Addressing the Challenge of *P*-Value and Sample Size when the Significance is Borderline: The Test of Random Duplication of Participants as a New Approach

Jose-Gaby Tshikuka[1,2,*], Mgaywa G.M.D. Magafu[1,3], Mooketsi Molefi[1], Tiny Masupe[1], Reginald B. Matchaba-Hove[4], Bontle Mbongwe[4] and Roy Tapera[4]

[1]*Department of Public Health Medicine, Faculty of Medicine, University of Botswana, Private bag 00713, Gaborone, Botswana*

[2]*Department of Health Sciences, National Pedagogic University, Kinshasa I, DRC, Republic of the Congo*

[3]*Department of Global Health, University of Washington, Seattle, USA*

[4]*School of Public Health, Faculty of Health Sciences, University of Botswana, Private Bag 0022, Gaborone, Botswana*

Abstract: The issue of borderline *p*-value seems to divide health scientists into two schools of thought. One school of thought argues that when the *p*-value is greater than or equal to the statistical significance cut-off level of 0.05, it should not be considered statistically significant and the null hypothesis should be accepted no matter how close the *p*-value is to the 0.05. The other school of thought believes that by doing so one might be committing a Type 2 error and possibly missing valuable information. In this paper, we discuss an approach to address this issue and suggest the test of random duplication of participants as a way to interpret study outcomes when the statistical significance is borderline. This discussion shows the irrefutability of the concept of borderline statistical significance, however, it is important that one demonstrates whether a borderline statistical significance is truly borderline or not. Since the absence of statistical significance is not necessarily evidence of absence of effect, one needs to double check if a borderline statistical significance is indeed borderline or not. The *p*-value should not be looked at as a rule of thumb for accepting or rejecting the null hypothesis but rather as a guide for further action or analysis that leads to correct conclusions.

Keywords: *P*-value, Sample Size, Statistical Significance, Borderline Significance, Participant Random Duplication.

INTRODUCTION

The *p*-value that indicates that an effect under study is statistically significant is a value that was established arbitrarily and by convention it should be < 0.05 [1,2]. This means that under the null hypothesis of no association, the probability of observing an effect as large as that found in the study population by chance and by chance alone is less than 5% [1-3]. This is the same as saying that chance is an unlikely explanation of the outcome [1-3]. But if the *p*-value is found to be ≥ 0.05, it is said that chance cannot be excluded as the likely explanation for the outcome, in which case the null hypothesis is not rejected and often the conclusion is that there is no effect or real association [4,5]. That is to say that when interpreting results from a study, we only have one of the two alternative conclusions – either the findings show that the explanatory variable under investigation has a statistically significant effect on the outcome ($p < 0.05$) or the explanatory variable does not have a statistically significant effect on the outcome ($p \geq 0.05$) [1-6].

A number of authors staunchly defend this viewpoint [7,8]. They argue that when the *p*-value is greater than 0.05, the null hypothesis should be simply accepted and that the explanatory variable shows no effect on the outcome and no other logical assertion should be considered. However, recent developments in biostatistics have shown that more experts are increasingly using such terms as “borderline significance”, “approaching significance”, “nearing significance”, and the like for outcomes whose *p*-values are slightly greater than or equal to 0.05 [9-12]. The criteria used to support *p*-values which are slightly greater than 0.05 as of borderline statistical significance or as of no statistical significance at all remain unclear to date and therefore the need for debating and clarifying this issue.

This paper discusses the challenge of *p*-value and sample size when the statistical significance is borderline and lays ground work for scientific reasoning to avoid misinterpretation of *p*-values when they are slightly greater than or equal to the statistical significance cut-off level of 0.05. Correct interpretation will prevent erroneous conclusions and misguided actions that may follow.

*Address correspondence to this author at the Department of Public Health Medicine, Faculty of Medicine, University of Botswana, Private bag 00713, Gaborone, Botswana; Tel: +2673554603;
E-mail: josegaby.tshikuka@mopipi.ub.bw

