Letters to the Editors

Changing the traditional *p*-value significance threshold to .005 does not decrease the number of statistically significant *p*-values in observational studies more than in randomised controlled trials

Sirs,

Changing the *p*-value significance level to 0.005 had been proposed to increase research reproducibility (1, 2). It was then reported this would decrease the number of significant *p*-values in randomised controlled trials (RCTs) by 29.3% (3). Later, a 49.0% decrease was observed among the RCTs in orthopedics (4). Being unaware of similar data among observational studies (OSs) where we hypothesised, with usually less stringent methodology, the effect of this change would be more pronounced.

We surveyed how the 0.05 to 0.005 threshold change would affect the primary end point related *p*-values in the abstracts of RCTs and OSs in the same journals in reference 3 (*New England Journal of Medicine, Lancet, JAMA* in 2017) and in another year, 2002. The additional distant year was arbitrarily chosen to assess reproducibility of, and to seek perhaps a temporal trend in, our finding. RCTs were also surveyed for both time periods. RCTs with pooled analyses, Bayesian or non-inferiority analyses and OSs reporting genetic associations were excluded. Our primary end point was the degree to which changing the *p*-value threshold would affect the number of significant *p*-values related to the primary study outcomes in RCTs *versus* in OSs. Comparisons were made by ORs with 95% CIs. *Post hoc* analyses were done to compare the number of significant *p*-values per manuscript for either study type at both years compared by Cohen's *d*. Sample sizes, taken as the total number of individuals studied in all groups, were tabulated as medians and IQRs.

M.O and Z.T.D identified and analysed the studies, disagreements were settled by H.Y. SPSS 21.0 software was used for statistical analyses.The threshold change decreased the number of significant *p*-values among the RCTs less than among the OSs by an OR of 0.72 (95% CI, 0.44 to 1.17) in 2017 and by an OR of 0.68 (95% CI, 0.49 to 0.96) in 2002. The proportion of statistically significant *p*-values were higher among the OSs at both years and at both significance thresholds (Table I). Also, the median samples sizes were noticeably larger among the OSs in either year (Table I). The decrease in the number of significant *p*-values in 2017 was similar between what was reported before (3) and the current study (29.3 *vs*. 30.6%) (Table I), supporting the validity of surveying abstracts only. Contrary to our hypothesis, changing the *p*-value threshold did not decrease the number of significant *p*-values more among the OSs. In fact, there was indication that,with the threshold change, the number of significant *p*-values decreased more among the RCTs in both years (Table I).

Our observation that the frequencies of both all and statistically significant *p*-values at either threshold were higher among the OSs than that among the RCTs for both years (Table I) might reflect less strictly defined multiple primary end points in OSs, yielding perhaps a publication bias with a greater number of significant *p*-values, and The practical necessity of limiting sample sizes in RCTs would also yield comparatively larger sample sizes among the OSs, in turn yielding a higher number of significant *p*-values. An important limitation of our work was that we surveyed only 3, high impact general medicine journals publishing studies with superior methodology and analyses. A similar survey among journals with lesser impact factors might well show considerably more decreases in the number of *p*-values declared significant both in RCTs and OSs. However, based on the results of this survey, we proposethat the effect of such a change will not be more pronounced, even then, among the OSs.

Acknowledgments

We thank Prof. Maarten Boers for his helpful comments during the manuscript preparation.

M. OZTAS[1] *MD*
Z.T. DINCER[1], *MD*
N. SUT[2], *MD*
H. YAZICI[3], *MD*

[1]Department of Medicine, Division of Rheumatology, Istanbul University-Cerrahpasa, Cerrahpasa Medical Faculty, Istanbul, Turkey;
[2]Department of Biostatistics, Trakya Medical Faculty, Trakya University, Edirne, Turkey;
[3]Academic Hospital, Istanbul, Turkey.

Please address correspondence to:
Mert Oztas,
Istanbul University-Cerrahpasa,
Cerrahpasa Medical Faculty,
Department of Medicine,
İUC. Cerrahpasa Tip Fakultesi Yerleskesi,
Kocamustafapasa Cd. No: 53 Cerrahpasa,
34098 Fatih/Istanbul, Turkey.
E-mail: dr.mertoztas@gmail.com

Competing interests: none declared.

References

1. BENJAMIN DJ, BERGER JO, JOHANNESSON M *et al.*: Redefine statistical significance. *Nat Hum Behav* 2018; 2: 6-10.
2. IOANNIDIS JPA: The proposal to lower p value thresholds to .005. *JAMA* 2018; 319: 1429-30.
3. WAYANT C, SCOTT J, VASSAR M: Evaluation of lowering the p value threshold for statistical significance from .05 to .005 in previously published randomized clinical trials in major medical journals. *JAMA* 2018; 320: 1813-5.
4. JOHNSON AL, EVANS S, CHECKKETS JX *et al.*: Effects of a proposal to alter the statistical significance threshold on previously published orthopaedic trauma randomized controlled trials. *Injury* 2019; 50: 1934-7.

Table I. Differences in the number of *p*-values at 0.05 and 0.005 thresholds in RCTs and OSs.

		RCTs	OSs	Effect sizes
2017	Number of articles	191	61	–
	Total number of *p*-values related to primary outcome(s)	254	169	–
	p ≤0.05 n (%)	173 (68.1)	145 (85.7)	–
	p ≤0.005 n (%)	120 (47.2)	113 (66.8)	–
	Reduction in the number of significant *p*-values* (%)	53 (30.6)	32 (22.1)	OR, 0.72; 95% CI, 0.44 to 1.17
	Mean number of *p*-values per article, n ± SD	1.3 ± 0.8	2.7 ± 1.9	Cohen's *d*=0.96
	Mean number of significant† *p*-values per article, n ± SD	0.9 ± 0.9	2.3 ± 1.6	Cohen's *d*=1.07
	Median number of sample sizes related to *p*-values, (IQR)	573 (293-1307)	6879 (570-165561)	
2002	Number of articles	169	100	
	Total number of *p*-values related to primary outcome(s)	322	278	
	p ≤0.05 n (%)	239 (74.2)	248 (89.2)	
	p ≤0.005 n (%)	127 (39.4)	168 (67.7)	
	Reduction in the number of significant *p*-values* (%)	112 (46.9)	80 (32.2)	OR, 0.68; 95% CI, 0.49 to 0.96
	Mean *p*-value number per article, n ± SD	1.9 ± 1.5	2.7 ± 2.3	Cohen's *d*=0.41
	Mean significant† *p*-value number per article, n ± SD	1.4 ± 1.4	2.4 ± 2.1	Cohen's *d*=0.56
	Median number of sample sizes related to *p*-values, (IQR)	135 (38 – 249)	2807** (688 – 19238)	

*(Number of *p*-values ≤0.05 - number of *p*-values ≤0.005) / Number of *p*-values ≤0.05. † all *p*-values ≤0.05.
**Sample sizes of the 3 primary endpoints were not well defined in one epidemiologic study.