GCAT-R

ISSN 2641-7154 (Print)
ISSN 2643-8151 (Online)

Global Clinical And Translational Research

Toward an integration of genomic, environmental, and social medicine

Volume 2, Number 1

March 30, 2020

Editorial Office
P O Box 341126
Bethesda, MD
Email: journal@gcatresearch.com
Tel: 202-609-8812
Website: www.gcatresearch.com

Publisher
The journal of *Global Clinical and Translational Research* (GCAT-R) is published and printed quarterly by the Global Clinical and Translational Research Institute, Inc, MD, USA
Website: gcatreserach.org

GUIDE FOR AUTHORS

Aims and Scopes

Global Clinical and Translational Research is a new peer-reviewed, multidisciplinary scientific journal dedicated to publishing articles regarding all areas of clinical and translational research in humans or animals.

The journal aims to promote a unified platform on research communications for basic scientists, medical doctors, clinical and health professionals, social scientists and soci-al workers to share the most recent advances in all areas of clinical and translational sciences.

The journal accepts original research articles, reviews, mini-reviews, correspondence, case reports, short notes, and rapid communications covering all aspects of clinical and translational research. Papers about novel applications of statistical methods or data science are also welcomed. Specific fields of the papers to be published include but not limited to

- Clinical research in human or animals, including iden-tification of genetic or molecular markers associated with diseases or traits that have clinical implications;
- Translational research that based on clinical research findings. For example, research findings provide evid-ence for health professionals to improve human heal-th conditions (T1 translational research), and provide targets for basic scientists to study the biological me-chanism (T2 translational research);
- Clinical trials to evaluate the effectiveness of treatme-nt including pharmacological and non-pharmacologi-cal approach such as traditional Chinese medicine (TCM), with a molecular study to identify biomarkers that could be used to evaluate effectiveness;
- Methodology papers that improve study design, statistical analysis;
- Population-based survey on human health;
- Negative results (We also publish summary data of negative results);
- Clinical and translational research protocols.

Review process

All manuscripts are subject to peer review and are expected to meet standards of academic excellence. Upon submi-ssion, the Editorial Office will consult with associate edit-ors to have a preliminary check on the manuscript subm-itted, and then decide to whether to send out for a peer rev-iew. All reviewers' identities will remain anonymous to the authors.

The journal implements a double-blinded peer review pro-cess, but the review can have access to the author infor-mation if they request and believe that could help evaluate the manuscript.

Other types of papers

Correspondence is directly related to methods and interpretation of data presented in the recent publication in our journal. While allowing to post comments on online, we also publish the correspondence as an e-content in our post categories. Commentaries and editorials are generally for invited only.

Format of manuscript

Title page

The following information should be included

- Paper title
- Full author names
- Affiliations for each author
- Email addresses and contact information on the corresponding author(s)

Abstract

Not exceed 200 words, should be structured as below for all research articles; other articles should use non-structured abstract.

- Background:
- Methods:
- Results:
- Discussion:
- Conclusion:

Keywords (3-5 keywords)

Maintext

Introduction. This section should be succinct, with no sub-headings.

Methods. Should contain all procedures with enough details so that they can be repeated.

Results. Both descriptive and analytic results.

Discussions. This should describe the implications and significance of the findings, also highlighting the limitations of the study.

Conclusion (optional).This should clearly explain the main conclusions of the work highlighting its importance and relevance.

Conflict of interest.

If there is no conflict of interest, authors should state, "The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper."

Acknowledgments (optional, including funding support if available).

References

Authors are responsible for ensuring that the information in each reference is complete and accurate. All references must be numbered consecutively and citations of references in the text should be identified using numbers in square brackets (e.g., "as discussed by Smith [1]"; "as discussed elsewhere [3, 4]"). All references should be cited within the text; otherwise, these references will be automatically removed. GCAT-R uses "Vancouver" style, as outlined in the ICMJE sample references (https://www.nlm.nih.gov/ bsd/uniform _requirements.html).

Supplementary material

Supplementary material, related to the work but not critical to the publication, is encouraged and made available as e-content.

Preparation of Figures

Upon submission of an article, authors are supposed to include all figures and tables in the Word file of the manuscr-ipt. Figures and tables should not be submitted in separate files. If the article is accepted, authors will be asked to pro-vide the source files of the figures. Each figure should be supplied in a separate electronic file. All figures should be cited in the paper in consecutive order. Figures should be supplied in bitmap formats (Photoshop, TIFF, GIF, JPEG, etc.). Bitmap images should be of 300 dpi resolution at least unless the resolution is intentionally set to a lower level for scientific reasons. If a bitmap image has labels, the image and labels should be embedded in separate layers.

Preparation of Tables

Tables should be cited consecutively in the text. Every table must have a descriptive title and if numerical measu-rements are given, the units should be included in the column heading. Vertical rules should not be used. A tab-le should be well annotated and easy to read and interpret with little reference to the text.