E-ISSN: 1929-6029/16

THE CONCEPT OF BORDERLINE STATISTICAL SIGNIFICANCE

The concept of borderline statistical significance looks irrefutable because if we consider two *p*-values of 0.6 and 0.06, both are not statistically significant at the conventional *p*-value cut-off of 0.05. The *p*-value of 0.06 is nevertheless closer to achieving statistical significance than the *p*-value of 0.6 and would likely do so if the study sample was made large enough or if the number of events of interest increased. Such *p*-values, in most cases, are indicative of associations but could not achieve statistical significance more likely because the study did not have enough power to detect existing differences. Again, this could possibly be due to inadequate sample size or insufficient outcome events. That is why it is important when planning and designing a study to use the appropriate sample size from the population under study to minimize chances of making a Type 2 error [3,5]. However, as it is usually the case, only a limited number of subjects is available for a study. This may be due to resource or time limitation, ethical issues, or rarity of the condition under study among other reasons [1-6]. All in all, it is recommended that the sample size at hand has enough power to minimize Type 2 error. In other words, it has to be ensured that a statistically significant effect is detected only if it truly exists within the population from which the sample was drawn so that correct inference about the population is made [1-5].

SAMPLE SIZE

Formulas used for sample size calculation are many [13-16]. The choice of a formula to be used depends on several factors including the question under study, study design, type of data to be collected, size of the difference to be detected between groups, the smallest effect of clinical interest and many others [1-5,13-16]. The values to be plugged in the formulas such as the strength of the association between the exposure and the outcome of interest are often obtained from the existing literature [17]. However, sometimes the values do not exist in the literature. In such cases, the values need to be defined by the investigator [17]. Depending on the investigator, the defined magnitude of association in similar study designs will certainly yield different sample sizes yet addressing the same question within the same population. As a result, larger sample sizes are more likely to achieve statistical significance compared to smaller sample sizes drawn from the same study population. This suggests that concluding that there is no statistical significance in a measure of association based solely on the conventional cut-off *p*-value of 0.05 may be misleading. Therefore, the conventional cut-off *p*-value of 0.05 should not be the rule of thumb for making conclusions but rather a guide to further action.

STUDY REPLICATION

Based on the argument made above, when a statistical analysis achieves a *p*-value of 0.06, for instance, the null hypothesis cannot simply be accepted to conclude that there is no association. To ensure that the conclusion is airtight we propose that the analysis be repeated once or twice using slightly larger sample sizes. Nonetheless, too large sample sizes beyond the estimated initial sample size should be avoided to prevent overpowering the study [18-20]. In a case where the lack of statistical significance in the original study was due to a smaller sample size, its replicate with an optimal sample size should undoubtedly achieve statistical significance. In contrast, if the lack of statistical significance in the original study was not a consequence of a small sample size, e.g. if it were a negligible or no effect, achieving statistical significance will remain a challenge until the replicate study uses a sample size unacceptably amplified to a level that allows even trivial effects to become statistically significant [21,22].

CONFIDENCE INTERVAL AROUND THE MEASURE OF EFFECT OR ASSOCIATION

Information on the direction of the *p*-value 0.06 (whether it can move up or down) is obtained by computing the 95% confidence interval (95% CI) around the point estimate. A wide 95% CI, say (0.9 – 7.34), suggests that the data are compatible with a true effect but that the sample size is simply not sufficient enough to have an adequate statistical power to exclude chance as a likely explanation of the outcome [1-6]. In such a case, if the study is replicated with an optimal sample size, the likelihood of obtaining statistical significance is irrefutable. In contrast, a narrow 95% CI, say (0.8 – 1.34), indicating a much smaller degree of variability, is much more informative about the true magnitude of the effect associated with the outcome [1-6]. That is to say, not every *p*-value of 0.06 or so should be treated as of borderline statistical significance. A *p*-value of 0.06 or so, with a narrow 95% CI would indeed add support that there is actually no true effect and calling it of a borderline statistical significance cannot be justified here. As mentioned above, when the *p*-value is equal to 0.06 and comes

with a wide 95% CI, there is a high probability that there is an association but the study did not have adequate power to exclude chance as the likely explanation of the outcome. Therefore, when investigating the role of chance in study findings, both the *p*-value and CI should be meticulously interpreted in order to accurately convey the message contained in the data and avoid misinterpretation and consequent misleading action. In case the 95% CI indicates that the sample size is not sufficient to rebut the null hypothesis the use of the term borderline statistical significance should be permitted to differentiate it from a *p*-value of 0.06 that would hardly achieve statistical significance even after replications of the study with a larger sample size [23]. Fisher too recommends that the *p*-value should be interpreted in the context of a series of experiments [20]. There is no reason why we should limit ourselves to one single experiment when a *p*-value is say 0.06 and fail to reject the null hypothesis. Instead, we should explore the results further, more especially when the CI indicates that an effect of potential interest is likely to show up if the degree of variability is reduced. In some cases, replication may be useful even for studies with *p*-values less than 0.05.

