In Search of the Significant p. Its Influence on the Credibility of Publications

En busca de la p significativa. Su influencia sobre la credibilidad de lo publicado

JORGE THIERER

Publishing study results in a peer-reviewed journal represents the ultimate goal of research in any field of science and it is obviously assumed that the results are correct and supported by a careful analysis.

A curious phenomenon is that most publications present "successful" studies, in which the initial hypothesis is ratified by its demonstration. A review by Fanelli refers that 84% of published studies demonstrate what the introduction presents as initial hypothesis. (1) This fact is remarkable. However, it is worrying that the amount of publications in which findings can be reproduced is low, from cancer (2) to psychological research. (3)

In Medicine, as in other disciplines, the hypothetico-deductive method is the basis of research. A plausible theory generates an "a priori" hypothesis which will be tested in a clearly designed experiment considered the best way to demonstrate it. A defined statistical test (according to the type of variable compared and the number and distribution of values) will result in a p value. But what is the meaning of p?

WHAT DOES P MEAN?

Any result of any variable is part of a probability distribution. It is probable that some results occur more frequently than others. The classical example of probability distribution is the Gaussian curve, where the values nearer the mean have greater probability of occurring, while those which are farther away from the mean occur much less frequently.

In any comparison between an experimental and a reference value, or between a value from a group compared with that from another group, it is possible that in each situation both values belong to the same probability distribution and, therefore, that hazard is responsible for the different results found.

The meaning of p is not the same in a significance test than in a hypothesis test. (4)

The p value is important in a significance test: it represents the probability of finding values (or differences between values) as those obtained or more extreme if the null hypothesis is true. In this type of test (known as Fisher´s test, in honor of Fisher, who reported the concept in a classical text in 1925) the value of p is a continuum, and the lower it is, the lower the probability of finding the explored values under the null hypothesis. Thus, for example, when exploring the difference of means between two groups, if the p value is 0.10, we will say that if the null hypothesis is true (i.e. if there is truly no difference between both groups), the probability of finding the difference obtained is 10%.

In a hypothesis test, which in general dominates the rules accepted for research, a binary approach is used to evaluate p; what matters, in general, is whether or not the p value is <0.05. This evaluation of p was proposed by Neyman and Pearson in 1933. Data are collected and a hypothesis test is postulated which considers the null hypothesis or an alternative hypothesis.

If chance led to different values within the same distribution, there is no significant difference between them. This is what the null hypothesis formulates. The opposite is the alternative hypothesis, which states that there is a difference between the value found and the reference one, or between the value in one group and that in the other. For this hypothesis, therefore, the values do not belong to the same distribution, bur to different populations.

How to decide which of the two hypotheses is the correct one? It all depends on the null hypothesis and whether to reject it or not. A statistical test for each case establishes, in the distribution of probabilities corresponding to the variable under study, where the value we are considering lies with respect to the one used for comparison, and clarifies whether there is or not a statistically significant difference.

This test informs us whether in the distribution of probabilities, the value we are interested in with respect to that of the comparator lies in the area of no rejection or of rejection of the null hypothesis. By a convention criterion it is assumed that the null hypothesis cannot be rejected if the value found lies within 95% of the probability distribution which has as reference the population or the comparator group. We accept that in this case the difference is due to chance. Conversely, if the value obtained lies outside this 95%, if it is part of the 5% extreme values, the null hypothesis is rejected.

When in the comparison between two values the test reports that the probability that they belong to the same distribution is less than 5%, we say that,

yes, the difference is statistically significant. This cut-off value to reject the null hypothesis (5%, 1 in 20, p=0.05) expresses a convention.

Furthermore, beyond the test, in the real situation, it is assumed that the null hypothesis may be true or false. Moreover, the statistical test may be right or wrong when describing this reality. Four situations are then postulated. (5)

- If there is a real difference and the test detects it by rejecting the null hypothesis, the result of the test is a true positive.
- If the difference is not real, and both values correspond to the same probability distribution, but the statistical test rejects the null hypothesis, establishing that the p value is <0.05, we are in the presence of a false positive result. We shall call this erroneous result an alpha or type I error. If the false positive rate for the test is 5%, its specificity, or true positive rate is 95%: 1-alpha error. The lower the alpha error value, the more specific the test is.
- If there is no difference, both values belong to the same distribution and the test does not reject the null hypothesis; in this case we are in the presence of a true negative.
- If there is a real difference, but the test does not detect it, i.e. it does not reject the null hypothesis, we are in the presence of a false negative result and thus consider that a beta error or type II error has been produced. If we consider a beta error, we obtain the power of the test, which represents its sensitivity or ability to detect a real difference. So, as an alpha error of 0.05 or 0.01 is usually admitted, the beta error defined a priori for a comparison can be 0.10 (for 90% sensitivity) or 0.20 (if 80% sensitivity is admitted).

