---

**Article:**
Dhaliwal, S. and Campbell, M.J. (2010) Misinterpreting p-values in research. Australasian Medical Journal, 1 (1). pp. 1-2. ISSN 1836-1935

https://doi.org/10.4066/AMJ.2009.191

---

Australasian Medical Journal, 2010, 1, 1-2

# Misinterpreting P-Values In Research

Satvinder S. Dhaliwal[1], Michael J. Campbell[2]

[1]Associate Professor and Director, Public Health and Epidemiology Directorate, School of Public Health, Curtin Health Innovation Research Institute (CHIRI) and Australian Technology Network (ATN) Centre for Metabolic Fitness, Curtin University of Technology, Bentley, Western Australia

[2]Professor of Medical Statistics, Medical Statistics Group, School of Health and Related Research, University of Sheffield, United Kingdom

**REVIEW**

Please cite this paper as: Dhaliwal SS, Campbell MJ. Misinterpreting P-Values In Research. AMJ, 2010, 1, 1-2. Doi 10.4066/AMJ.2009.191

**Corresponding Author:**
Satvinder S. Dhaliwal
Associate Professor and Director, Public Health and Epidemiology Directorate, School of Public Health, Curtin Health Innovation Research Institute (CHIRI) and Australian Technology Network (ATN) Centre for Metabolic Fitness, Curtin University of Technology, Bentley, Western Australia
s.dhaliwal@curtin.edu.au

**Abstract**

The overuse of p-values to dichotomize the results of research studies as being either significant or non-significant has taken some investigators away from the main task of determining the size of the difference between groups and the precision with which it is measured. Presenting the results of research as statements such as “p < 0.05”, “p > 0.05”, “NS” or as precise p-values has the effect of oversimplifying study findings. Further information regarding the size of the difference between groups is required. Presenting confidence intervals for the difference in effect, of say two treatments, in addition to p-values, has the distinct advantage of presenting imprecision on the scale of the original measurement. A statistically significant test also does not imply that the observed difference is clinically important or meaningful, and their meanings are often confused.

**Key Words**
p-value, confidence interval, clinical significance, equivalence test

Current practice and the overuse of p-values to dichotomize the results of research studies as being either significant or non-significant has taken some investigators away from the main task of determining the size of the difference between groups and the precision with which it is measured. The convention of using the 5% level of significance has led investigators and students to be complacent in their thinking and hence ignore the size of the difference between groups.

In the testing of hypotheses, test statistics are calculated from the information contained in the sample data. As a simple example of a hypotheses test which involves the comparison of two groups (for example the effects of two treatments), the null hypothesis which states the equality of two means or proportions is tested against the alternative where the two means or proportions are unequal. That is, it tests if the difference between the two groups is large relative to the size of variability determined from the data. Depending on the test performed, the calculated test-statistic is compared against its respective distribution. The *p-value* is the probability that the test statistic takes on the calculated or a more extreme value when the null hypothesis is true.

The *p-value* is not a yes/no answer. The larger the difference between the two groups relative to the size of the variability, the smaller the *p-value*. The smaller the *p-value*, the greater the evidence is against the null hypothesis which states the means or proportions are equal.

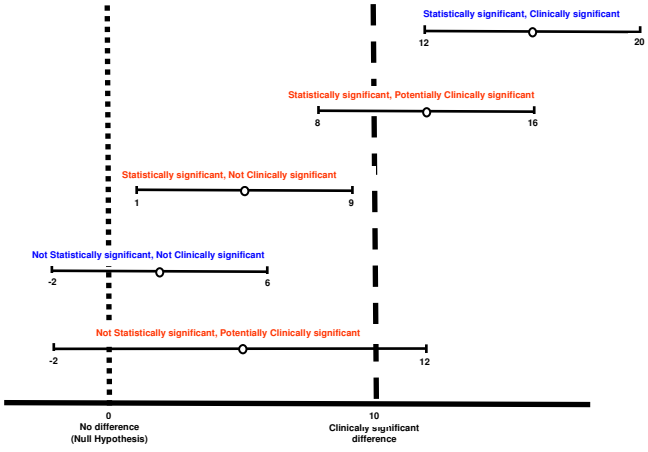
The *p-value* is then usually compared to the level of significance (or α) which is conventionally set at 5% to determine if the difference observed is statistically significant and, a decision is made as to whether or not to reject the null hypothesis of equality. The level of significance, or α, is the probability of committing a type I error or the probability of making the incorrect decision of rejecting the null hypothesis that the two groups are equal when they are in fact equal in effectiveness. An alternative way of looking at this comparison of *p-value* against α is that if there is only a 5% change of a difference occurring by chance then we can confidently (95% of the time) accept that the effect we have observed is unlikely to have arisen by chance and hence conclude that the finding is statistically significant. If we lower the probability of accepting an effect as genuine, with a smaller α, we are essentially increasing the probability that we will say that there is no effect, when in fact one genuinely exists.