TEST OF RANDOM DUPLICATION OF PARTICIPANTS

Replication of an original study with a slightly larger sample size as discussed above is not an easy task mainly because of practical considerations mentioned in this paper. Therefore, when the *p*-value is of borderline statistical significance, say $p = 0.06$ with a wide 95% CI, to avoid misinterpretation and the resulting misleading actions we propose the use of the table of random numbers to randomly duplicate a small number of participants (1-10%) of the original sample size and use the new sample size for a repeat analysis. By duplication, we mean the creation of two different records for the same study participant. The *p*-value 0.06 will become statistically significant only if it was truly of borderline statistical significance. It will not become statistically significant unless the sample size is excessively amplified allowing for any trivial effect to become statistically significant [22,23]. To illustrate this concept, let us look at two sets of published data where authors judiciously used the term borderline statistical significance.

STUDY ONE

In evaluating the association between parasite burden and wasting among children, Tshikuka and colleagues used three multivariate sub-models [24]. One of the sub-models showed that wasted children were more likely to be infected with *Ascaris lumbricoides* [adjusted odds ratio (AOR) = 3.63; 95% CI: 1.18-11.16; $p < 0.05$)], and possibly with *Trichuris trichiura* (AOR = 2.56; 95% CI: 0.9-7.34; $p = 0.07$). The children were more likely to be younger (AOR = 0.25; 95% CI: 0.13-0.48; $p < 0.001$) and have diarrhoea (AOR = 18.73; 95% CI: 6.80-51.56; $p < 0.001$). Although *T. trichiura* achieved only a *p*-value of 0.07, the authors concluded that there was a possible association between wasted children and *T. trichiura*. That was not only because the *p*-value 0.07 was close to the cut-off *p*-value of 0.05 but primarily because the size of 95% CI suggested that the data were compatible with a true association but the sample size did not have adequate power to exclude chance as the likely explanation of the outcome. A narrow CI would have added support that there was actually no true association or effect and the use of the term borderline statistical significance in that case would have been incorrect. This reasoning suggests that the absence of statistical significance does not necessarily provide evidence of the absence of effect [10,11]. Had the authors replicated this study with optimal sample size they would have certainly turned the association between wasted children and *T. trichiura* into a statistical significance ($p < 0.05$), hence, justifying their failure to accept the null hypothesis.

Since the study was conducted several years back and circumstances have obviously changed ever since, replicating this study today would suffer various types of biases. Nonetheless, to verify the outcome of that study, data were obtained from the authors [24]. We used a table of random numbers and randomly duplicated 40 of the 558 participants and added them to the original sample to get a new sample of 598 children. Using the new sample size of 598 children in the analysis, the association between *T. trichiura* and wasting became evident (AOR = 2.02; 95% CI: 1.012-5.10; $p < 0.05$), which to us sufficiently justifies the use of $p = 0.07$ as a borderline statistical significance.

STUDY TWO

The same authors as in study one above used the term borderline statistical significance recently while investigating the association between HIV/AIDS, tuberculosis and malaria-specific mortalities and socio-demographic and economic factors in 3 multivariate sub-models [25]. Year of admission to hospital was retained in the malaria and HIV sub-models and was

statistically significant (AOR^1 = 1.85; 95% CI: 1.55 - 2.19, $p < 0.05$; and AOR^2 = 2.15; 95% CI: 1.61-2.85, $p < 0.05$ respectively) but was not statistically significant in the tuberculosis sub-model (AOR^3 = 1.32; 95% CI: 0.99 - 1.75, $p = 0.05$). The authors, however, forced this explanatory variable into the tuberculosis sub-model on the borderline statistical significance ticket and concluded that patients admitted in the recent year for any of the three illnesses including tuberculosis were more likely to die than those admitted in the previous year. Once again, the reasons why they failed to accept the null hypothesis were not well stated in the article. We obtained the data from the authors and performed a random duplication of 16 of their 1325 participants and added 16 to the original sample of 1325. The new sample size of 1341 was analysed and year of admission to hospital turned out to be statistically significant in all the three sub-models including the tuberculosis sub-model (AOR^3 = 1.36; 95% CI of 1.03-1.80, $p < 0.05$) indicating that forcing of this variable into the tuberculosis sub-model with a borderline significant *p*-value of 0.05 was well justified.

