Moving beyond p-value

Augusto Di Castelnuovo,[1] Licia Iacoviello[2,3]

[1]Mediterranea Cardiocentro, Napoli, Italy; [2]Department of Epidemiology and Prevention, IRCCS Neuromed, Pozzilli, Italy; [3]Department of Medicine and Surgery, University of Insubria, Varese, Italy

INTRODUCTION

Scientific literature is overflowing of significance testing and p-values.[1] P-value states how discordant the observed finding is with a null hypothesis. P<0.05 indicates that an association greater than that detected would happen less than 5% of the time under a null hypothesis of no association. As a widespread belief, it is thought that p-value provides the probability that chance alone produced the detected association. This is not true.[1] P-value is the probability of the data (D) assuming that the null hypothesis (H) is true. In formal statistical language, p-value=p(D|H). The probability that chance alone creates the observed association is p(H|D) (=the probability of the null hypothesis given the data).[2]

Imagine you obtained a result for a coagulation test associated with thrombosis. What you now want to know is the probability that given the finding ("the test is associated"), the null hypothesis ("there is no association") is true; this probability is also known as "False Positive Risk" and is equal to p(H|D). Lower this probability is, more confident you are in the correctness of the test conclusion.

I am afraid that a lot of us, including myself, have trusted in the past that p-values from our tests had provided such "reassuring" probability. Unfortunately, this is not the case: p-value is not the false positive risk.

This is not the unique misuse/misinterpretation of the p-value.[1]

Correspondence: Augusto Di Castelnuovo, Mediterranea Cardiocentro, Via Orazio, 2, 80122 Napoli, Italy.
E-mail: dicastel@ngi.it

Key words: Editorial; p-value.

Contributions: The authors contributed equally.

Conflict of interest: The authors declare no potential conflict of interest.

Funding: None.

Received for publication: 24 March 2022.
Accepted for publication: 29 March 2022.

Licensee PAGEPress, Italy
Bleeding, Thrombosis and Vascular Biology 2022; 1:30
doi:10.4081/btvb.2022.30

STATISTICAL INFERENCE IN THE 21st CENTURY

In 2019, the American Statistical Association published a special issue containing 43 papers,[3] which exhaustively discussed the topic, and tried to provide alternatives to go beyond p-value. In an accompanying editorial,[4] the perils of misuse and misinterpretation of p-values and significance testing were well expressed: a) don't base your decisions merely on whether or not an association or an effect was found to be "statistically significant"; b) don't believe that an association exists (or is null) just because it was (it was not) statistically significant; c) don't believe that your p-value gives you the probability that chance alone produced the observed association; d) don't conclude anything about scientific or practical importance of your data only based on statistical significance (or lack thereof).

Following these convincing and authoritative pronouncements by statistician's community, several Journals changed their guidelines for statistical reporting.[5,6] A firm claim for retiring of statistical significance characterises these updating.

This is also the line of this editorial, which sketches the statistical guidelines of Bleeding, Thrombosis and Vascular Biology Journal.

DON'T

What not to do? We have never to assume that there is 'no association' or 'no difference' or 'no effect' just because a p-value is larger than a threshold such as 0.05 (or, in the same way, because a confidence interval includes zero or one –depending on the metric).[3]

We have not to conclude that two studies are in conflict because one had a statistically significant result, and the other did not. For example, the risk factor *alpha* is associated with a certain outcome with an identical point estimate in both studies A and B (Figure 1), and confidence

