2003.
8. Petrie A, Sabin C. Systematic reviews and metaanalysis. In: Medical statistics at a glance. Malden, MA: Blackwell; 2005. pp. 116–8.
9. Sheikh JI, Yesavage JA. Geriatric depression scale (GDS): Recent evidence and development of a shorter version. In: Clinical gerontology: A guide to assessment and intervention. New York: Haworth Press; 1986. pp. 165–73.
10. du Prel JB, Hommel G, Rohrig B, Blettner M. Confidence interval or p-value: Part 4 of a series on evaluation of scientific publications. Dtsch Arztebl Int. 2009;106:335–9.
11. Sterne JA, Davey Smith G. Shifting the evidence-what's wrong with significance tests? Br Med J. 2001;322:226–31.