Perspective

Beyond P-value: the Rigor and Power of Study

Fengyu Zhang[*], Claude Hughes

Received June 29, 2019, Accepted December 6, 2019

ABSTRACT

There have been a series of recent discussions and debates on the p-value and statistical significance. These discussions, including publications of more than 40 papers in a special issue of the *American Statistician,* provide an excellent opportunity to think about some technical measures for practical implementation in grant applications and publications. While several factors have been discussed, it may be the rigor of a study that determines the p-value for reporting study results and judging a consistent replication of research. Both p-values and power, which integrate Fisherian and Neyman-Pearson methods, should be used for hypothesis testing. We propose new criteria, which can be implemented without fundamental changes in existing statistics, to reduce false positives and irreplicability of studies that are either inadequately powered or overpowered.

KEYWORDS

P-value, rigor of study, statistical power, statistical hypothesis testing, statistical significance

Since Ronald A. Fisher introduced the test of significance in 1925, the p-value has been the "gold standard" measure for statistical hypothesis testing. While the Fisher's hypothesis test has had critics such as those from Neyman and Pearson since its origin, recent discussions have focused on the misuse or misinterpretation of the p-value, leading some authors to recommend lowering the p-value threshold(1, 2), or to abandon statistical significance statements altogether(3). The main reasons motivating such arguments are that most of the published results that have passed the p-value threshold are false positives(4) or non-replicable (5). This accelerating argument has also been raised in other fields such as psychology, behavioral sciences, and biomedical sciences (5-7).

The discussions about the p-value are drawing continued attention to this issue. As far back as 2005, John Ioannidis, an epidemiologist from Stanford University, pointed out some possible reasons for the presence of false positives in publications. Multiple additional commentaries have been published in various areas remarking on study design, data collection, power, uncontrolled heterogeneity, and problems in statistical analysis (8, 9). In February 2014, Regina Nuzzo published a highly viewed article on the statistical error (10), which seemingly initiated another round of discussions and debates on the p-value. Some proposed a ban on the use of p-value, which arguably may not help solve the problem of a replicability crisis in research(5). The American Statistical Association (ASA) released a statement on March 7, 2016 (Box1), to address some of the concerns about the misuse and misconception of the p-value. These discussions were expected to have more influence, but it seemed to have had little impact on the practice of research following the ASA statement, according to a follow-up with journal publications a year later (11).

Box 1 American Statistical Association Statement on Statistical Significance and *P*-Values.

1) P-values can indicate how incompatible the data are with a specified statistical model.
2) P-values do not measure the probability that the studied hypothesis is correct or the likelihood that the data were produced by random chance alone.
3) Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4) Proper inference requires full reporting and transparency.
5) A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6) By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

[*] Correspondence to: F Zhang, Email:zhangfy@gcatresearch.org

There have been further discussions about statistical significance and inference. In March 2019, scientists published a commentary in *Nature* "rise up against statistical significance," a conclusion that is made based on the p-value. The review received more than 800 signatories in about a week from scientists across continents(3), although they and other investigators(12) expressed no desire to ban the use of p-value or other measures of statistics. Concurrently, the journal of the *American Statistician* de-voted a special issue of more than 40 papers to "Statistical Inference"(13). These papers propose various statistical measures alternative to the p-value, and they create "knock-on" needs for a feasible plan to implement such practices in research, notably by educating scientists, reviewers, editors, funding agencies, regulators and industry on how to adapt and interpret such new measures beyond the p-value.

In addition, the recently emerging fields of biomedical research with a large-scale "Omics"(14) may require sophisticated data processing such as normalization, variant calling, and analytics where advanced statistical tools are needed to produce a correct p-value. Here we discuss several key concepts associated with the use of p-value and propose new criteria for statistical inference in research and other applications, which are organized as follows: 1) sampling and hypothesis testing, 2) determinant of the p-value, 3) interpretation of the p-value, and 4) report on rigor and power of a study to improve replicability in research.