Proofs and Print

The corresponding author will receive proofs by email. Some editorial modification will be made after the manuscript is accepted, so it is important to check proofs carefully. Corrected proofs must be returned to the publisher within 48 hours of receipt. The publisher will do everything possible to ensure prompt publication. It will, therefore, be appreciated if the manuscripts and figures conform fro- m the outset to the style of the journal.

Copyright and Permission

When publishing in *Global Clinical and Translational Research*, the author(s) agree to assign the copyright of the article to the journal. We request that permission should be sought from the journal as the rights holder to produce any substantial part of copyrighted work. To obtain permi-ssion, please contact the editorial office as below.

The author(s) may choose to publish with us as an open-access article under the terms and conditions of the Creati-ve Commons Attribution License (CC BY 4.0, https://creat-ive-commons.org/licenses/by/4.0/), which permits unre-stricted use, distribution, and reproduction in any medium, provided the original work be properly cited. Article publi-shed under this condition is subject to article-process char-ge or open-access fee, which will be paid at the time when manuscript is accepted.

The use of general descriptive names, trade names, trademarks, and so forth in this publication, even if not specifically identified, does not imply that the relevant laws and regulations do not protect these names.

While the advice and information in this journal are believ-ed to be true and accurate on the date it is going to press, neither the authors, the editors nor the publisher can ace-pt any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Disclosure Policy

A competing interest exists when professional judgment concerning the validity of research is influenced by a seco-ndary interest, such as financial gain. We require that our authors reveal any possible conflict of interest in their submitted manuscripts.

Clinical Studies

Global Clinical and Translational Research aims to comply with the recommendations of the International Committee of Medical Journal Editors (ICMJE) on trials registration. Therefore, authors are requested to register the clinical trial presented in the manuscript in a public trials registry and include the trial registration number at the end of the abstract. Trials must be registered prospectively before patient recruitment has begun. We promote transparency in clinical trial reporting.

In addition, the journal endorses the Principles and Guide-lines for Reporting Preclinical Research by the National Institutes of Health, which a considerable number of journals have agreed to endorse (https: //www.nih.gov/researc-h-training/rigor-repro-ducibi-lity/principles-guidelines-re-porting-preclinical-resea-rch).The journal also promotes the core set of standards for rigorous reporting of study design (Adapted from Landis et al., Nature 2012).

CONTENTS

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 genomewide 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.

CONFLICT OF INTERESTS

The authors declare no conflict 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

Commentary

A Comment on “Beyond P-value: the Rigor and Power of Study"

Helena Chmura Kraemer

“As ye sow. So shall ye reap”: For almost 100 years, researchers have been taught that the be-all and end-all in data-based research is the p-value. The resulting problems have now generated concern, often from us who have long so taught researchers. We must bear a major responsibility for the present situation and must alter our teachings.

Despite the fact that the Zhang and Hughes paper is titled “*Beyond* p-value”, the total focus remains on statistical hypothesis testing studies (HTS) and p-values(1). Instead, I would propose that there are three distinct, necessary, and important phases of research:

1) Hypothesis Generation Studies (HGS) or Exploratory Research (2-4);
2) Hypothesis Testing Studies (HTS);
3) Replication and Application of Results.

Of these, HTS is undoubtedly the most important, but without HGS, HTS is often weak and wasteful, and without Replication and Application, the results of HTS are often misleading.

HYPOTHESIS GENERATING STUDIES

HGS is done largely on existing data sets: public or clinical records, data from completed HTS, etc., and occasionally on datasets collected specifically for exploration. The goal of HGS is to generate specific strong and important hypotheses for future testing and to gain an understanding of the population and measures necessary for designing valid and powerful tests of those hypotheses. There are no conclusions from such studies, no tests, no p-values, only descriptive statistics, effect sizes, and mathematical modeling results.

If HGS results are mistakenly reported as conclusions, the false positive rate is unacceptably high. Currently, such HGS often report conclusions using invalid p-values, for authors fear that otherwise, reviewers might label such studies derisively as "fishing expeditions" and refuse publication. This is the source of many, perhaps most, of the false-positive results in the research literature(5, 6). Well-done and correctly reported HGS deserve respect and publication.

One crucial task in HGS is the definition of an appropriate effect size (ES) that takes on the value 0 or less if the hypothesis generated is not true (H_0), and increases in size the stronger the true hypothesis (H_1) (e.g., Cohen’s d, (HR-1)/(HR+1), where HR is the hazards ratio, risk difference, correlation coefficients).

What must also be determined from HGS information (especially from the impact of previously done HTS on related issues) is the critical value (CV) of that effect size, the value of the selected ES>0, below which the hypothesis may be true but for clinical purposes, trivial, meriting no further attention. For example, it might be decided in a randomized clinical trial comparing two treatments, that any Cohen’s d under .1 (CV=.1), or in a prediction study, that a correlation less than .2 (CV=.2) is of no great interest or importance, even though each is greater than its null value of zero.