For the same number of observations alpha and beta errors operate in an opposite manner. As with any diagnostic test, if the sensitivity or power increases, hence the beta error decreases. On the contrary, if the alpha error decreases, looking to be more specific (for example, choosing p <0.01) the power of the study to detect a difference is reduced. Certainly, the exploratory phase of a study seeks greater sensitivity (greater power is admitted and hence, more false positives are admitted) while to confirm the hypothesis specificity is favored a priori. (6)

As we can see, Fisher's concept understands the p value as a continuum and Pearson's concept in a dichotomous way. Although there is growing consensus that the dichotomous concept can lead to serious errors and lacks flexibility to draw adequate conclusions about the study phenomena (if p=0.049, we accept the difference, while if p=0.051, we say there is no difference), the truth is that in evidence assessment and articles submitted for publication, the p value of 0.05 is still the universally accepted criterion to decide whether a result is positive or not. And although a p value of 0.005 has been postulated, and it is true that it reduces the false positive rate, it does not solve the problem definitively. (7) The use of a cut-off point seriously reduces the possibility of reexploring the result; if a study has led to p=0.04, who will repeat the experiment to see whether it is a false positive result? If p=0.08, how many will start a new study with the idea that it could have been a false negative finding?

In the last years a series of warnings have been reported regarding erroneous concepts about the significance of p and the error of trusting a hypothesis test, (8, 9) as well as ardent defenses of its use, (10) which will be the basis of a future Editor's Page. Nevertheless, we repeat, the concept of p <0.05 is still the cornerstone on which the hypothesis of a study with positive or negative result lies, and in the search of achieving the desired value a series of situations may arise conspiring against the credibility of findings.

Let us first consider a disruptive statement. Fifteen years ago, Ioannidis declared in a provocative article- that most published research findings were false. (11) A boutade or an indisputable truth? How to justify it? Let us assume that among 1,000 scientific hypotheses 10% are true and the remaining 90% not. If we work with an alpha level of 5% (i.e. a ratio of false positives is accepted when the 5% null hypothesis is true), over the 900 erroneous hypotheses a p value <0.05 will be obtained, and therefore, the null hypothesis will be rejected in 45. We will accept 45 times a hypothesis that is not correct over 900 explored ones! And if we work with 80% power to detect false positives, from 100 correct hypotheses only 80 will be detected. Hence, from 45+80=125 hypotheses accepted, 45 (36%) will be false.

If power were only 20%, false positive findings would rise to 53%; but if the proportion of correct hypotheses were not 10% but 50%, with 5% alpha error and 20% beta error, the proportion of false positive findings would be 11%. This means that the ratio of true and false positive findings over the total number of hypotheses depends on the accepted alpha and beta errors and on the proportion of truly correct hypotheses; i.e. the proportion of false null hypothesis. Knowing that the number of times we may be facing a false positive largely exceeds 5%, despite having used p <0.05, certainly demands extreme care to preserve the validity of published results.

MECHANISMS THAT ALLOW REACHING A SIGNIFICANT P VALUE

Multiple comparisons

If we explore two groups and assess a series of baseline characteristics looking for significant differences between them, and for each comparison we establish p <0.05 to reject the null hypothesis, i.e. we accept up to 5% of false positives, making several comparisons significantly increases this rate.

For example, if we make 10 comparisons assuming for each a p value <0.05 to define statistical significance, the probability of finding a positive result

purely by chance increases to 40%. In this context, assuming $p<0.05$ as significant is clearly erroneous. Therefore, when multiple comparisons are made, it is recommended to work with a lower than normal p value.

