Biometrical Journal **xx** (2017) xx, zzz–zzz / DOI: 10.1002/bimj.200100000

Contribution to the discussion of the paper by Stefan Wellek: A critical evaluation of the current p-value controversy

Alessio Farcomeni [*,1]

[1] Department of Public Health and Infectious Diseases *(Sapienza - University of Rome)*
Piazzale Aldo Moro, 5
00185 Roma
Italy

Received zzz, revised zzz, accepted zzz

Computation and thresholding of significance levels is ubiquitous in applied research, but p-values are often poorly interpreted. Wellek (2017) is very welcome in providing biostatisticians' view on the p-value debate. Here I will try to give additional remarks on the p-values versus measures of effect size (Section 2.2) issue and on the issue of multiplicity; on one particularly common and deleterious misuse of p-values in applied research; and finally point out one underused flexibility regarding p-value interpretation. First of all, it is undoubtful that applied research has been overfocused on statistical significance and has neglected biological significance. Not only small p-values have been interpreted as measures of effect size. Effect sizes might sometimes have been correctly reported and interpreted, but they were negligible. Several clinical trials have been published, and drugs approved, where the significant improvement in terms of median mortality was just few days. This maybe together with high costs for the experimental treatment and no improvements in terms of patient comfort and quality of life. For this reason, even if $p < 0.05$, a finding should be discarded if the effect size does not simultaneously exceed a pre-fixed biological significance threshold. This is very hard to fix, being context-dependent, but medical societies could prepare careful statements establishing minimal thresholds. If simultaneous statistical *and* biological significance becomes a requirement, research might report more useful and more likely replicable findings. Testing methods balancing effect size and statistical significance are proposed in Finos and Farcomeni (2011) and Farcomeni and Finos (2013). One very frequent misuse of p-values is, in my opinion, the interpretation of lack of statistical significance as evidence in favor of the null model (acceptance of the null hypothesis). Equality can be defined only up to a margin. Once again, medical societies could issue statements about recommended margins of equivalence and of non-inferiority/non-superiority. Well defined thresholds for biological significance can serve two scopes: that of defining meaningful findings (when $p < 0.05$ and the effect size exceeds a pre-specified threshold) and that of defining equivalence (when $p > 0.05$, sample size is large enough to claim sufficient power at the margin, and the effect size does not exceed the threshold). In his very careful list of suggestions, Stefan Wellek recommends reporting multiplicity and a list of what p-values are used for testing and what only for descriptive purposes. This is connected to the widespread practice of p-hacking: multiple testing without adjustment and reporting of the multiplicity. Researchers might explore the data they have recorded by evaluating several associations, and then report only one or two significance tests in their paper. The practice of p-hacking decreases the chances of replication, and therefore discredits the scientific community. In order to discourage p-hacking, journals might require that an approved protocol, study plan, and/or grant application is submitted together with papers. This in order to check correspondence between primary objectives and originally planned endpoints. Finally, I would like to point out that in several cases Type III error control is achieved as a free lunch together with Type I control (e.g., Finner (1999) and references therein). When a two-sided hypothesis is rejected and the

*Corresponding author: e-mail: alessio.farcomeni@uniroma1.it

effect size is, say, positive, one can often additionally conclude that data disagree with all models with non-positive effect size. The error rate is unchanged.

Conflict of Interest *The author declares no conflict of interest*

References

Farcomeni, A. and Finos, L. (2013). FDR Control with Pseudo-Gatekeeping Based on a Possibly Data Driven Order of the Hypotheses. *Biometrics* **69**, 606–613.

Finner, H. (1999). Stepwise multiple test procedures and control of directional errors. *Annals of Statistics* pages 274–289.

Finos, L. and Farcomeni, A. (2011). k-FWER control without multiplicity correction, with application to detection of genetic determinants of multiple sclerosis in Italian twins. *Biometrics* **67**, 174–181.

Wellek, S. (2017). A critical evaluation of the current p-value controversy. *Biometrical Journal* page in press.

