Acknowledged statistical help and a better use of *p*-values: a proposal

E. Dincses[1], G. Guzelant[1],
G. Hatemi[1], N. Sut[2], H. Yazici[3]

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

Elif Dincses, MD*
Gul Guzelant, MD*
Gulen Hatemi, MD
Necdet Sut, MD
Hasan Yazici, MD

**These authors contributed equally to this study.*

Please address correspondence to:
Prof. Hasan Yazici,
Department of Rheumatology -
Internal Medicine,
Academic Hospital,
Uskudar, Istanbul 34668, Turkey.
E-mail: hasan@yazici.net

Received on November 11, 2018; accepted in revised form on March 26, 2019.

Key words: *p*-value, acknowledged statistical help, effect size, confidence interval

Competing interests: none declared.

ABSTRACT

Objective. *The p-value is commonly misused. We hypothesised that* a close cooperation *with a statistician would go along with a more proper use of p-values. We considered a close cooperation present, when a statistician was a co-author, or a formal statistical help was acknowledged in a study report.*

Methods. *Randomised controlled trials published in 2015-16 in 4 widely read rheumatology journals were searched for a close cooperation with a statistician, the inclusion of effect sizes, confidence intervals, exact rather than relative p-values and the omission of p-values in tables depicting trial entry data.*

Results. *There were only 28/133 (21%) articles in which a formal statistical help was acknowledged (Group I). The rest (Group II) gave no acknowledgement of a close cooperation. Reporting of effect sizes (96% vs. 71%) and exact p-values (88% vs. 69%) were more in Group I (p=0.01, and p=0.08, respectively).*

Conclusion. *While a formal acknowledgement of a close cooperation was notably infrequent at 21%, this went along with improvement in some aspects of p-value reporting. If substantiated by further studies, we propose that a formally acknowledged statistical help should improve p-value reporting. Like all professionals, statisticians would like their name/office to be formally associated with their good work.*

Introduction

P-values are commonly misused and misunderstood (1). A *p*-value is the probability of getting a difference or association at least as large as the one observed in a study under the assumption that no such difference/association was present in the first place, the so called null hypothesis (2). It only gives us the probability of the study results being due to sheer chance (a Type 1 or alpha error) like in tossing a fair coin.

Based solely on the centuries old theory of big numbers (3), the *p*-values are very dependent on sample sizes and what we are given as a very statistically significant *p*-value could be clinically most unimportant with very large sample sizes. In order to circumvent this, it is essential the investigators also give the effect sizes (4). The converse is also true where inadequate sample sizes can give us statistically non-significant results (a Type II error) or more importantly, *fickle p-values* (5).

It stands to reason that active participation of a formal statistician in reporting scientific work would bring better use. With this work we aimed to test the hypothesis that a closer cooperation with a statistician would be associated with improved *p*-value reporting with special emphasis on effect sizes, the reporting of which is mandatory to appreciate the numerical and clinical importance of a statistical significance or a non-significance we are reporting.

Materials and methods

Two observers (ED and GG) screened, both by reading and electronic scanning, the full-texts and supplementary materials of all the randomized controlled trials (RCT) published in 4 widely read rheumatology journals over a 2-year period (2015-2016) Our choice of 4 widely read journals was mainly based on our individual judgement of their impact combined with their likelihood of publishing RCTs. Thus we included Annals of the Rheumatic Diseases (ARD), Arthritis Care and Research (ACR), Arthritis and Rheumatology (A&R) and Rheumatology Oxford (RO) in our survey.

We limited our survey to only the *p*-values associated with the main outcome results of RCTs. Furthermore, we surveyed in addition to the effect sizes, our main emphasis, only two other issues related to p value reporting among the rather long list (3). We defined the close cooperation as the inclusion of a statistician among the co-authors and/or a declaration of formal statistical help in the studies surveyed.

We specifically tabulated: a. whether the RCT was a drug trial or not, and whether it was industry sponsored in either case; b. the presence of a statistician among the co-authors – a co-author who had a work address in statistics or a related unit – and/or formal acknowledged statistical help in the methods; c. the inclusion of effect sizes for the primary outcome (at least for 1 primary outcome if there were several) and the associated confidence intervals (CI). Also tabulated were giving relative (*p*< or *p*>) instead of exact (*p*=) *p*-values and the erroneous inclusion of *p*-values in tables depicting trial entry data in RCTs since these tables, by definition, display randomised features (4).