Let us assume we made 20 comparisons establishing a priori $p<0.05$. Among the 20 comparisons only one resulted with $p<0.05$, but we understand that it can be a false positive. Applying the Bonferroni correction 0.05 is divided by 20 resulting in $p=0.0025$ which is the necessary p value to reject the null hypothesis for each comparison. This approach considerably decreases the false positive rate, and should certainly be applied; but it is also true that with this correction, a p value of 0.003 becomes non-significant.

Usually, we do not know all the analyses that have been performed. If over 10 predictive models tested only one is adequate, that is the one reported. How is it possible to know that another 9 tests surreptitiously sleep the dream of the righteous? How to interpret the statistical significance of the positive finding if we ignore from how many comparisons it emerges and, therefore, what is the chance of facing a false positive?

In this context, the fallacy of incomplete evidence, or "Cherry Picking", justly refers to selecting variables or relationships that help to validate a hypothesis, suppressing those that could not demonstrate it. It is a variation of the multiple comparison phenomenon, with a biased selection of the evidence. It is also called fallacy of incomplete evidence.

It is not easy to solve the problem. (12) Perhaps, part of the tension lies in the need to reject or not the null hypothesis, discarding findings in which $p \geq 0.05$. Another point deals with differentiating whether the variable which appears as statistically significant is independent or not from those explored. If it is not, clearly the adjust must be made; if it is, we could accept the result obtained.

P HACKING

This definition involves a series of procedures aimed at obtaining a significant p value, when this was nor achieved after the initial analysis. (13) It is understood that the p values must be close to 0.05. These actions would not be performed if initially $p=0.45$. We assume that an initial value that prompts this conduct oscillates between 0.05 and 0.10, or even, up to 0.20.

Among these actions we may identify: (4, 14)

a) Analyzing only one subgroup of data: This is only justified if there is a strong reason to exclude the rest, either because they are irrelevant for the postulated question, or because there are doubts about the information collected. Nonetheless, this procedure is extremely doubtful when carried out as a post hoc analysis, without a priori clear protocol definitions.
b) Excluding outliers (data far away from the sample mean). It is an acceptable practice if these data are wrongly collected, outside the initial inclusion criteria. It is not an acceptable practice if the data are correct and there is no reason to remove them, other than their exclusion reduces sample variability and as the standard error decreases, the statistical power increases, reducing false negatives, but increasing false positives.
c) Standardizing (for example, adjusting weight by height) and transforming logarithmically non-normal distribution data. These mechanisms increase the ability to achieve a significant p value, and their use is not objectionable, but must be declared a priori. If they are applied when the initial analysis was not satisfactory, they increase the chance of false positives because significance is sought a second time, establishing more than one comparison.
d) Increasing sample size. This is the most common advice authors from a study that did not achieve significant results receive. As we know, sample size is defined by four factors: the alpha error or acceptable false positive rate (p value), the beta error, or false negative rate, and hence the power of the test (1-beta error), the magnitude of the explored effect and its standard deviation which expresses effect variability. The lower the alpha error, the beta error and the magnitude of the effect, and the greater the variability, the higher the necessary sample size. If it cannot be demonstrated that a certain difference or an effect are statistically significant, it emerges that perhaps this is due to lack of sufficient power to do so. As the effect values and their variability are data that cannot be modified with the available sample data, and a p value <0.05 is sought, the only resource is enhancing power by increasing sample size.

Then, it is possible to calculate the number of observations needed to reach $p<0.05$, with the effect or difference found, and their variability. Although it can be postulated that the procedure is correct, it is not less true that we do not know whether the effect found and its variability are true estimates of the population effects and that, consequently, we may be doing a great effort to demonstrate something that is not real. On the other hand, when we return to seek significance with a larger number of data, we are increasing the false positive rate, since we are granting many observations the possibility of being tested in more than one opportunity.