While these examples support the use of the term borderline statistical significance, they both illustrate the fact that we should not consider the *p*-value as a rule of thumb on which all our conclusions must be based. Rather, the *p*-value should be interpreted carefully in the context of other factors like the CI. In fact, there are circumstances where we have to consider an effect as significant even though its *p*-value is greater than or equal to the conventional *p*-value of 0.05. For instance, in cases of established facts like the cause-effect relationship of smoking and lung cancer.

ESTABLISHED FACTS AND CLINICAL SIGNIFICANCE

The association between smoking and lung cancer is a well-established fact [26]. The questions is: while investigating the relationship between smoking and lung cancer in a given population, if for some reasons, say, inadequate sample size or insufficient outcome events, the association between smoking and lung cancer achieves only a *p*-value of 0.06 or so, should we just accept the null hypothesis and conclude that there is no effect? As alluded to above, such an assertion would be aberrant and misleading. Since this is a well-established fact, it is legitimate to interpret such a *p*-value as of a borderline statistical significance and accept the outcome [26,27]. Nonetheless, this case differs from, say, a *p*-value of 0.06 when investigating the relationship between variables such as shoe size and the risk of having caesarean section which is not an established fact [28,29].

In addition, a finding that does not have statistical significance can still be clinically significant and vice versa. In the medical field, clinical significance denotes superiority of a treatment, diagnostic method or procedure compared to existing one(s), or even to a placebo. Whereas statistical significance is a mathematical concept, clinical significance is a practical concept in that a treatment, diagnostic method or procedure's effect is assessed and judged based on its significance clinically [2-4]. Statistical significance is largely used to inform clinical findings and the absence of which does not necessarily indicate the absence of clinical significance [1-6,21]. For instance, in comparing the effectiveness of topical chloramphenicol in preventing wound infections after minor surgery to a placebo, Heal and colleagues found a statistically significant improvement in the treatment group compared to the placebo group ($p < 0.01$) but the outcome was deemed not to be clinically significant because the reduction in infection incidence was less than the smallest effect of clinical interest [30]. On the other hand, it is possible to have an effect that can change current clinical practice even though it is not statistically significant.

CONCLUSION

The absence of statistical significance is not necessarily evidence of absence of effect. Therefore, the *p*-value should not be looked at as a rule of thumb for accepting or rejecting the null hypothesis but rather a guide for further action or analysis leading to the correct conclusion. Investigating CIs of the measure of association, replicating the study using an optimal sample size, performing a random duplication of participants as explained herein, searching the literature for established facts are further measures that should be considered when the *p*-value is borderline in order to establish whether there is true effect or not. We hope that suggestions made in this paper including our theory of random duplication of participants will stimulate a productive discussion and add to improved statistical reasoning to avoid misleading conclusions and ill-advised public health actions.

REFERENCES

[1] Morman GR, Streiner DL. PDQ statistics. 2nd ed. St. Louis: Mosby 1997.

[2] Bowers. Medical statistics from scratch. 2nd ed. West Sussex, England: John Wiley & Sons Ltd 2008.

[3] Kirkwood BR, Sterne JAC. Essential medical statistics. 2nd ed. Victoria, Australia: Blackwell Science Ltd 2005.

[4] Peacock JL, Peacock PJ. Oxford handbook of medical statistics. London: Oxford University Press 2011.

[5] Gordis L. Epidemiology. 3rd ed. Philadelphia: Elsevier Saunders 2004.

[6] Lang TA, Secic M. How to report statistics in medicine: annotated guidelines for authors, editors and reviewers. 2nd ed. Philadelphia: Sheridan Press 2006.

[7] Guyatt GH, Oxman AD, Kunz R, *et al*. GRADE guidelines 6. Rating the quality of evidence -- imprecision. J Clin Epidemiol 2011; 64(12): 1283-1293. http://dx.doi.org/10.1016/j.jclinepi.2011.01.012

[8] Alderson P, Chalmers I. Survey of claims of no effect in abstracts of Cochrane reviews. BMJ 2003; 326(7387): 475. http://dx.doi.org/10.1136/bmj.326.7387.475

[9] Khan I, Sarker SJ, Hackshaw A. Smaller sample sizes for phase II trials based on exact tests with actual error rates by trading-off their nominal levels of significance and power. Br J cancer 2012; 107(11): 1801-1809. http://dx.doi.org/10.1038/bjc.2012.444

