To P or not to P. That is the wrong question. Suggestions on reporting uncertainty

A few years ago, an article in the first issue of *Bleeding, Thrombosis and Vascular Biology*, argued that biomedical research should move beyond the mechanical use of the p-value.[1] That message was not new, but it was timely. It remains timely today.[2]

The p-value is still often treated as a gatekeeper: below 0.05, a finding is "real"; above 0.05, it is "negative". This habit is simple, familiar, and convenient. It is also misleading. A p-value does not tell us the probability that the null hypothesis is true. It does not tell us the probability that chance alone produced the observed association. It does not measure the size, importance, credibility, or clinical relevance of an effect.[3] It is a measure of how incompatible the observed data are with a statistical model, usually a null model, under a set of assumptions. Greenland and colleagues catalogued 25 distinct misinterpretations of p-values, confidence intervals, and power that remain widespread in the literature, and many of them survive precisely because the threshold habit substitutes for careful reading of numbers.[4]

This distinction is not semantic. It affects how papers are written, reviewed, read, and remembered.

In vascular biology, thrombosis, hemostasis, and cardiovascular medicine, many studies are observational, mechanistic, exploratory, translational, or based on relatively small clinical samples. In these settings, the overuse of "statistical significance" can distort interpretation in both directions. A small effect, precisely estimated in a large dataset, may achieve a very low p-value but have little clinical meaning. A larger effect, estimated with uncertainty in a smaller study, may fail to cross the 0.05 threshold but still be biologically plausible and worth reporting carefully. The width of the confidence interval in this second scenario is also informative about study precision: a very wide interval may reflect limited sample size, few events, or design constraints, and this limitation deserves explicit discussion, not just a note that the result was "not significant".

For this reason, we strongly suggest that Authors submitting to* Bleeding, Thrombosis and Vascular Biology *use the p-value with restraint and avoid treating it as the main result of a study.[5] A suggestion extended to Section Editors and Reviewers.

The primary object of interpretation should be the effect estimate: a risk ratio, hazard ratio, odds ratio, mean difference, regression coefficient, or other measure appropriate to the research question. The estimate should be reported with its confidence interval, and the interval should be interpreted as a range of values reasonably compatible with the data, not as a device for declaring success or failure.[6]

Several simple rules may help.

1. Do not write that there is "no association," "no difference," or "no effect" only because $p>0.05$. A result with a hazard ratio of 1.50 and a 95% confidence interval from 0.95 to 2.05 is not evidence of absence. The point estimate of 1.50 is the value most compatible with the observed data; values closer to 1.50 are more compatible than values near the interval boundaries, and this compatibility decreases progressively toward the extremes. While the lower bound of 0.95 means that a hazard ratio below 1.0 cannot be excluded, such values sit at the tail of the compatibility range and are far less supported by the data than values above 1.0. The result is therefore better described as an imprecise but directionally informative estimate, pointing toward increased risk, than as a null finding.
2. Do not claim that two studies disagree only because one p-value is below 0.05 and the other one is above 0.05. Two studies may have similar point estimates and overlapping confidence intervals even if only one is conventionally "significant." In such cases, the data may be largely consistent. The apparent conflict is created by the threshold, not by the evidence.
3. Avoid using "significant" as a synonym for "important." Statistical significance and scientific relevance are different ideas.[7] The practical meaning of a result depends on the magnitude of the effect, the outcome under study, the quality of the design, the risk of bias, biological plausibility, consistency with previous evidence, and the potential clinical or public-health consequences.
4. Report exact p-values when they are useful, but avoid decorating results with stars, bold type, or labels that divide findings into "significant" and "non-significant." A table should not guide the reader's eye only toward $p<0.05$. It should help the reader understand the pattern, size, and uncertainty of the findings. This practice, now adopted as formal policy by several journals, also discourages selective emphasis on results that cross an arbitrary threshold.
5. Distinguish *prespecified* analyses from *exploratory* analyses. In confirmatory settings, especially clinical trials or analyses with a clear primary hypothesis and a prespecified statistical plan, p-values may have a defined role, particularly when type I error and multiplicity have been properly addressed. In secondary, subgroup, biomarker, omics, or hypothesis-generating analyses, p-values should be interpreted more cautiously. Without a clear plan for multiplicity, isolated $p<0.05$ findings should not be presented as definitive evidence. A related problem, worth naming explicitly, is selective reporting within a paper: presenting only the subset of results that reached a threshold, while suppressing or minimizing those that did not, distorts the picture of the evidence just as much as publication bias does at the journal level.
6. Describe uncertainty in plain scientific language. Instead of writing "the association was not statistically significant," Authors might write: "The estimate suggested a higher risk, but the confidence interval was wide and included both no clear association and a clinically relevant increase." Instead of writing "there was a significant protective effect," they might write: "The estimate suggested a lower risk; the confidence interval was compatible with a modest to moderate reduction." These formulations are longer, but they are more honest and more accurate.
7. In selected settings, authors may consider complementing p-values with other ways of expressing evidence. *S-values,* for example, translate p-values into bits of information against the null model, and *Bayesian approaches* can make the role of prior evidence