Then, this increase in the false positive rate occurs under the assumption that the null hypothesis is true. If the null hypothesis is false, that is, if there is really a difference, what has been done by increasing power is to decrease the false negative rate. Is the null hypothesis true or false? That is the question. We do not know! Hence, among the mechanisms displayed to reach the sought p value, this is perhaps the least criticized. However, it is suggested that it would be adequate to mention this topic in the Discussion section of a paper, informing the readers of the difficulties encountered.

p hacking originates from a flexible analysis, in which the author has many degrees of freedom to achieve his purpose. (15) The author may, after the initial analysis, resort to one or more of the above-described procedures, and it is very difficult to detect its practice once the work has been published. (16)

HYPOTHESIZING AFTER THE RESULTS ARE KNOWN (HARKING)

In the last 30 years a new way of approaching research has progressively expanded. This is what Kerr, in his liminal article, (17) termed HARKing (Hypothesizing After the Results are Known). This practice consists in presenting in the introduction of a research study a hypothesis generated post hoc, once the data have been analyzed, as if it had been formulated a priori, before collecting them. A three-row, two column contingency table presents the 6 situations that result when contrasting the hypothesis before and after knowing the data,

The underlying idea of HARKing is that it does not matter if the sustainable hypothesis after the analysis has been previously formulated. What matters is that it is plausible. Under this assumption, among a, c and e hypotheses the most plausible may be chosen and postulated in the Introduction as if it had been done a priori. From the combination of pre- and post-study conditions, a series of variants emerge for Kerr, the most feasible being: suppress the plausible hypotheses that could not be demonstrated (if a priori a and b had been postulated, mention only a, generating the illusion that only what was thought was effectively demonstrated) or extend this condition to the nonplausible ones (present hypotheses a and c in the introductions) and even to those that had not been anticipated. The common condition to all these variants is that results are the ones that define which hypotheses will be presented in the Introduction appearing to be supported by evidence.

When can we suspect HARKing? A criterion is that the hypothesis presented as formulated a priori does not clearly emerge from the theory or paradigm in which the study is framed. This could be explained by lack of essential variables that would have been considered to demonstrate the hypothesis, provided their a priori condition is true, or because it is suspected that the methods used are not the ones that would have been thought in that circumstance. The truth is that some forms of HARKing we have mentioned are employed as much as the hypothetico-deductive approximation. In fact, abandoning a hypothesis -if the data so suggests it- although a priori it has been considered plausible, and focus in hypotheses that were not considered, but appear to be plausible, is a conduct recommended by many textbooks.

	Hypothesis after knowing the results	
Hypothesis before performing the study	Plausible	Nonplausible
Anticipated and plausible	a	b
Anticipated and nonplausible	c	d
Not anticipated	e	f

Why is HARKing resorted to? The are many reasons. In the first place, it is understood that there must always be a theory and a hypothesis that have preceded the study. And even when Popper has pointed out that ideally a theory should be contested, the truth is that ideally a theory formulated be confirmed by the study; it is what readers expect. And what researchers believe: in a survey, 89% of respondents understood that a test that confirms the hypothesis provides information, but only 39% thought the same of the test that does not confirm it. The communication of confirmatory results satisfies the need of researchers, readers and journal authorities of providing a report that soothes and gives the false illusion of scientific progress. If the result is positive, the study has greater chance of being published; more publications ensure more access to grants and success in the academic world. Publication rewards novelty and the idea of the original hypothesis validated by a study with "happy ending". Due to this bias of retrospective confirmation it is believed that the unexpected finding was something that in reality had always been known. That the hypothesis presented has not been a priori formulated is often a minor subject for most interested parties. And there are even editors, generally from journals which charge authors for their publication, who suggest this procedure as a way of facilitating it.

This preference for novel, fact-confirmed hypotheses takes root with the bad press the null hypothesis has. Studies pointing out that there is no difference, that new and known results are the same, that an intervention does not modify the established outcome, that a recently developed diagnostic method does not improve traditional resources, are not studies that motivate readers, imbued by the belief of constant progress.

The null hypothesis "must" be refuted at any cost. It is worth pointing out that, when several significance tests are simultaneously performed, all with $p<0.05$, there is a marked increase in the false positive rate (type I error), of which HARKing is a very frequent source. Hence, and ignoring there is a high chance of a false positive rate, the study result becomes an accepted truth.

In contrast to the hypothesis generated prospectively (called "prediction") is the one generated post hoc (which Norwich termed "accommodation"). It could be argued that, if both were equally examined and against the same body of evidence, there would be no difference in credibility between them. But it is more frequent that when a post hoc hypothesis is generated, the interpretation of the evidence becomes biased in favor of the most recent findings, creating a narrative supporting outcomes that were no suspect-

ed at the beginning of the study. A basic difference between HARKing and p hacking is that in the first case the initial hypothesis is not the one that is finally presented, and in the latter the hypothesis is preserved, but the initially expected analysis changes. (18)

Clearly HARKing is not as serious as data falsification or plagiarism. In fact, there are no ethical codes in research that explicitly condemn it. However, it is clear that its consequences can be taxing: information is lost, application of statistical methods is more lenient, and communication of results is not honest. A finding supported by $p < 0.05$ may lead to the defense of a line of research that leaves relevant information in the periphery.