[10] Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ 1995; 311(7003): 485. http://dx.doi.org/10.1136/bmj.311.7003.485

[11] Alderson P. Absence of evidence is not evidence of absence: We need to report uncertain results and do it clearly. BMJ 2004; 328(7438): 476-477. http://dx.doi.org/10.1136/bmj.328.7438.476

[12] Graham CD, Rose MR, Hankins M, Chalder T, Weinman J. Separating emotions from consequences in muscle disease: comparing beneficial and unhelpful illness schemata to inform intervention development. J Psychosom Res 2013; 74(4): 320-326. http://dx.doi.org/10.1016/j.jpsychores.2012.09.012

[13] Schlesselman JJ, Schneiderman MA. Case control studies: design, conduct, and analysis. J Occup Environ Med 1982; 24: 879.

[14] Armitage P, Berry G, Matthews JNS. Statistical methods in medical research, 4th ed. Oxford: Blackwell Science Ltd 1994.

[15] Altman DG. Practical Statistics for Medical Research. London: Chapman & Hall/CRC 1990.

[16] Arya R, Antonisamy B, Kumar S. Sample size estimation in prevalence studies. Indian J Pediatr 2012; 79(11): 1482-1488. http://dx.doi.org/10.1007/s12098-012-0763-3

[17] Fox N, Hunn A, Mathers N. Sampling and sample size calculation. London: The NIHR RDS for the East Midlands/Yorkshire & the Humber 2007.

[18] Tukey JW. Analyzing data: sanctification or detective work? Am Psychol 1969; 24: 83. http://dx.doi.org/10.1037/h0027108

[19] Tukey JW. The philosophy of multiple comparisons. Stat Sci 1991; 100-16. http://dx.doi.org/10.1214/ss/1177011945

[20] Fisher RA. Statistical methods for research workers. 13th ed. New York: Hafner 1958.

[21] Sedgwick P. Clinical significance versus statistical significance. BMJ 2014; 348. http://dx.doi.org/10.1136/bmj.g2130

[22] Mayrent SL, Ed. Epidemiology in medicine. Boston: Little Brown and Company 1987.

[23] Wainer H, Robinson DH. Shaping up the practice of null hypothesis significance testing. Educ Res 2003; 32(7): 22-30. http://dx.doi.org/10.3102/0013189X032007022

[24] Tshikuka JG, Gray-Donald K, Scott M, Olela KN. Relationship of childhood protein-energy malnutrition and parasite infections in an urban African setting. Trop Med Int Health 1997; 2(4): 374-382. http://dx.doi.org/10.1111/j.1365-3156.1997.tb00154.x

[25] Tshikuka JG, Okenge L, Lukuka A, *et al*. Severity of outcomes associated to illnesses funded by GFATM initiative and socio demographic and economic factors associated with HIV/AIDS, TB and malaria mortality in Kinshasa Hospitals, DRC. Ethiop J Health Sci 2014; 24(4): 299-306. http://dx.doi.org/10.4314/ejhs.v24i4.4

[26] Doll R, Hill AB. Mortality in relation to smoking: ten years' observations of British doctors. Br Med J 1964a; 1(5395): 1399-1410. http://dx.doi.org/10.1136/bmj.1.5395.1399

[27] Doll R, Hill AB. Mortality in relation to smoking: ten years' observations of British doctors. Br Med J 1964b; 1(5396): 1460-1467. http://dx.doi.org/10.1136/bmj.1.5396.1460

[28] Frame S, Moore J, Peters A, Hall D. Maternal height and shoe size as predictors of pelvic disproportion: an assessment. Br J Obstet Gynaecol 1985; 92(12): 1239-1245. http://dx.doi.org/10.1111/j.1471-0528.1985.tb04869.x

[29] Mahmood TA, Campbell DM, Wilson AW. Maternal height, shoe size, and outcome of labour in white primigravidas: a prospective anthropometric study. BMJ 1988; 297(6647): 515-517. http://dx.doi.org/10.1136/bmj.297.6647.515

[30] Heal CF, Buettner PG, Cruickshank R, *et al*. Does single application of topical chloramphenicol to high risk sutured wounds reduce incidence of wound infection after minor surgery? Prospective randomized placebo controlled double blind trial. BMJ 2009; 338: a2812. http://dx.doi.org/10.1136/bmj.a2812

Received on 15-02-2016 Accepted on 18-07-2016 Published on 16-08-2016

http://dx.doi.org/10.6000/1929-6029.2016.05.03.7