This is an author produced version of Misinterpreting p-values in research.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/11208/

**Article:**

http://dx.doi.org/10.4066/AMJ.2009.191
Misinterpreting P-Values In Research
Satvinder S. Dhaliwal1, Michael J. Campbell2

1Associate 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

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

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

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

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.
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.

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. It was decided that a difference of less than 2 cm was not meaningful. The results are presented in the box below:

<table>
<thead>
<tr>
<th>Australia (n=3234)</th>
<th>United Kingdom and Ireland (n=495)</th>
<th>Mean difference (95% confidence interval)</th>
<th>P-value from Independent samples t-test</th>
<th>Equivalence test, using equivalence bounds of ± 2cm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean: 96.5</td>
<td>Mean: 89.4</td>
<td>2.07 (0.06 – 2.07)</td>
<td>0.038</td>
<td>Equivalent</td>
</tr>
<tr>
<td>Std Dev:10.7</td>
<td>Std Dev: 10.1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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


CONFLICTS OF INTEREST
The authors declare that they have no competing interests.

PEER REVIEW
Commissioned, not externally peer reviewed.