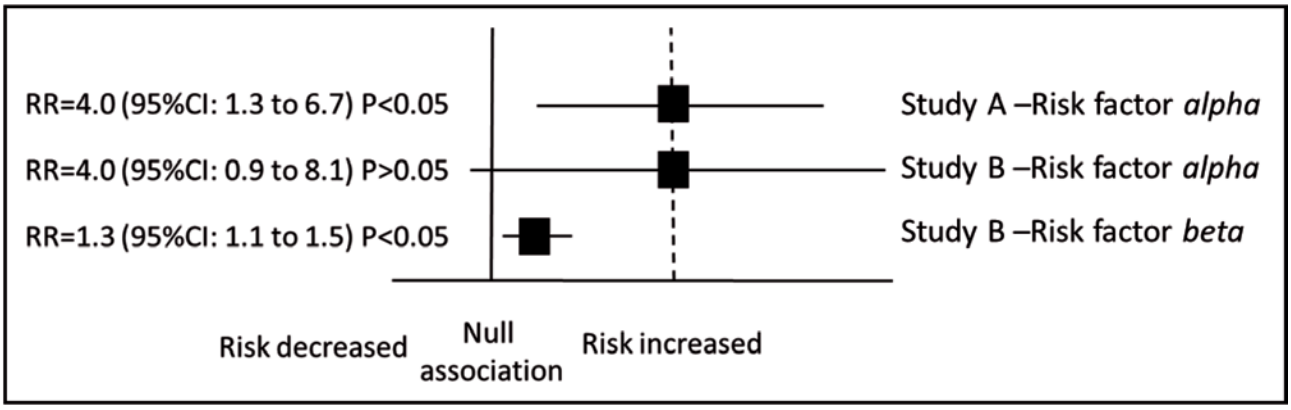

Figure 1. Relative risks and 95% confidence intervals (invented data) for two risk factors (*alpha* and *beta*) in two different studies (A and B). Black squares represent the relative risk (RR); horizontal lines denote the 95% confidence intervals (95%CI).

intervals largely overlap: the two studies are not divergent in their conclusions, even if the association is 'significant' in study A but not in study B.

We have to stop the use of p-values in a dichotomous way and must retire statistical significance.[6]

DO

Unfortunately, there is no single solution for awesomely replacing of significance testing and p-value. Several alternatives have been proposed, and it is expected that many of them will be increasingly used for statistical inference in scientific research. Detailed discussion of these alternatives is outside the scope of this editorial. The reader will find plenty of information in other sources.[3,4]

MEASURE OF THE EFFECT AND HYPOTHESIS TESTING

A vital thing to argue is that the measure of the effect is more important that the hypothesis testing. Magnitude, precision, direction, plausibility, consistency, repeatability and clinical or practical utility have to be the key features to be investigated for an effect (or association, or difference), much more than a p<0.05. Note that only precision and, to a lesser extent, magnitude, are linked to p-values.

We should be more confident with uncertainty.[4] We have to report point estimate (=observed effect) and confidence intervals, and describe the practical inferences suggested by all values within the interval, especially the point estimate and the limits. We must become familiar with the fact that all values within the interval are compatible with the data. For example, in the study B (Figure), the risk factor *beta* is associated with an increased risk (for a certain outcome) ranging from 10% to 50% (point estimate equal to 30%). In the same study and for the same outcome, the risk factor *alpha* (observed effect equal to 300%) is compatible with: a) a risk greater than 50% for the large majority of the interval; b) a risk between 0% to 50% for another portion of the interval and c) a null risk or protection for a very small fraction of the interval. Concluding that *beta* is a risk factor since p<0.05 but *alpha* is not as p>0.05 is unreasonable.

The magnitude of the effect (or association or difference) is very important. Focusing on statistically significant but small effects (therefore most of the times having negligible clinical impact) and ignoring large effects (potentially of clinical interest) because the latter are not statistically significant is an improper approach.

A FINAL QUESTION

If I were on a crashing airplane, I would rely more on a pilot's manoeuvre that would reduce the risk of crashing by a non-statistically significant 90% rather than one that would reduce the risk by a significant 10%.

And you?

REFERENCES

1. Wasserstein R, Lazar N. The ASA's statement on p-values: context, process, and purpose. Am Stat 2016;70:129-33
2. Krueger JI, Heck PR. Putting the p-value in its place. Am Stat 2019;73:122-8.
3. Statistical inference in the 21st Century: a world beyond p<0.05. Am Stat 2019;73.
4. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond p<0.05. Am Stat 2019;73:1-19.
5. Harrington D, D'Agostino RB Sr, Gatsonis C, et al. New guidelines for statistical reporting in the journal. N Engl J Med 2019;381:285-6.
6. Amrhein V, Greenland S, McShane B. Retire statistical significance. Nature 2019;567:305-7.