The ES from HGS for any hypothesis considered strong and important enough to go on to HTS should be much greater than its CV, for ES is often overestimated in HTS (capitalization on chance). Why would one spend time and resources on a hypothesis that under the best of circumstances is likely of trivial clinical significance, 0< ES <CV?

HYPOTHESIS-TESTING STUDIES

HTS starts with a strong 'a priori' hypothesis (from HGS). Then the guidance provided by Zhang and Hughes applies. Sampling, design, measurement, and analysis plans are based on knowledge gained in HGS. The proposed analysis (validity predicated on the HGS) will likely result in a valid p-value. The power computations should use the CV (from HGS) to set the sample size large enough so that the probability of rejecting H_0 for any ES>CV is greater than, say, 80% (power). If the ‘a priori’ significance level is .05, then the probability of a false positive is less than 5%, and the probability of missing a clinically significant true positive (ES>CV) is much less than 20%. All this occurs *before* the data to be used in the HTS is accessed ('a priori').The study is then executed as designed (with “fidelity”), and the p-value computed as proposed. The statistical results of an HTS are then expressed by that p-value, along with a sample estimate of the ES, with its 95% confidence interval.

REPLICATION OR APPLICATION

No matter how well designed and executed a single study, no matter how small the p-value or large the ES, independent confirmation and replication is necessary. But what is replicated? Certainly not the p-value or the event that the p-value<.05, for those largely depends on sample size, not the strength of the hypothesis (7). The ES should replicate over studies.

If multiple studies testing a hypothesis (same population, research question, outcome measure), all valid for that purpose, each report an estimate of the effect size, there should be no significant heterogeneity among those effect sizes. Using meta-analytic methods, one can then compute the pooled effect size and its 95% confidence interval.

If effect sizes are heterogeneous (indicating non-replication), the question is whether all studies included addressed the same research question, and, if so, whether they were all valid for that research question (see, e.g.,(8)). The so-called "apples and oranges" and "garbage in, garb-age out" problems continue to plague meta-analysis (9, 10).

If there is replication (relatively homogeneous effect size), there are 4 possibilities for the pooled results:

1) The 95% confidence interval contains only ESs greater than CV. This result is both statistically significant (p-value<.05) and clinically significant. No further studies are needed.
2) The confidence interval contains only ES less than CV. This result may or may not be statistically significant, depending on whether the null value of zero lies within the confidence interval. However, the ES is not clinically significant. Time and resources should not be wasted on further such studies of this issue.
3) The confidence interval does not contain ES=0, but contains both ES both greater than and less than CV. This result is statistically significant, but not necessarily clinically significant. More studies are needed to settle that issue.
4) The confidence interval contains ES=0 as well as ES> CV.

Situation 4 may well happen with an individual study, particularly an inadequately powered one. However, with three or more valid and adequately powered studies, Situation 4 should rarely, if ever, happen in meta-analysis. If valid, but inadequately powered, studies are included in the meta-analysis, it may take far more studies to arrive at Situations 1 or 2. If invalid studies are included, there may never be a correct resolution to the question. However, studies included in meta-analysis often address different research questions (heterogeneous effect sizes), and many are not valid or adequately powered (8).

Zhang and Hughes focus only on HTS, and within that context, we largely agree. A few points of disagreement, however, with regard to the Table 1:

First, Odds Ratio (OR) is not a viable choice for ES. There is no sample size large enough to have more than 80% probability of detecting any OR>CV, whatever CV>1 is chosen (11, 12). Moreover, any non-null OR is un-interpretable in terms of clinical impact and is often grossly misleading (13-16).

Second, both significance level (α) and power are primarily determined by sample size, not by ES. I would argue that no decision could be based on these alone.

Finally, genome-wide association studies should not be regarded as HTS, but as HGS. When a specific gene or gene combination is found in such exploration that predicts disorder status, then a HTS should be proposed, designed and executed to test the hypothesis that that specific gene or gene combination actually predicts disorder status in the population of interest.

Historically, these p-value problems took on greater salience when statistical computation packages became more readily accessible and very powerful. Then multiple testing, appropriate in HGS, but questionable in HTS, and 'post hoc' hypothesis testing became more common. These issues now take on even greater salience with the advent of “big data” into medical research. With sample sizes in the thousands and, sometimes, millions, even the most trivial finding has p-value<.5.

The choice is either to fix the problems we have created or to give up on statistical hypothesis testing and p-values altogether. Personally, I am willing to do so, but not eager to give up what has long been (properly used) a valuable tool.

Helena Chmura Kraemer, Ph.D.
Professor of Biostatistics in Psychiatry (Emerita)
Department of Psychiatry and Behavioral Sciences
Stanford University
Stanford, CA
Email: hckhome@pacbell.net

CONFLICT OF INTERESTS

The author declares no conflict of interest regarding the publication of this paper.