SAMPLING AND HYPOTHESIS TESTING

Sampling is the first crucial concept for the use of applied statistics in the 20th century, and notably, Fisher made pioneering contributions to all three components: parameter estimation (15), statistical hypothesis testing(16), and experimental design (17). Because most extant populations of interest are too large to study *in toto*, cannot be adequately assessed with certainty, or are theoretically infinite, it is neither feasible nor necessary to examine an entire population. The sampling technique helps to determine a representative sample, which is then used to make statistical inferences on the population parameters of interest. In his influential book, "Statistical Methods for Research Worker," published in 1925(16), Fisher describes the statistical test to test the difference between the observed and expected using a Chi-Square statistic, which had been found by I.J. Bienayme in 1850s and by F.R. Hekmert and E. Abbe in 1860s and 1870s, and was rediscovered by K Pearson in 1900(18). Based on the Tables that are constructed to show the deviation and corresponding different values of probability, Fisher considers p=0.05 or 2-fold (1.96) standard deviations as a limit in judging if the departure is significant or not.

In practical research, one has to define a study population and scientific hypothesis to be tested with a sample from the defined study population. Statistical hypothesis testing should be the first carried out based on the collected data of the sample. Under the "null hypothesis H_0", a statistic, T_{obs} is calculated from an empirical sample that is selected to represent the study population based on a known distribution $f(T|H_0)$; For a two-sided alternative hypothesis, the p-value is defined as $P(|T|\geq|T_{obs}|$,under H_0)=2*$P(T \geq |T_{obs}|$, under H_0), where T is a random variable following distribution $f(T|H_0)$. While having proposed a combined p-value of multiple independent tests for the same hypothesis(16), Fisher had never quantified the statistical power explicitly (17).

Fisher's hypothesis testing mainly considers p-value, but it pays little attention to the size of effect, power, and information on the study population. Soon after Fisher published work that includes test research data or the test of significance in 1925 (16, 19), Jerzy Neyman (1894-1981), a Polish mathematician and statistician, and Egon Pearson (1895-1980), a British statistician, challenged Fisher's approach by developing the test of a statistical hypothesis (20). This led to the development of the Neyman–Pearson lemma of statistical hypothesis testing published in 1933 (21), which, together with a series of papers, provided a consistent logical basis for statistical hypothesis testing. Neyman-Pearson hypothesis testing deals with two different hypotheses. i.e., the test statistic T has two different probability density functions of $f(T|H_0)$ under the null hypothesis and $f(T|H_1)$ under the alternative hypothesis. This forms a framework of statistical power, false positives, and false negatives. While multiple factors such as sizes of sample and effect and sample variation affect the p-value, the Neyman-Pearson lemma of hypothesis testing is complementary to Fisher's hypothesis testing. As Eric L. Lehman expressed, "despite basic philosophical differences, in their main practical aspects, the two theories are complementary rather than contradictory" (22).

DETERMINANTS OF *P*-VALUE

The sampling and sample size affect the estimation of sampling error and therefore influence the p-value for statistical inference. In social sciences such as economics, social, or demographic studies, it is relatively feasible to obtain a representative sample from a study population. However, in biomedical research, especially hospital-based research on patients, it is challenging to do a random or representtative sampling due to constraints on the feasibility of defining a sampling frame or performing subject recruitment and biospecimen collection. For example, in a genome-wide association study of a human complex disorder, patients may be recruited from clinics and hospitals where patients may be from diverse locations due to patterns of referral or care networks, whereas healthy controls are often from the respective local communities. That divergence may cause problems of sample representativeness, which may affect the replicability of research or create population stratifications, which is notoriously known for producing false-positive findings, particularly in the study of genomics and other "Omics" in human populations.

Even with random sampling, the samples from simple random and multistage cluster random sampling require different statistical methods for parameter estimation. In cluster sampling, special attention should be paid to the estimation of standard error, because non-independence of subjects within the same cluster or community may lead to an underestimation of the standard error for a parameter estimate when classical methods are used.

In addition to the sample and effect sizes that determine the power to detect an actual effect, the research process itself, including data collection, measurement, batch effect, data processing, and statistical analysis, all influence the p-value (4, 14). Errors in any step of the research process may lead to a biased estimate of the parameter or the associated standard error. Therefore, to assure that a p-value is correct, independent replications, for which Fisher had initially considered combining multiple independent tests, are recommended. A proper independent replication study should be performed under the same conditions, notably including that a sample should be from the same study population. A combined p-value of multiple independent tests for the same hypothesis may increase the power to detect a real effect. However, results from a purposively selected sample, especially out of the study population, are often misleading.