For effect size reporting we also tabulated: a. whether effect sizes were not given at all; b. whether the effect sizes were specifically indicated as such or c. whether they could be calculated from the data presented (5).

We classified the articles into two groups according to their acknowledged collaboration with or without a statistician (Groups I and II). We compared the frequency of the aforementioned variables between the two groups using the Yates or Fisher's exact chi-square tests. An arbiter (HY) decided the final tabulation when there were conflicts between the 2 observers not reconciled among themselves. The Mantel-Haenszel test was used for comparison of groups (statistician *vs.* non-statistician) and trial types (drug *vs.* non-drug), controlling for the third confounding factor, the sponsor (industry sponsored, not industry sponsored). Also, odds ratios (OR) and 95% CI were calculated.

Results

The total number of RCTs published in these 4 journals was 133 (62 in ARD, 21 in ACR, 32 in A&R and 18 in RO) (Supplement). The arbiter made the final decision in 23 conflicts between the 2 observers. 20/23 (87%) of the discrepancies were related to not classifying studies that reported secondary outcomes of RCTs such as quality of life that were published as a separate article as a drug-trial. None of the conflicts were related to the presence of acknowledged statistical help (Supplementary data).

Acknowledgement of formal statistical help

In Group I 28 (21%) RCTs a formal help was acknowledged. Among these 28 RCTs, a statistician was a co-author in 25 and statistical help was acknowledged in the text in 3. In Group II 105 (79%) RCTs no statistical help was acknowledged.

Table I. Distribution of Group I and Group II trials and being a drug or non-drug trial, industry sponsored or not.

	Group I (Statistician)	Group II (Non-statistician)
Drug trials (n=98)		
Industry sponsored	15	60
Not industry sponsored	8	15
Non-drug trials (n=35)		
Industry sponsored	0	4
Not industry sponsored	5	26

Table II. The differences in *p*-value reporting between Group I and Group II.

	Group I (Statistician) (n=28)	Group II (Non-statistician) (n=105)	Differences between Group I and Group II
Effect size reporting, n (%)	27 (96)	75 (71)	OR=10.8 (95% CI 1.4-83)[¶]
Given directly, n (%)	2 (7)	5 (5)	OR=1.5 (95% CI 0.2-8.3)
Can be calculated (given HR, OR, RR, ß coefficient), n (%)	25 (89)	70 (67)	OR=4.1 (95% CI 1.1-14.7)[#]
Confidence intervals for effect size reporting, n (%)	16 (57)	43 (41)	OR=1.9 (95% CI 0.8-4.4)
Reporting exact *p*-values, n (%)	23/26 (88)	63/91 (69)	OR=3.4 (95% CI 0.9-12.2)[¥]
Inclusion of *p*-values for the baseline data, n (%)	2/26 (8)	20/91 (22)	OR=0.3 (95% CI 0.06-1.3)

[¶]*p*=0.01. [#]*p*=0.03. [¥]*p*=0.08.

Drug, non-drug, sponsored, and non-sponsored trials

The majority of the publications were drug trials (98/133, 74%). 75/98 (76%) of these trials acknowledged a sponsor while this was true for only 4/35 (11%) of the non-drug trials (*p*<0.0001, OR=25.2, 95% CI 8.0–79.1) (Table I).

There were no significant differences in the frequency of drug trials between Group I (23/28, 82%) and Group II (75/105, 71%) (*p*=0.36, OR=1.8, 95% CI 0.6–5.2).

When the number of industry sponsored trials in Group I and Group II were compared using Mantel Haenzel statistic, there was a non-significant trend for more industry sponsored trials in Group II (*p*=0.11, OR=3.3; 95% CI 0.9–11.9).

Frequency of reporting effect sizes, confidence intervals, exact p-values and p-values for baseline randomisation data

Table II shows the association of being a Group I or Group II RCT and the 4 surveyed aspects of *p*-value reporting. We observed that effect sizes and exact *p*-values were more frequently given in Group I. On the other hand, there were no significant differences in CI reporting and *p*-values for baseline data.