REFERENCES

1. Zhang F, Hughes C. Beyond p-value: the rigor and power of study. Glob Clin Transl Res. 2020;2(1):1-6.
2. Behrens JT. Principles and procedures of exploratory data analysis. Psychological Methods. 1997;2(2):131-60.
3. Greenhouse JB, B.W. J. Exploratory statistical methods, with applications to psychiatric research. Psychoneuroendocrinology. 1992;17(5):423-41.
4. Tukey J. Exploratory Data Analysis. Reading MA: Addison-Wesley; 1977.
5. Ioannidis JPA. Contradicted and initially stronger effects in highly cited clinical research. Journal of the American Medical Association. 2005;294(2):218-28.
6. Ioannidis JPA. Why Most Published Research Findings are False. PLoS Medicine.2005;2(8):696-791.
7. Hedges LV, Olkin I. Vote-counting methods in research synthesis. Psychological Bulletin. 1980;88:359-69.
8. Mirosevic S, Jo B, Kraemer HC, Ershadi M, Neri E, Spiegel D. "Not just another meta analysis": Sources of heterogeneity in psychosocial treatment effect on cancer survival. Cancer Medicine. 2019;8(1):363-73.
9. Wortman PM. Judging Research Quality. In: Cooper H, Hedges LV, editors. The Handbook of Research Synthesis. New York: Russell Sage Foundation; 1994. p. 97-109.
10. Ioannidis JPA. The Mass Production of Redundant, Misleading, and Conflicted Systematic Reviews and Meta-analyses. The Milbank Quarterly. 2016;94(3):485-514.
11. Kraemer HC, Blasey C. How Many Subjects? Statistical Power Analysis in Research (Second Edition). Los Angeles, CA: Sage Publications; 2015.
12. Cohen J. Statistical Power Analysis for the Behavioral Sciences. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
13. Newcombe RG. A deficiency of the odds ratio as a measure of effect size. Statistics in Medicine. 2006;25:4235-40.
14. Kraemer HC. Reconsidering the Odds Ratio as a Measure of 2X2 Association in a Population. Statistics in Medicine. 2004 ;23(2):257-70.

15. Kraemer HC, Kazdin AE, Offord DR, Kessler RC, Jensen PS, Kupfer DJ. Measuring the potency of a risk factor for clinical or policy significance. Psychological Methods. 1999;4(3): 257-16.
16. Sackett DL. Down with odds ratios! Evidence-Based Medicine. 1996;1:164-6.

How to cite:

Kraemer, HC. A Comment on “Beyond P-value: the Rigor and Power of Study." Glob Clin Transl Res. 2020;2(1):7-9.
DOI: 10.36316/gcatr.02.0022.

Commentary

Improve Reproducibility by Using Statistical Methods Appropriately

Shiying Wu

The author (1) made some good suggestions, such as involving a statistician throughout the entire study process, from the study design to statistical reporting, and a more detailed report on statistical methods used, including the study design, power calculation, effect, and sample size.

There are several common and important aspects of the issue of false positives and reproducibility that can be further clarified. Among them are:

First. The false positives in biomedical studies are often due to a lack of knowledge of appropriate statistical methods. In particular, the false positives in "-omics and genome-wide association studies are often due to failure to account for multiple comparisons properly. In these types of studies, a large number of features are under examination, and the number of true positives could be very small if any. Instead of using p-values, Beniamini, and Hochberg (1995) developed a method to control the false discovery rate explicitly, and the method itself is called “false discovery rate” (FDR). Their method is widely used. Split sample validations and other cross-validation methods are also widely used to reduce the number of false positives and increase the reproducibility of the studies. In this case, the validation step is critical in reducing false positives and improving reproducibility.

Second. The “overpower” of a statistical test discussed in the paper (1). It is due to a lack of a proper specification of effect size in the null hypothesis. Let us say you want to see if a new drug on hypertension is effective. Your hypothesis may be the blood pressure difference in means between treatment and control being 0 mmHg. Given enough sample size, a real difference of 1 mmHg will be detected with a substantially small p-value. However, 1 mmHg is not medically significant and thus does not justify a new drug. Thus, a small p-value from your test does not lead to a meaningful conclusion in this case. Let us say that a difference of 10 mmHg or more is sufficient to justify the new drug medically. Then the right "null hypothesis" should be the difference being less than 10 mmHg. When the hypothesis is properly specified, the large sample size will not lead to incorrect inference. "Overpower" exists only in the sense of wasting resources.

Third. Should a smaller p-value be chosen as a significance level, or should the p-value be abandoned? This can be considered at two levels: First, under the current statistical framework, the p-values as control of false-positive errors are valid when used properly. However, the values of 0.05 or 0.01 were chosen somewhat arbitrarily by Fisher to simplify the distribution tables. A different p-value can undoubtedly be used when proper justification is provided. For example, in multiple comparisons problems like those in -omics and genome-wide association studies, the FDR method and cross-validation are often used to control false positives. In this case, the threshold for p-values is dictated by the desired false discovery rate and is no longer 0.05 or 0.01. However, blindly reducing p-value threshold increases the number of false negatives, which could be more expensive in the end. Be aware of the risk. Second, a decision to take a certain action or not should be ideally based on minimizing the risk given a well-thought loss function, instead of an arbitrary statistical significance level. The FDA's Accelerated Approval regulations can be thought of as an effort towards this direction. When there is no alternative treatment (thus no opportunity cost), a chance of saving lives out-weighs wasting money and suffering some side effects. This can be formalized by using an asymmetric loss function as opposed to the square loss function implicitly used in a typical statistical analysis from where the p-value is obtained(Given the form of the loss of function is subject to different choices, a sensitivity analysis of the choices may be needed). Only in this sense, p-values can be abandoned and replaced by minimizing the risk function. That said, people in the research community are not used to think decisions in terms of loss functions, let alone to achieve some level of consensus. Hence, there is a long way to go in this direction. Mean-while, we need to keep using p-values appropriately and justify the chosen threshold.