INTERPRETATION OF *P*-VALUE

The p-value derived from one sample statistic only states that a chance to observe such as result if the "null hypothesis H_0" was correct. How the p-value is obtained determines its interpretation. In the first use by Fisher, the p-value was used to measure the difference between observed and expected data (i.e., compatibility with model)(16). However, the p-value is not used to measure the probability that the hypothesis is correct (ASA Statement No 1-2).

A small p-value below the significant threshold is only a reflection of a statistical outcome based on one sample. The decision made may have to be associated with sampling error. For example, with 100 independent samples from a study population, one can obtain 100 *p*-values to test a hypothesis. If there are 95 samples with a *p*-value of less than 0.05, one can feel more confident in rejecting the "null hypothesis H_0" at an error rate of 0.05 (type I error). That says testing for a statistical hypothesis requires consistent independent replications using samples from the same study population and with the same "recipe" to keep a type I error under control. When a type I error is specified, and a p-value passes the threshold, the quantitative magnitude of the p-value would have little to do with the decision on a study.

The interpretation of the p-value should consider the type of design for a study. In principle, the experimental study is less subject to confounding effect than an observational study, in which the later merely assesses association but not causation. In observational research, whenever a significant or small p-value is observed, further investigation with an advanced study design will be warranted. When a statistical "null hypothesis H_0" is rejected with sufficient evidence, it warrants further evaluation of the scientific hypothesis of interest; even though "scientific conclusion and policy decision should not be based on only whether the p-value passes the specific threshold" (ASA Statement No. 3). When multiple independent tests of significance have been performed, and few or none of them are individually significant, Fisher recommends a combined p-value (16) to provide aggregate evidence.

More importantly, interpretation of the p-value must be in the context of statistical power along with the size of the effect, because p-value does not measure the size of effect. As in the ASA released statement (No 5-6), "the p-value does not measure the size of effect or importance of the results, and is not a good measure of evidence regarding a hypothesis." The practice of current statistical hypothesis testing is mostly a Fisherian test of significance but does not consider power or effect size, in which Neyman and Pearson initially proposed and argued against Fisher's statistical testing. In the example that Amrhein et al. discussed (3), two studies with the same hazard ratio (HR=1.2), one was very significant (95% CL:1.09-1.33; *p*=0.0003); whereas the other was just "non-significant" (95% CL: 1.03-1.48; *p*=0.097). If one knew the power of each study, which could be from the post hoc power analysis, one should have been able to explain the difference between these two studies. It appeared that the second study was with a large confidence interval, which could be due to a smaller sample size that led to the "non-significant" (*p*<0.097).

Also, using power and effect size in statistical hypothesis testing would prevent false positives that might be caused by an overpowered study. Some recent genome-wide association studies of complex human disorders often end up with a large sample size. A large number of variants are detected at a stringent level of threshold ($p<5x10^{-08}$) but often with small effect sizes (23, 24). Further examination of these variants finds they are primarily located in regions with a rare copy number of variations(23), which might not be those the genome-wide association study is designed to discover. The actual genetic variants associated with the diseases may have been hidden. It may be that the large sample size alone drives the finding of significant outcomes.

REPORT ON RIGOR AND POWER OF STUDY

Recent commentaries about the use of statistical significance and p-value may or may not enhance the replicability of scientific research. In the special issue by the journal of the *American Statistician* "Statistical Inference in the 21st Century: a World Beyond *p*<0.05", more than 40 papers proposed a range of solutions, including some new measures as alternatives to the use of p-value(13). Investigators or editors will need to consider how to unify these proposals into research practices or publications. As an initial step, we suggest the following measures to reduce false positives and to improve replicability in research.

Strengthen rigor of study design and conduct throughout the research process

The rigor of a study is the key to the appropriate use of p-value and replicability of research. A rigorous study design needs to be implemented throughout the entire research process from sampling, measurement of outcome and covariates, data quality control to statistical analysis. A statistician should be first involved in study design, even starting with defining a study population as well as sampling and recruitment plans. The initial design is critical for the replicability of research. An investigator can determine a smaller and more homogenous population to study when the study outcome is anticipated to be heterogeneous or to cope with budget constraints. It is crucial to make sure those samples for discovery and replication should be from the same defined population.