Discussion

Our survey showed that the quality of reporting *p*-values was better in those manuscripts where there was an acknowledged statistical help from the standpoints of reporting effect sizes and exact *p*-values while the same could not be said for reporting the CIs of effect sizes. We became aware of one other study which addressed the issue of more proper reporting of statistical parameters with acknowledged help of statisticians. Jaykaran *et al.* had assessed the function of statisticians in clinical trials published in Indian medical journals (6). They concluded that the presence of a statistician did not seem to contribute much to proper reporting of statistical parameters. However, it seemed to be associated with lower sample sizes and somewhat lower primary endpoints. The issue of effect size reporting, which we emphasized in our survey, did not seem to be addressed by Jaykaran *et al.* (6). What is interesting is the small percentage of trials, 13/68 (19%) in which a formal statistical help was not acknowledged. While these authors do not emphasise this issue in their report, the reported percentage is very close to what we noted, 28/133 (21%) in our work. It should be noted the small sample sizes in the quoted work is a limitation when comparing the findings among the trials with and without statistical help and a similar, perhaps a somewhat less limitation, is also present in ours.

While the main aim of our work was to understand whether acknowledgement of formal statistical help was associated with improved *p*-value reporting, our survey to our surprise showed, as we just referred to, that the relatively small percentage of RCTs had acknowledged statistical help as we defined it. We first analysed whether this acknowledgement depended on being a drug trial and being sponsored. As expected the majority of the drug trials were sponsored (Table I) while this did not seem to significantly affect whether a statistical help was acknowledged.

In addition to the small number of studies in Group I, a further limitation of our work was the scope of possible improvements with formal statistical help. An important issue is a proper study design, a parameter which we did not address in our survey, and to which a statistician would most expectedly contribute. A further issue is the limited scope of our survey to only 4 specific issues related to the proper use of *p*-value reporting. We had chosen to do this both for the conceptual importance of these issues and our expectation of low observer variability in the recognition of their proper use. Finally, we included only 4 major rheumatology journals. Expanding this study to cover an increased number of journals from different fields may increase our understanding of the magnitude of this problem.

Our finding of an undesirable low frequency of a formal acknowledgement of statistical help is important and it leads us to propose a different approach to the current *p*-value misuse. We are aware that expert statistician colleagues do help both in study design and reporting without being acknowledged. The paucity of proper acknowledgement both in our and Jaykaran *et al.* survey also of RCTs (6) is strongly supportive of our awareness and might be due to 1. These colleagues themselves thought their contribution was not of enough importance to be formally acknowledged; 2. The study investigators thought similarly; 3. The statistician colleagues who gave unacknowledged help were not actually experts; 4. They were indeed experts and did recognise the improper use of the *p*-values in the study report. However, they looked the other way, perhaps to please the investigators and the journal editors to publish significant results.

Although it was previously spotlighted, the improper use of the *p*-value still continues (7) and there have been many attempts to improve the situation including stopping their use altogether (8). In these attempts at improvement the approach has mainly been trying to educate the prospective authors and journal editors about the proper *p*-value use. We propose a different approach which is to ask from the journal editors to require from their prospective authors to indicate who was mainly responsible, either as a co-author or in the text, for the statistical analyses. Our hope is that a formal share of burden would lessen the improper use of the *p*-value.

Finally, it has recently been reported that investigators not infrequently "make inappropriate requests to their statistical consultants regarding the analysis and reporting of their data" (9). We envision what we propose would also lessen the amount of compliance of these consultants to such requests in issues outside the improper use of *p*-values.

References

1. GOODMAN SN: Aligning statistical and scientific reasoning. *Science* 2016; 352: 1180-1.
2. ROBINSON DH, WAINER H: On the past and future of null hypothesis significance testing. *J Wildl Manage* 2002; 66: 263-71.
3. 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.
4. SENN S: Testing for baseline balance in clinical trials. *Stat Med* 1994; 13: 1715-26.
5. CHAVALARIAS D, WALLACH JD, LI AH, IOANNIDIS JP: Evolution of reporting p values in the biomedical literature, 1990-2015. *JAMA* 2016; 315: 1141-8.
6. JAYKARAN, CHAVDA N, YADAV P: Proper reporting of statistical parameters in clinical trials published in Indian Medical Journals. Is inclusion of statistician play any significant role? *J Young Pharm* 2011; 3: 167-8.
7. WASSERSTEIN RL, LAZAR NA: The ASA's statement on p-values: context, process, and purpose. *Am Stat* 2016; 70: 129-33.
8. TRAFIMOW D, MARKS M: Editorial. *Basic Appl Soc Psych* 2015; 37: 1-2.
9. WANG MQ, YAN AF, KATZ RV: Researcher requests for inappropriate analysis and reporting: A U.S. survey of consulting biostatisticians. *Ann Intern Med* 2018; 169: 554-8.