In addition, I do not entirely follow the decisions in Table 1. On Decision 2): The lack of power only suggests an increased probability of false negatives. Since we are observing a positive, I do not know why "null hypothesis H_0" is not rejected.

On Decision 3): Having a high power is not a problem per se, statistically, and should not lead to a wrong inference. If heterogeneity is a concern, subgroup analysis should be performed in both Decisions 1) and 3), because you may be able to detect it if it is substantial. Thus, I expect Deci-sions 1) and 3) be the same unless a different justification is provided.

Shiying Wu, PhD
SAS Institute
Cary, NC
USA
E-mail: shiying.wu@sas.com

CONFLICT OF INTERESTS

The author has declared no conflict of interest regarding the publications of this paper.

REFERENCES

1. Zhang F, Hughes C. Beyond p-value: the rigor and power of study. Glob Clin Transl Res. 2020;2(1):1-6.
2. Benjamini Y, Hochberg Y (2005) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol. 1995, 57: 289-300.

How to cite:

Wu S. Improve reproducibility by using statistical methods appropriately. Glob Clin Transl Res. 2020;2(1):9-10.
DOI: 10.36316/gcatr.02.0023.

Commentary

A Comment on "Beyond P-value: the Rigor and Power of Study”

Angelean Hendrix

Interdisciplinary work has unique challenges regardless of the fields involved. A free flow of ideas and advances can be almost impossible due to the vast amount of information available and the use of different scientific languages. Pharmaceutical development faces this challenge on a daily basis. Though the statistical community has been debating and improving their methods for decades, implementation outside of the field and academia has been slow to follow. One of the first statistical measures that I personally used was the p-value, a principle that all experimentalists must follow if their work is to be taken seriously, but statisticians know it is not enough and results are routinely published that are not replicable. “Beyond P-value: the Rigor and Power of Study” by Zhang and Hughes (1) boils down the American Statistical Associations position from 2016 into terms that are easily understandable to a wide audience of scientists. The proposed criteria have the potential to easily fix many problems associated with poorly designed studies and goes further to strongly suggest engagement with statisticians at every step from bench to bedside. The proposal is well presented, easily understood and long overdue.

Changes to statistical design, as proposed, address a critical scientific need, but a larger communication issue exists in interdisciplinary work. Fundamentally, it is impossible to stay current on all the advances in one’s own discipline much less those adjacent. Proficiency should not replace expertise, and, when the implementation of significant changes in theories is delayed for years, we must consider the consequences to the end beneficiaries of our work: the patients. Increased engagement with statistical and mathematical experts during the earliest stages of animal study design can streamline an entire development program and avoid future problems of reproducibility by helping make sure every study is rigorously designed and powered.

Angelean Hendrix PhD
Metabolism – Pharmacokinetics/Pharmacodynamics
The Covance Laboratories
Madison, WI.

CONFLICT OF INTERESTS

The author has declared no conflict of interests regarding the publications of this paper.

REFERENCES

Zhang F, Hughes C. Beyond p-value: the rigor and power of study. Glob Clin Transl Res. 2020;2(1):1-6.

How to cite:

Hendrix A. A Comment on "Beyond P-value: the Rigor and Power of Study”. Glob Clin Transl Res. 2020;2(1):9-10.
DOI: 10.36316/gcatr.02.0024

Authors' reply

Reply to Comments on "Beyond P-value: the Rigor and Power of Study"

Fengyu Zhang*, Claude Hughes

We received commentaries that included comments on our paper. Here we provide replies to the comments in the order received. To avoid confusion, we made a revision in the Table 1 to assure that level of significance and power at specified effect size are used appropriately.

Reply to Kraemer

We agree with Helena Kraemer's comments (1) on our paper "Beyond p-value: the rigor and power of study"(2), and that she also discussed some additional cases such as hypothesis-generating studies (HGS) or exploratory research.