Research and quality assurance protocols should be developed to guide the implementation of research as designed. Some of these processes have been suggested in the publications that recently appeared in the special issue of the *American Statistician* — for example, getting statisticians more involved throughout the entire research process as well as preparation of publications (25). Rigorous statistical analysis should always be assured. Sampling and the measurement of study outcomes determine the selection of appropriate methods for statistical analysis. A study with multi-stage cluster sampling survey should employ hierarchical modeling (26, 27).

Furthermore, strengthening the rigor of study design and conduct requires systematic training of investigators in biomedical research. While recent developments in biotechnology and bioinformatics tools have provided an excelent opportunity to facilitate the discovery that can affect clinical practice and development of novel therapeutics, the same biotechnology advances create challenges for designing and conducting studies with rigor. These opportunities are also challenges that require funding agencies and institutions to provide increased funding for training their investigators. Other related research training should be provided for both mentors and trainees (28).

Increase transparency in report

Transparent reports on research help improve research replicability by providing details essential to the "recipe" needed for performing replication of the study being reported. If a subsequent investigator cannot ascertain the specific methods used in a study published in the literature, and then it is less likely that a consistent replication will be obtained. One group of scientists has issued a call for transparent reporting of preclinical studies, and they proposed a set of standards for reporting a rigorous design (29), which will improve proper statistical inference (ASA Statement No. 4). Some journals such as Nature Research Journals have started implementing the report standards as a checklist when a research manuscript is submitted.

Transparent reports, especially in terms of study design and statistical analysis, are essential in clinical and preclinical research (29, 30). Even though most biomedical grant proposals list a statistician, the need is to demonstrate that the statistician will be more involved throughout the research process from design, to conduct, to the publication of study outcomes. In terms of papers, editors should obtain advice from a statistical reviewer earlier rather than waiting until the review process is complete. Publishing results from a biomedical study with a detectable flaw in the design or statistical analysis, especially in high profile journals, is detrimental to scientific research *per se* and poses an unnecessary risk to human health.

Require power, effect size, and 95 percent of confidence limits in reports

Strengthening the rigor of a study and increasing transparency in reports may take time to implement. We propose to report the power and size of effect concurrently with the p-value, which may help reduce false positives. This proposal, in theory, makes hypothesis testing more complete, an integration of the Fisherian and Neyman-Pearson methods. While not explicitly describing the power for statistical testing, Fisher did state when one study is not significant, an additional sample should be added, which leads to the development of the Fisher combined method (16). By reporting the power of a study, one can avoid publishing a false positive that could result from an underpowered or overpowered study, both of which may lead to incorrect conclusions or be detrimental. With an underpowered study, the probability of detecting a real effect is low, and some of the non-replications are due to low statistical power (31, 32), which is more sensitive to error. When a significant p-value is obtained in an underpowered study, the false-positive risk will still be considerable (33); and the positive findings may be most likely due to misuse of statistics or non-statistical cause.

In contrast, an overpowered study may show a significant p-value for what is not a real effect, or the result is not meaningful, usually with smaller effect size and smaller p-value, in which a large sample size drives the statistical significance. In clinical trials, it means that "a trial has the power to detect smaller difference and still be statistically significant." (34) The over-powered study is often seen in recent genome-wide association studies of complex human disorders or randomized clinical trials that pull various heterogeneous samples for meta-analysis (35, 36). Therefore, an overpowered study also causes false positives and should be avoided because it wastes resources, in particular for clinical trials (37).

We propose the criteria for making a statistical decision on the "null hypothesis H_0" based on both p-value and power, with a specified or meaningful effect size (Table 1). When a p-value meets the threshold (e.g., 0.05, or 0.01), and a study power is between 80% and 99 %, one can decide on rejecting the "null hypothesis H_0" (*Decision 1*). When a p-value meets a threshold, but from a study with low statistical power (<80%), one should make no decision on the "null hypothesis H_0," and independent replication is needed (*Decision 2*). When a p-value meets the threshold, but from a study with power >99% or higher, a sensitivity or subgroup analysis should be carried out to examine the heterogeneity of the study sample (*Decision 3*). For example, in a genome-wide association study of complex human disorders, an analytical sample could include samples from multiple populations, in which the frequency of genetic variants of interest, such as single nucleotide polymorphisms (SNPs), could vary. Therefore, subgroup analysis can help reveal sample genetic heterogeneity, which has demonstrated in a previous study(38). Besides, 95% confidence limits should be required to report in practice.