explicit.[8] These tools are not necessary in most papers, and they do not solve poor design or selective analysis. Their value is greatest when they help readers understand the strength and limits of the evidence without returning to a binary threshold.

8. Consider multiplicity correction when the number of simultaneous comparisons is large, the tests are correlated, and all are performed on the same dataset, as is typical in omics, genome-wide, or high-dimensional biomarker analyses. In these settings, the probability of at least one false positive under the null hypothesis becomes substantial, and methods such as *Bonferroni correction* or *Benjamini-Hochberg* false discovery rate control provide a useful safeguard. What matters more than the correction itself is transparency: Authors should state clearly how many comparisons were made, which were prespecified and which were not, and interpret each result in the context of the full set of analyses. A corrected p-value does not validate a given finding, it only makes a chance artifact less likely.

The limits of threshold-based thinking become particularly clear when comparing two associations. Suppose a study reports a hazard ratio of 1.35 (95% CI: 1.02-1.79, p=0.037) in one subgroup and a hazard ratio of 1.28 (95% CI: 0.96-1.70, p=0.091) in another. The first would typically be labelled "significant," the second "non-significant," and the two presented as qualitatively different. Yet the point estimates are close and the confidence intervals overlap substantially: a formal test for the difference between the two HRs yields a p=0.78. The apparent contrast is an artifact of the threshold. The quantity that actually matters here is the difference between the two hazard ratios, and that difference is small, imprecisely estimated, and compatible with chance. Gelman and Loken called this the *difference between significant and non-significant is not itself a statistically significant* problem.[9] Formally testing the interaction does not mean falling back on a pass/fail judgment: it means asking the right question, estimating the right quantity, and reporting the result, including its uncertainty, rather than inferring heterogeneity from a comparison of two arbitrary labels.

None of these points is an invitation to lower scientific standards. It is just the opposite. Moving away from automatic threshold-based language requires Authors to think more, not less. It requires attention to study design, measurement error, confounding, selection bias, model assumptions, missing data, multiplicity, and prior evidence. A small p-value cannot repair a weak design. A large p-value cannot make an imprecise but potentially important finding disappear.

Nor is this Editorial a call to ban p-values. Used with care, they may be certainly useful. The problem is not the p-value itself, but the excessive weight placed on one arbitrary threshold. The phrase *statistically significant* often gives a false impression of certainty, while *non-significant* often suggests absence of evidence when the study may simply be underpowered or imprecise.

For *Bleeding, Thrombosis and Vascular Biology*, a reasonable standard is quite simple: report estimates, report uncertainty, avoid dichotomous language, and interpret results in the light of biology and design. Authors should make clear what their data suggest, what they do not show, and how much uncertainty remains.

Science rarely advances by asking whether p is just below or just above 0.05. It advances by estimating effects, judging their credibility, comparing them with prior knowledge, and deciding whether they matter. That is the kind of statistical reporting we strongly encourage.

REFERENCES

1. Di Castelnuovo A, Iacoviello L. Moving beyond p-value. Bleeding Thromb Vasc Biol 2022;1:30.
2. van Zwet E, Gelman A, Greenland S, et al. A new look at P values for randomized clinical trials. NEJM Evid 2024;3:EVIDoa2300003.
3. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond p<0.05. Am Stat. 2019;73:1-19.
4. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 2016;31:337-50.
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.
7. Greenland S, Mansournia MA, Joffe M. To curb research misreporting, replace significance and confidence by compatibility: a preventive medicine golden Jubilee article. Prev Med 2022;164:107127.
8. Greenland S. Valid p-values behave exactly as they should: some misleading criticisms of p-values and their resolution with s-values. Am Stat 2019;73:106-14.
9. Gelman A, Stern H. The difference between "significant" and "not significant" is not itself statistically significant. American Statistician 2006;60:328-31.

Augusto Di Castelnuovo
Research Unit of Epidemiology and Prevention, IRCCS Neuromed, Pozzilli (IS), Italy

Giovanni de Gaetano
Editor-in-Chief, *Bleeding, Thrombosis and Vascular Biology*