There are studies when p-value(s) are calculated or reported, but they may not be applicable for formal statistical hypothesis testing. In such a case, misuse and misinterpretation of p-value may occur. Our paper mainly focu-sed the discussion on how p-values should be appropriately used and interpreted for statistical hypothesis testing studies (HTS), and then we proposed new criteria to reduce the possible misuse of p-values and reduce irreplicability in research. As we mentioned, "the interpretation of the p-value should consider the type of study"(2). We are also more than willing to agree with that one should not be eager to give up something until there has been a systematic effort to fix the problems. Our paper attempted to present actionable steps toward reducing the misuse or misinterpretation of the p-value that has been taught or used nearly for a century after Fisher's initial use in 1925, or the notion to abandon "significance" that Fisher introduced and associated the term "significance" with a small p-value (3).

The hypotheses generating studies (HGS) highlighted by Kraemer (1) are worthy of further comment. HGS are usually based on use of an existing database that often times was not formally designed for research or use of datasets that were conveniently collected and are explored for novel hypotheses. Nowadays, more and more inquires are made into sets of "big" data, such as those from electronic health records (EHR) or from "real world" clinical treatments in biomedical studies. When dataset from such resources lack an approach to an objective sampling from a defined study population, they should be used with care. As we pointed out in our paper on "Sampling and Hypothesis Testing" (2), p-value and hypothesis testing are established based on sampling, a technique used to obtain a sample to represent a study population objectively. Therefore, the parameter estimation, calculation of statistic, and hypothesis testing are all based on an objective sample, which helps to correctly estimate the sampling error associated with an outcome of study. If a study has not used an objective sampling, then the investigators should be cautious about using the p-value to decide against "null hypothesis H_0". We agree with Dr. Kraemer that for HGS, no conclusion should be made based on p-value if one wants to calculate or abandon it.

In terms of replication and application of results, Kraemer discussed both homogenous and heterogeneous cases of effect size for combining results (1). Her guidance is constructive in informing how to make the decision based on multiple replication studies. We strongly propose that when there is heterogeneity in effect size, no combined analysis should be performed, unless one has an adequate number of studies or replications of the same kind to estimate the random-effect appropriately. Dr. Kraemer's comments on the Table 1 are helpful and worthy of some explanations. We agree with the first point that "the odds ratio (OR) is not a viable effect size." and that ORs have often been misused and interpreted inappropriately. We listed (2) "coefficient or OR" under the column "effect" in the Table 1 to indicate two types of parameter estimates for continuous and non-continuous outcomes, not for a real effect size. Strictly speaking, these should have been labeled as parameter or estimate. We had made a note underneath the Table 1, including relative risk and hazard ratio (should have included other estimates such as correlation coefficient as well). Investigators should consider their individual study design in order to specify an effect size for power calculation and to use these new criteria for hypothesis testing.

We also agree with Dr. Kraemer on the second point that (1)"both significance level (α) and power are primarily determined by sample size, not by ES." These are the critical factors as we proposed to use level of significance (α) and power under pre-specified effect size to avoid studies that are too large or too small. In the paper (2), we stated, "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)". Therefore, level of significance (α), statistical power, and a pre-specified or meaningful effect size are all required components to implement the new criteria to decide on the "null hypothesis H_0." Effect size is critical for computing power, which should not be based on a post-hoc power

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

analysis. We did not list an effect size in the table because the choice of an effect size is a variable with study. Only after a meaningful effect size is specified, can appropriate power be calculated and divided into three categories to implement the new criteria. A large sample at some specified effect size could lead to an overpowered study. To avoid further confusion, we would propose a revised Table 1 (Table 1R) as follows.

Table 1R. Criteria for "Null hypothesis H_0" Based on P-value and Power at Specified Effect Size[a]

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

a. Could include regression coefficient, odds ratio (OR), relative risk (RR), HR (hazard ratio), or correlation coefficient.

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

Reply to Wu

In response to our recent paper (2) "Beyond p-value: the rigor and power of a study"), Shiying Wu discussed (4) "false positives and reproducibility that can be further clarified." We certainly agree with his first point that "false positives in biomedical studies are often due to a lack of knowledge of appropriate statistical methods." Misuse of statistics is commonly seen and is a significant problem to solve. His second point is that "an overpowered study is due to a lack of a proper specification of effect size in the null hypothesis." We also agree with this remark and believe that the proposed new criteria in our paper may help to reduce the conduct and reporting of some overpowered and underpowered studies. In our article (2), we stated, "We propose the criteria for making a statistical decision on the null hypothesis H0 based on both p-value and power, with a specified or meaningful effect size (Table 1)". The effect size is one criterion required for the power calculation and hypothesis testing. By specifying an effect size, one can judge whether a study is overpowered, regardless of whether the *p*-value is significant or not. Therefore, considering power can avoid *p*-value cha-sing in the conduct of research.

In terms of proposals for lowering the level of or abandoning the p-value, Dr. Wu noted that one should reduce the level of the p-value to accommodate multiple testing. We again agree and add that in a well-designed GWAS, a very low p-value should be used to control for this issue of multiple testing. In contrast, in some clinical studies, particularly randomized clinical trials, investigators may need to limit multiple testing because lowering the threshold for a p-value would be costly and potentially require more resources than are available.