Presenting the results of research as statements such as “p < 0.05” and “p > 0.05” or “NS” has the effect of oversimplifying study findings. Precise p-values also do not provide any further information regarding the size of the difference between groups.

A statistically significant test does not imply that the observed difference is important or meaningful. It is advisable to represent difference observed between means or the strength of association or relationship between variables as a standardised measure referred to as an effect size. The use of effect sizes to provide an objective measure of the importance of the observed effect or importance of a research finding is highly recommended. It is possible for small or unimportant effects to be statistically significant (low *p-values*) when the number of subjects used in the study is large. Or it is possible for an important or meaning effect to be non-significant when it is of clinical significance. Statistical significance does not necessarily imply clinical significance, and their meanings are often confused.[1]

The 95% confidence interval, usually calculated during analyses, gives the range of values within which the population value is expected to lie. Shorter confidence intervals, which can be achieved with larger sample sizes, indicate higher precision in the estimation of the population value. Presenting confidence intervals for the difference in effect, of say two treatments, in addition to p-values, has the distinct advantage of presenting imprecision on the scale of the original measurement. Confidence intervals also can be used to generalise the results of the research study to the wider population.[2]

The figure below illustrates the difference between statistical significance and clinical significance. In a study to compare the effect of a drug versus placebo to reduce systolic blood pressure where a mean difference of 10mmHg is considered clinically meaningful, this figure illustrates the interpretation of confidence intervals in relation to a clinically relevant difference. If the confidence interval for the difference does not include zero, the difference is statistically significant. Confidence intervals in red-font are to be interpreted with caution. If the confidence interval lies in the range of 0 to 10, then it lies in a region of clinical indifference and confidence intervals that include 10 in its range could be potentially clinically significant.




**Statistical significance and clinical significance (adapted from Campbell et al, 2007)**

Equivalence tests allow the comparison of groups to determine if the difference is within a small acceptable range, as defined by the equivalence bounds. Two groups are considered equivalent if their difference is within the clinically acceptable range specified by the investigator. In equivalence tests, the null hypothesis states that the two groups are non-equivalent and is tested against the alternative hypothesis of equivalence.[3]

Example: To compare the waist circumference (cm) measurements of adult men who were born either in Australia or United Kingdom and Ireland in order to determine if the same waist circumference cut-points can be used for the assessment of obesity as required in the definition of the metabolic syndrome.[4] It was decided that a difference of less than 2 cm was not meaningful. The results are presented in the box below:

| *Australia (n=3234)* | *United Kingdom and Ireland (n= 495)* | *Mean difference (95% confidence interval)* | *P-value from Independent samples t-test* | *Equivalence test, using equivalence bounds of ± 2cm* |
|---|---|---|---|---|
| Mean: 90.5<br>Std Dev:10.7 | Mean: 89.4<br>Std Dev: 10.1 | 1.07 (0.06 – 2.07) | 0.038 | Equivalent |

The difference between the two groups is statistically significant (p=0.038) but not meaningful since the difference between the mean of the groups is only 1.07cm! This difference is less than the measurement error calculated for waist circumference measurements (1.84cm). Furthermore, the 95% confidence interval lies largely in the region of clinical indifference. The two groups are also found to be equivalent with the specified bounds using the Equivalent Test.

In conclusion, when presenting research findings in scientific papers it is recommended to include confidence intervals or effect sizes for major findings when appropriate. Alternative tests such as equivalence tests should be considered when comparing groups, especially with large sample sizes.

## References

1. Campbell MJ, Machin D, Walters SJ. Medical Statistics. A textbook for the health sciences. Fourth edition. John Wiley & Sons Ltd. 2007.
2. Gardner MJ, Altman DG. Statistics with confidence – Confidence intervals and statistical guidelines. BMJ, London. 1989.
3. Wellek S. Testing statistical hypothesis of equivalence. CRC Press LLC, Boca Raton. 2003.
4. Dhaliwal SS, Welborn TA. Measurement error and ethnic comparisons of measures of abdominal obesity. Preventive Medicine 2009, 49, 148-152.

## CONFLICTS OF INTEREST

The authors declare that they have no competing interests.

## PEER REVIEW

Commissioned, not externally peer reviewed.