Table 1. Criteria recommendation for making a decision on "Null hypothesis H_0" based on p-value and power

	Effect [a]	α^b	Power	Decision
1)	Beta or OR	0.05, 0.01	80-99%	A judgment against the "null hypothesis";
2)	Beta or OR	0.05, 0.01	<80%	No decision. Need an independent sample for replication
3)	Beta or OR	0.05, 0.01	>99%	No decision. Need a sensitivity analysis or subgroup analysis

a. OR, odds ratio; RR, relative risk; and HR, hazard ratio; Beta is the regression coefficient;

b. The threshold for p-value, the specification should take into account multiple testing.

Advocate and conduct independent replications and interventional studies

Whenever it is possible, in particular for claiming a discovery, an observational study or non-clinical trial should have independent replication. This practice has been implemented in performing genome-wide association studies of complex human disorders over the past decade (39). Ideally, a replication sample should be at least from the same study population. Results from previously published research, which might be from a sample in a different study population, are helpful but may not be able to serve as an independent replication for a specific study. For example, one performs genetic research of a human disorder in the US population but uses the data from another country for replication. It could be appropriate if previously published results from a different study population provided a consistent replication, but this may not falsify the finding from one's study. Sometimes people are trying to validate results from a well-designed study by a study with a relatively small sample or a large sample from a combined or meta-analysis, which could hide the actual findings (35). Replication should be conducted under similar conditions and within the same study population; otherwise, anyone can claim to falsify other's conclusions.

Evidence from a longitudinal or interventional study is stronger in making a causal inference, which is very important for moving the statistical hypothesis testing to a scientific hypothesis. In the field of biologic or genetic studies, the true finding can be validated eventually at a level of biological function, but one needs to obtain a reliable tar-get to pursue further investigation of the underlying bio-logy. In clinical trials, independent replication is costly and time -consuming, but the analysis of additional outcomes from the same study sample should always be encouraged. A clinical study with follow-up at multiple time points provides a degree of self-replication.

Perform rigorous statistical analysis

Statistical analysis is the key to step to produce the final p-value. Some have pointed out that the false-positives are due to problems in statistical analysis (14), which is often performed by those without adequate training or related expertise. When reporting a study, one should specify statistical software used for data analysis. If using a non-routine method or tool, the authors should provide results from an independent tool, in particular for genetic variant calling in exome- or whole-genome sequencing. In general, it is strongly recommended to use reliable software, which has undergone related quality assurance and for which the computational methods are described somewhere.

In summary, we further discussed the p-value and rigor of a study and proposed new criteria for reducing false-positive findings in research. The requirements can be implemented within existing research practices in statistics. We also encourage the inclusion of professional statisticians more comprehensively throughout design, conduct, and publication of research. These are critical for clinical and biomedical research, in which some studies are costly, and a misleading result may be detrimental to human health.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest regarding the publication of this paper.

ACKNOWLEDGMENT

The authors thank for anonymous reviewers who provided comments, which allowed improving this article.

REDERENCES

1. Ioannidis JPA. The Proposal to Lower P Value Thresholds to .005. JAMA. 2018;319(14):1429-30.
2. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers E-J, Berk R, et al. Redefine statistical significance. Nature Human Behaviour. 2018;2(1):6.
3. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature. 2019;567 (7748): 305-7.
4. Ioannidis JP. Why most published research findings are false. PLoS Med. 2005;2(8):e124.
5. Savalei V, Dunn E. Is the call to abandon p-values the red herring of the replicability crisis? Front Psychol. 2015;6:245.
6. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, et al. Redefine statistical significance. Nat Hum Behav. 2018;2(1):6-10.
7. McNutt M. Journals unite for reproducibility. Science. 2014; 346(6210):679.
8. Goodman S, Greenland S. Why most published research findings are false: problems in the analysis. PLoS Med. 2007; 4(4):e168.
9. Shrier I. Power, reliability, and heterogeneous results. PLoS Med. 2005;2(11):e386; author reply e98.
10. Nuzzo R. Scientific method: statistical errors. Nature. 2014; 506(7487):150-2.
11. Mathew R. The ASA's p-value statement,one year on. Significance. 2017;14(2):38-41.
12. McShane BB, Gal D, Gelman A, Robert C, Tackett JL. Abandon Statistical Significance. The American Statistician. 2019; 73 (sup1):235-45.
13. Wasserstein R, Schirm A, Lazar N. Moving to aWorld Beyond "p<0.05". The American Statistician. 2019;73 (sup1):1-19.

