Why not to (over)emphasize statistical significance

Olaf M Dekkers

Leiden University Medical Centre, Leiden, the Netherlands

Correspondence
should be addressed
to O M Dekkers

Email
o.m.dekkers@lumc.nl

Abstract

P values should not merely be used to categorize results into significant and non-significant. This practice disregards clinical relevance, confounds non-significance with no effect and underestimates the likelihood of false-positive results. Better than to use the *P* value as a dichotomizing instrument, the *P* values and the confidence intervals around effect estimates can be used to put research findings in a context, thereby taking clinical relevance but also uncertainty genuinely into account.

*European Journal of
Endocrinology*
(2019) **181**, E1–E2

In a recent *Nature* paper, leading methodologists called for a ‘retirement of statistical significance’, arguing that researchers should not mention in their scientific paper whether a statistical test result is (non)significant (1). To underline the urgency of the message, the call was signed by more than 800 researchers. The call to abandon statistical significance as a marker for the presence/absence of an effect is not new, but clearly unsuccessful thus far (2).

To see why these authors argue to abandon reference to statistical significance, let us start with three hypothetical studies (Fig. 1). The studies compare the same new (and likely very fancy) drug to placebo for treatment of Cushing’s disease, the outcome being biochemical cure. The effect estimate is expressed as hazard ratio (HR) with accompanying 95% confidence intervals (95% CI). For the sake of the argument, we assume that the three studies are unbiased. All studies find the same effect estimate, an HR of 2.0, and thus suggest that the new drug doubles cure probability. However, the three studies differ in cohort size, study 1 being the largest, study 3 the smallest. As study size is inversely related to the width of the confidence interval, the confidence interval for study 3 is the widest. The 95% confidence interval from study 3 crosses the line of no effect (HR of 1.0), the *P* value of study 3 will be >0.05 ; the *P* values from studies 1 and 2 will be <0.05 . If only the *P* value would be used to determine whether

there is an effect, the conclusion from study 3 (‘no effect’) would contradict the conclusion from studies 1 and 2 (‘there is an effect’). That clearly seems odd, and the main source for this apparent contradiction occurs when only *P* values are used to decide whether there is an effect.

As can be seen from the figure, there is a direct connection between study size, confidence interval and *P* value. In smaller studies, confidence intervals get wider and effect estimates are less likely to reach statistical significance. Putting it differently: a small study can only reach statistical significance in case the effect is large. And this is a major disadvantage of the *P* value: it is a measure that relates to both effect size and study size. Studies with the same effect estimate can therefore have different *P* values (as can be seen from the figure); studies with different effect sizes can have the same *P* value if the sample size differs. This shows that a *P* value is never a marker of the magnitude of effect, and in line, never a marker of relevance of an effect.

This does not mean that *P* values should not be used or not reported. The main point is that *P* values should not merely be used to categorize results into significant and non-significant. In research practice, however, *P* values are often used to distinguish true null-hypothesis (‘there is no effect’) from non-true null-hypothesis (‘there is an effect’). This practice has three disadvantages:

Figure 1

Graphical representation of effects of three hypothetical studies.

- *It disregards clinical relevance.* Especially in database-driven studies with huge sample sizes, even the smallest difference between groups can reach statistical significance. Consider a study comparing potassium levels between men and women in a large register-based study; even a difference between 4.01 and 4.02 mmol/L could reach a P value <0.05 if the sample size is large enough. Obviously, the clinical relevance of the difference could be questioned; if however the result section only mentions that the difference is ‘statistically significant’ this clinical irrelevance may go unnoticed.
- *It confounds non-significance with no effect.* Conversely, the confidence interval is wide by default in small studies, and the effect needs to be large to reach significance. Take for example study 3 in the table. Claiming no effect because the P value is >0.05 , would disregard the huge uncertainty in the effect estimate. This uncertainty is captured in the 95% confidence interval that ranges from 0.75 to 5.33. Claiming no effect because of non-significance, comes down to only highlighting that 1.0 is included in the confidence interval while disregarding that the study cannot rule out a large effect. So, absence of evidence is not evidence of absence, especially for smaller studies (3). The practice of clinical research however shows that P values >0.05 are often used to substantiate the claim of no effect. In an analysis based on 791 articles, 51% wrongly assumed that ‘non-significant’ equals ‘no effect’ (1).

- *It underestimates the likelihood of false-positive results.* Although very often used and reported, most researchers do not have a clear formal understanding of P values (4). Most often, a P value of 0.05 is loosely interpreted as ‘a 5% probability that the null-hypothesis is correct’, and thus a 95% probability that there is actually an effect. Without going into technical details (see (1, 2) for details), this incorrect interpretation of the significant P value largely overestimates the P value as marker of truth. It can be shown that the probability that the null-hypothesis is true (‘there is no effect’) in an observational study with $P < 0.05$ is closer to 50% than to 5% (5). This means that researchers should be aware that the risk of false-positive claims is considerable despite a significant P value.

There are thus many reasons not to overestimate the role of P values in clinical research. Better than to use the P value as a dichotomizing instrument, discerning true findings from null-findings, the P values and the confidence intervals around effect estimates can be used to put research findings in a context, thereby taking clinical relevance but also uncertainty genuinely into account.

Declaration of interest

Olaf M Dekkers is a Deputy Editor for *European Journal of Endocrinology*. O M D was not involved in the peer review or editorial process for this paper, on which he is listed as an author.

Funding

This research did not receive any specific grant from any funding agency in the public, commercial or not-for-profit sector.

References

- 1 Amrhein V, Greenland S & McShane B. Scientists rise up against statistical significance. *Nature* 2019 **567** 305–307. (<https://doi.org/10.1038/d41586-019-00857-9>)
- 2 Sterne JA & Davey Smith G. Sifting the evidence-what’s wrong with significance tests? *BMJ* 2001 **322** 226–231. (<https://doi.org/10.1136/bmj.322.7280.226>)
- 3 Altman DG & Bland JM. Absence of evidence is not evidence of absence. *BMJ* 1995 **311** 485. (<https://doi.org/10.1136/bmj.311.7003.485>)
- 4 Rosendaal FR. The p-value: a Clinician’s disease? *European Journal of Internal Medicine* 2016 **35** 20–23. (<https://doi.org/10.1016/j.ejim.2016.08.015>)
- 5 Ioannidis JP. Why most published research findings are false. *PLoS Medicine* 2005 **2** e124. (<https://doi.org/10.1371/journal.pmed.0020124>)

Received 11 July 2019

Accepted 12 July 2019