In analysis of a clinical trial, a core recommendation is to apply a statistical test first to the primary outcome. According to a follow-up examination of 203 clinical trials (5) published in three major medical journals in 2017, 272 primary outcomes (1.34 per trial), 174 were significant at p<0.05, but only 71% (123/174) remain significant at p< 0.005. The proportion of significance at the new threshold (p<0.005) was 68.6% (59/86) in the industry-funded studies, which was significantly higher than 33% (38/115) in others. The funding source was the only significant characteristic associated with the rate of new significance after adjustment for other covariates. Wu also cited the case of "the FDA's Accelerated Approval regulations" when there is no alternative treatment for patients, which can be considered as an effort towards abandoning the p-value. In the case of no available approved medication, there may be no data available for ma-king a comparative evaluation of the new treatment. If data is accumulated and analyzed, it should belong to hypothesis-generating studies, as Kraemer discussed in her comment.

We would like to further reply to the questions about the new criteria we proposed for statistical hypothesis testing. First, we agreed with the "lack of power suggests an increased probability of false-negative study." When one positive result is observed with an underpowered study, it is likely that the observed positive is not real or it reflects an error because a study with a small sample is more sensitive to mistakes, which could be just measurement or sampling errors. In this case, more samples can assure that the statistical power reaches a minimal level such as 80%. Other authors have also discussed underpowered studies (6, 7) in which they noted that the error rate at a significance level of 0.05 may still be considerable. Regarding this issue, decision 2 in our Table 1 will help avoid publishing an underpowered study. In some situations, it may be impossible to add additional samples in which case the investigators should note in the study report or publication the fact that the study was underpowered and the underlying reason(s). To illustrate, in the conduct of some clinical trials, under powering may sometimes be caused by participants/subjects being lost to follow-up.

Finally, we would like to take this chance to elaborate more about the issue of overpowered studies, regarding the remark that "having a high power is not a problem *per se*, statistically, and should not lead to a wrong inference." *First,* an overpowered study is defined as a study with a large sample size at a specified effect size, regardless of how small the p-value is or if a study achieves a significant result or not. Generally, we all understand that people are less motivated to report a non-significant result. Research resources have limits, so overpowered research should be avoided because it may be costly and wasteful of resources, in particular for clinical and "Omics" research. *Second,* a large sample is likely to be accompanied by high heterogeneity. Therefore, findings detected in such an overpowered study with a large sample size tend to have a small effect size. More importantly, sometimes an overpowerred study may hide some valid findings with meaningful effect size(8) that could or may have been discovered with an adequate but homogenous sample. For our decision 3 proposed in the new criteria for hypothesis testing (2), when the criteria are met, one should perform a subgroup analysis to assess heterogeneity. The sample should be divided into different groups where each group plausibly represents possible diverse study subpopulations. We should usually insist that without subgroup analyses or sensitively analysis, no decision should be made in the case for the Decision 3 of the proposed Table 1. In commenting our paper, Dr. Kraemer has provided insightful comments on the heterogeneous and homogenous effect size for replication and application (1). Excellent examples have been reported demonstrating that subgroup analyses might give beneficial information for detecting sample heterogeneity (9). Also, ancillary analysis has been required as a relevant item in the checklist for preparing a report of a randomized clinical trial (10).

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interest regarding the publication of this article.

ACKNOWLEDGEMENTS

The authors thank all authors for three commentaries.

REFERENCES

1. Kraemer HC. A comment on “Beyond p-value: the rigor and power of study". Glob Clin Transl Res. 2020;2(1):7-9.
2. Zhang F, Hughes C. Beyond p-value: the rigor and power of study. Glob Clin Transl Res. 2020;2(1):1-6.
3. Fisher R. Statistical Methods for Research Worker. Edingburg: Oliver and Boyd; 1925.
4. Wu S. Improve reproducibility by using statistical methods appropriately. Glob Clin Transl Res. 2020;2(1):10-1.
5. 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(17):1813-5.
6. Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p-values. R Soc Open Sci. 2014; 1(3):140216.
7. Loiselle D, Ramchandra R. A counterview of 'An investigation of the false discovery rate and the misinterpretation of p-values' by Colquhoun (2014). R Soc Open Sci. 2015; 2(8):150217.
8. Ioannidis JP. The Mass Production of Redundant, Misleading, and Conflicted Systematic Reviews and Meta-analyses. Milbank Q. 2016;94(3):485-514.
9. International Schizophrenia C, 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.
10. Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340:c869.

How to cite:

Zhang F and Hughes C. Reply to comments on "Beyond p-value: the rigor and power of study**"**. Glob Clin Transl Res. 2020;2(1):13-15. DOI: 10.36316/gcatr.02.0025.

