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. Rese-

arch 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.