ESC European Society of Cardiology

European Heart Journal: Acute Cardiovascular Care (2024) **13**, 181–182
https://doi.org/10.1093/ehjacc/zuad152

STATISTICAL SPOTLIGHT
Educational Papers

The *P*-value spectrum: from 'absence of evidence' to 'evidence of difference'

Xavier Rossello [1,2,3]*

[1]Cardiology Department, Institut D'investigació Sanitària Illes Balears (Idisba), Hospital Universitari Son Espases, Carretera de Valldemossa 79, 07120 Palma, Spain; [2]Clinical Research Department, Centro Nacional de Investigaciones Cardiovasculares (CNIC), C. de Melchor Fernández Almagro 3, 28029 Madrid, Spain; and [3]Faculty of Medicine, Universitat De Les Illes Balears (UIB), 07122 Palma, Spain

Online publish-ahead-of-print 14 December 2023

Medical research is often dogmatically classified as 'positive' or 'negative' based on whether or not a *P*-value for the main comparison achieves a value of less than 0.05.[1] For many years, *P*-values have been misinterpreted and misused.

What is a *P*-value?

The majority of statistical analyses involve comparisons, most obviously between treatments or between groups of subjects. The numerical value corresponding to the comparison of interest is often called the effect, or association. In the null hypothesis, we state that the effect (or association) of interest is zero (e.g. cardiovascular mortality is the same for men and women). This statistical null hypothesis is commonly the negation of the research hypothesis (the alternative hypothesis usually is formulated as an effect or association different than zero).[2] Following the same example, the alternative hypothesis is that cardiovascular mortality is not the same for men and women.

The *P*-value is the probability of obtaining a difference as great as that actually observed if the null hypothesis were true.[3] In the previous example, this refers to the observed difference in cardiovascular mortality between men and women. By having a sample and not the entire population of men and women, we might observe some differences by chance. In interpreting *P*-values, we relate our data to the likely variation in a sample due to chance when the null hypothesis is true in the population. So the *P*-value is the probability of observing such a difference in cardiovascular mortality between men and women, or even larger, when there is actually no difference. Importantly, the smaller the *P*-value, the stronger the evidence against the null hypothesis (e.g. the more convincing the evidence is that a truly difference might exist).

Why is medical literature so obsessed with *P*-values?

Hypothesis tests are a means to make decisions with an acceptance or rejection of the null hypothesis based on the *P*-value. A cut-off is therefore chosen to decide whether the null hypothesis is rejected (if smaller than the cut-off).[4] Some medical research is indeed concerned with decision making (e.g. randomized controlled trials to evaluate treatment efficacy). However, much medical research is not. If we evaluate the risk of a particular disease with regard to a given condition (e.g. sex and history of myocardial infarction), we put the focus on the aetiological path of the disease. Let's put this in plain English. In a randomized controlled trial to evaluate treatment efficacy, we test the hypothesis about whether treatment A reduces outcomes compared with treatment B (so we need an unambiguous decision rule to decide when treatment A would be considered different than treatment B). However, in an observational study, we might want to test how likely is for a risk factor (e.g. sex and history of myocardial infarction) to cause a disease.[5] In the first scenario, we need a tool with a pre-specified threshold to make a decision about efficacy, to then subsequently generalize the use of the treatment. In the second scenario, we are rather interested in using a *P*-value as a measure of the strength of evidence against the null hypothesis, rather than as an aid to decision-making. We cannot change our biological sex or previous history of myocardial in order to reduce the risk of the disease (the focus is rather on understanding the disease, on how likely some factors to influence it).

How did we get here: a historical and contemporary perspective

The idea of significance testing was introduced by RA Fisher, who saw the *P*-value as an index measuring the strength of evidence against the null hypothesis. Neyman and Pearson introduced the hypothesis testing formulation and proposed a more dogmatic view of the *P*-value, which was focused on a decision rule for interpreting the results of an experiment (e.g. randomized controlled trial) in advance.[4]

Because of the current requirements of regulatory bodies, and the historical appetite of medical journals and readers for positive news, the use of the *P*-value became dominated by a division of results into significant and non-significant, regardless of the original research question (e.g. testing an intervention vs. assessing a risk factor).