Introduction to the editors

Editor-in-Chief

Claude Hughes, MD, Ph.D. holds current Board Certifications in Obstetrics and Gynecology and Reproductive Endocrinology and Infertility from the American Board of Obstetrics and Gynecology. Since joining Quintiles /IQVIA in 2001. Dr. Hughes has served as a Medical Advisor on clinical trials or in due diligence assessment teams that evaluated pharmaceuticals, devices or tests for multiple medical indications. Before joining Quintiles, Dr. Hughes held academic, research, administrative and clinical practice positions for 15 years in divisions of reproductive endocrinology & infertility in departments of obstetrics & gynecology and clinical and research centers within university-affiliated medical centers. His leadership roles included Director of the Reproductive Hormone [hormone assay service] Lab at Duke University for ten years; Section Leader, Department of Comparative Medicine at Wake Forest University, Director of the Center for Women's Health at UCLA-Cedars Sinai Medical Center, and Vice President & Chief Medical Officer at RTI International.

Fengyu Zhang, Ph.D., is currently Chief Scientific Officer and Co-director of Global Clinical and Translational Research Institute, Bethesda, MD and holds distinguished adjunct professorships at the Second Xiangya Hospital of Central South University in Changsha and Peking University Huilongguan Clinical Medical School in Beijing, and invited distinguished visiting professorships at multiple universities. He had served as director of statistical genetics and senior genetic epidemiologist at National Institute of Mental Health's Genes, Cognition and Psychosis Program, and Investigator in genetics, bioinformatics and epigenetics at Lieber Institute for Brain Development, Johns Hopkins Medical Campus, MD. Dr. Zhang is a member of the American Association for the Advancement of Science, International Society for Developmental Origins of Health and Disease, the Society for Neuroscience, the Society for Biological Psychiatry, and Internationals Society of Psychiatric Genetics. His research interests include the etiology of human complex disorders, pharmacogenomics, population health and methodology in clinical and translational research.

Associate Editors

Donald Mattison, MD, is Chief Medical Officer and Senior Vice President of Risk Sciences International. Dr. Mattison also serves as Associate Director of the McLaughlin Centre for Population Health Risk Assessment at the University of Ottawa, Canada. He has held academic, clinical and research appointments, including; Senior Advisor to the Director of the Eunice Kennedy Shriver National Institute of Child Health and Human Development, Medical Director of the March of Dimes; Dean of the Graduate School of Public Health at the University of Pittsburgh, Professor of Obstetrics and Gynecology and Interdisciplinary-nary Toxicology at the University of Arkansas for Medical Sciences, and Director of Human Risk Assessment at the FDA National Center for Toxicological Research. He was elected a Fellow of the American Association for the Advancement of Science, a Fellow of the New York Academy of Medicine, and a member of the Institute of Medicine, Distinguished Alumni of Augsburg College and a Fellow of the Royal Society of Medicine.

Hong-Wen Deng, Ph.D., is Endowed Chair and Professor of Biostatistics at School of Public Health and Tropical Medicine at the Tulane University. Deng's work is published in more than 500 peer-reviewed publications including journals such as Nature, New Engl J of Medicine, American Journal of Human Genetics, Endocrine Review, PLoS Genetics, Human Molecular Genetics, Molecular Psychiatry, Bioinformatics, and Molecular Cell Proteomics

Xu-Feng Huang, MD, Ph.D., DSc, is Distinguished Professor, School of Medicine; Theme Leader of Mental Health (Clinic) of Illawarra Health and Medical Research Institute; Director, Centre for Translational Neuroscience; Faculty of Science, Medicine, and Health, University of Wollongong, Australia.

Dr. David St Clair, MD Ph.D., is Professor in Psychiatry at the University of Aberdeen in Scotland, UK and Honorary Consultant Psychiatrist with NHS Grampian. He has made a significant contribution to the genetics of schizophrenia and is part of both the International Schizophrenia and SGENE consortia.

Susan Sumner, Ph.D., is Professor of Nutrition at the University of North Carolina at Chapel Hill and Director of the NIH Eastern Regional Comprehensive Metabolomics Resource Core. Her expertise is in metabolism and metabolomics, and broad applications in studies of diet, smoking, cancer, diabetes, obesity, cognitive development, liver disease, natural products, maternal and child health, and the environmental influence of disease complements the nutrigenomics research.

Riqiang Yan, Ph.D., is Professor and Chair of Neuroscience at The University of Connecticut School of Medicine. He is Morris R and Ruth Graham Endowed Chairs in biomedical sciences, recipient of MetLife Award for Medical Research (the prestigious award for Alzheimer's Research), Award for outstanding science at Cleveland Clinic Foundation and Ralph Wilson Award.

Hui Zhang, Ph.D., is Associate Professor at St Jude Children's Research Hospital, TN. His research interests focus on longitudinal data analysis, survival data analysis, computational neuroscience and count data in next-generation sequence.

Youwen Zhou, MD Ph.D., is Professor and physician-scientist in dermatology at the Department of Dermatology and Skin Science of the University of British Columbia. Dr. Zhou is the past president of Canadian Society of Investigative Dermatology and served as an ad board member for CIHR Institute of Musculoskeletal Health and Arthritis (IMHA). Dr. Zhou has received multiple national and international awards, including Barney Usher Award for Outstanding Achievements from the Canadian Dermatology Association.