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 interests 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.