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 (capita-lization 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 INTEREST

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.