14. Leek JT, Peng RD. Statistics: P values are just the tip of the iceberg. Nature. 2015;520(7549):612.
15. Fisher RA. On the mathematical foundations of theoretical statistics. PhilosophicalTransactions of the Royal Society of London Series A, Containing Papers of a Mathematical or Physical Character. 1922;222(1):309-36.
16. Fisher R. Statistical Methods for Research Worker. Edingburg: Oliver and Boyd; 1925.
17. Lehmann E. Fisher, Neyman, and the Creation of Classical Statistics: Springer; 2011.
18. Cowles M. Statistics in Psychology: An Historical Perspective: Taylor & Francis; 2005.
19. Perezgonzalez JD. Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. Front Psychol.2015;6:223.
20. Neyman J, Pearson ES. On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. Biometrika. 1928:175-240.
21. Neyman J, Pearson ES. IX. On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society of London Series A,Containing Papers of a Mathematical or Physical Character. 1933; 231 (694-706):289-337.
22. Lehman EL. The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two?. Journal of the American Statistical Association 1993; 88(424):1242-9.
23. Xia L, Xia K, Weinberger DR, Zhang F. Common genetic variants shared among five major psychiatric disorders: a large-scale genome-wide combined analysis. Glob Clin Transl Res. 2019;1(1):21-30.
24. Schizophrenia Working Group of the Psychiatric Genomics C. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014;511(7510):421-7.
25. Locascio J. The Impact of Results Blind Science Publishing on Statistical Consultation and Collaboration The American Statistician. 2019;73(Sup1):346-51.
26. Steele F, Diamond I, Wang D. The determinants of the duration of contraceptive use in China: a multilevel multi-nomial discrete-hazards modeling approach. Demography. 1996; 33 (1):12-23.
27. Short S, Zhang F. Use of maternal health services in rural China. Popul Stud (Camb). 2004;58(1):3-19.
28. Finkel A. The road to bad research is paved with good intentions. Nature. 2019;566:297.
29. Landis SC, Amara SG, Asadullah K, Austin CP, Blumenstein R, Bradley EW, et al. A call for transparent reporting to optimize the predictive value of preclinical research. Nature. 2012; 490(7419):187-91.
30. Schulz KF, Altman DG, Moher D, Group C. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340:c332.
31. Trafimow D, Amrhein V, Areshenkoff CN, Barrera-Causil CJ, Beh EJ, Bilgic YK, et al. Manipulating the Alpha Level Cannot Cure Significance Testing. Front Psychol. 2018;9:699.
32. Bishop D. Rein in the four horsemen of irreproducibility. Nature. 2019;568:435.
33. Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p-values. R Soc Open Sci. 2014;1 (3): 140216.
34. Hochster HS. The power of "p": on overpowered clinical trials and "positive" results. Gastrointest Cancer Res. 2008; 2 (2): 108-9.
35. Ioannidis JP. The Mass Production of Redundant, Misleading, and Conflicted Systematic Reviews and Meta-analyses. Milbank Q. 2016;94(3):485-514.
36. Nikpay M, Goel A, Won HH, Hall LM, Willenborg C, Kanoni S, et al. A comprehensive1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nat Genet. 2015;47(10):1121-30.
37. Case LD, Ambrosius WT. Power and sample size. Methods Mol Biol. 2007;404:377-408.
38. International Schizophrenia Consortium, Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009;460(7256):748-52.
39. Kraft P, Zeggini E, Ioannidis JP. Replication in genome-wide association studies. Stat Sci. 2009;24(4):561-73.

How to cite this article:
Zhang, F and Hughes CL. Beyond p-value: the rigor and power of study. Glob Clin Transl Res. 2020; 2(1):1-6.
DOI:10.36316/gcatr.02.0021