P-values are (many times) misinterpreted and misused

(1) A *P*-value does not represent the probability that the alternative hypothesis is true. A *P*-value of 0.01 (1%) does not mean that the null hypothesis is 1% likely to be true (a drug is as effective as placebo), and the alternative hypothesis is 99% likely to be correct (a drug is more effective than placebo).
(2) A *P*-value >0.05 means that there is insufficient evidence that the null hypothesis is true. However, it is incorrect to infer that there is

* Corresponding author. Tel: +34 871 205050 (ext. 64527), Email: fjrossello@ssib.es

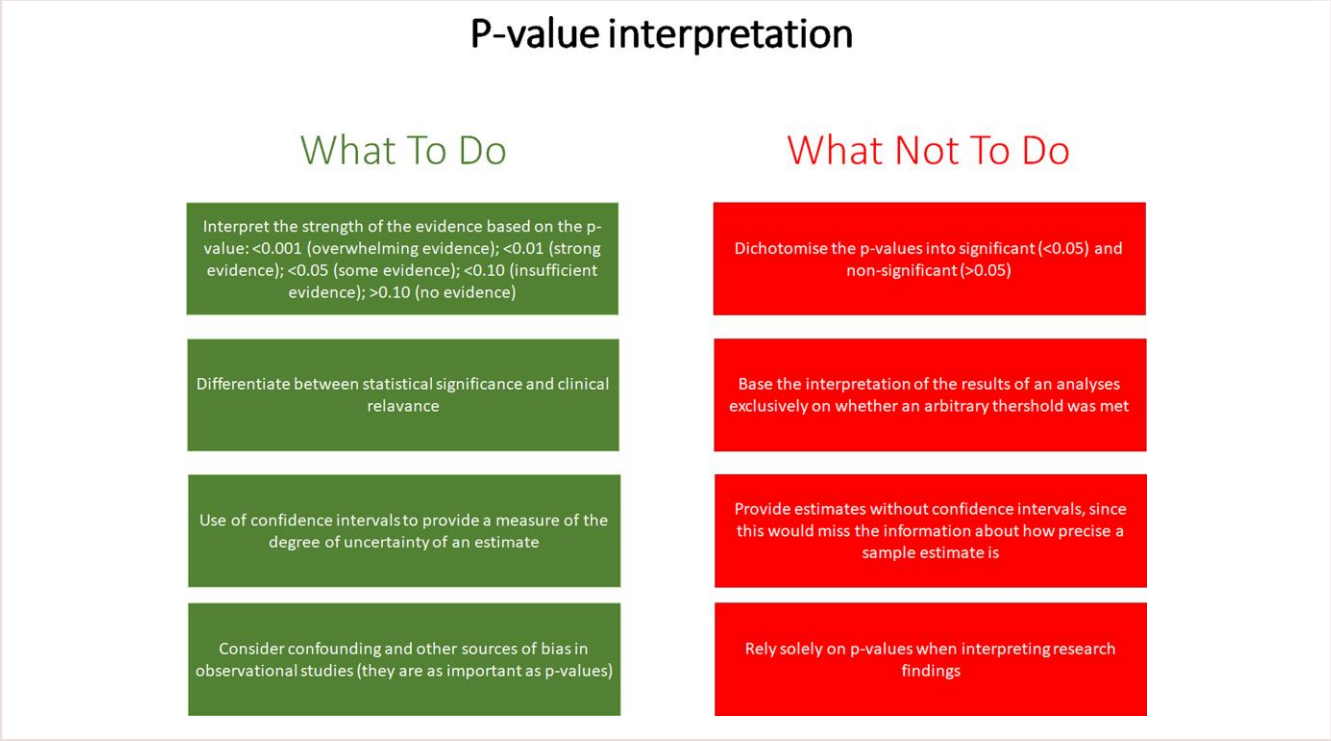

Figure 1 What to do and what not to do when interpreting *P*-values.

evidence that the null hypothesis is true since the 'absence of evidence' is not the same as 'evidence of absence'.

(3) *P*-values are often used for hypothesis testing (to evaluate whether it is below a pre-specified bar), whereas sometimes we are rather interested in the strength of the evidence against the null hypothesis.

(4) A *P*-value <0.05 does not necessarily translate into clinically meaningful differences, which are worth acting on.

Conclusions

The goal of medical statistics should be to provide an evaluation of certainty or uncertainty regarding the size of an association. A slavish focus on whether or not a *P*-value is above or below 0.05 is a rather simplistic approach to the interpretation of the results of an analysis. Some recommendations about what to do, and what not to do are shown in *Figure 1*.

Funding

None declared.

Conflict of interest: None declared.

Data availability

There are no original data in this work.

References

1. Ioannidis JPA. The proposal to lower P value thresholds to.005. *JAMA* 2018;**319**: 1429–1430.
2. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, *et al.* Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol* 2016;**31**:337–350.
3. Pocock SJ, McMurray JJ, Collier TJ. Making sense of statistics in clinical trial reports: part 1 of a 4-part series on statistics for clinical trials. *J Am Coll Cardiol* 2015;**66**: 2536–2549.
4. Sterne JA, Davey Smith G. Sifting the evidence-what's wrong with significance tests? *BMJ* 2001;**322**:226–231.
5. Rossello X, González-Del-Hoyo M. Survival analyses in cardiovascular research, part I: the essentials. *Rev Esp Cardiol (Engl Ed)* 2022;**75**:67–76.