Since the objective of demonstrating statistical significance has been achieved, the belief in what is presented is no longer important, favoring an attitude focused on the result, leaving aside the quest for truth. Pretending "to have found what was sought" neglects the possibility of recognizing serendipitous findings, (19) falls into the adoption of very narrow theories that justify exactly the findings or, conversely, faced with the inability to accurately explain them, builds diffuse theories that might generally explain almost everything and nothing in particular, abandoning the opportunity of generating alternative hypotheses. Different from falsification or plagiarism, it is not easy to detect or demonstrate the practice of HARKing, and the answer cannot be to believe that everything we read comes from it.

FINAL CONSIDERATIONS

We have only presented some situations that may lead to increase the publication of falsely positive results, purposefully leaving aside clearly dishonorable situations, as falsification and plagiarism. We focus on those in which, unintentionally, or because it is thought that good science is being performed interpreting that the actions do not conspire against the credibility of what is reported, the decisions adopted graze the assumptions which underlie the scientific method.

The proportion of researchers that practice them is high. For example, in a survey including 494 researchers in ecology and 313 in evolutionary biology, not reporting variables that were analyzed but did not reach statistical significance was acknowledged by more than 60% of them, the addition of new observations to increase power by more than 40% and having practiced HARKing by 50%. (14)

Why are these practices so common? Undoubtedly, there is an individual component: not all authors practice them. On the other hand, there are very honest researchers who do not find in their practice a really objectionable attitude. And there are those who understand that they are a normal part of the research procedure. Several cognitive biases (of observation, confirmation, retrospective comprehension, or belief) (20) may explain the unconscious trend to privilege, in the statistical analysis or when writing an article, the information that confirms their own beliefs, disregarding what questions them, and also to weave the speech that implicitly leads to publication.

But beyond personal aspects, we may stop -and this seems to be even more decisive- in the current system of ideas. If only positive results are rewarded, if the $p<0.05$ religion is uppermost, if those who most frequently reach this result have easier access to publication, if studies dedicated to replicate already published results are discouraged, if "publish or perish" is the first commandment, but the publication only rewards a significant p, the practices we have described will certainly not disappear.

A different attitude of editors, capable of accepting initial exploratory research communications, with the ability to detect HARKing and demand authors to acknowledge it, and of accepting studies that replicate already published results, are all actions that may lead to a more credible medical literature.

Beyond this point, and in general, if it were accepted that the finding may have originated from exploration, and the assumptions used to carry it out as well as the lines of research that might be derived from it were adequately explained, there would be enough academic space to publish and contribute to the progress of knowledge, although p were ≥ 0.05 (though, of course, we cannot avoid demanding up to what p level we would admit to publish a finding. Perhaps this should depend on the validity of the hypothesis, the importance of the clinical finding, and the certainty that it is a line of investigation that continues).

Then: on the one hand, a more favorable attitude for quality research, even if does not reach $p < 0.05$ And on the other? A firmer attitude, so when $p<0.05$ is declared it can be assured that the finding is true and not a false positive. In this sense, recommending publications in peer-reviewed journals is one of the available solutions. The decision of renowned reviewers as experts in their field of knowledge and blinded to authors' names may contribute to detect anomalies in data presentation and ensure the transparency of what is published.

Encouraging authors to acknowledge when they have added observations, eliminated variables or changed their hypothesis guided by unexpected results will give credit to publications. To be able to communicate what generally occurs in research, a road far from being free of obstacles, may illuminate the personal trajectory.

There are even firm proposals to submit the material used as basis for the publication to be reviewed even after its release. (21) Pre-registering the research protocol, clearly identifying the inclusion and exclusion criteria, the endpoints and how the work on the different variables will be carried out is, certainly, a very strong restriction to the also unconscious temptation of a posteriori manipulation of the data, changing the hypothesis or the endpoints. (22)

In the same line, an even more radical proposal, that few journals have put into practice, consists in submitting the database used to perform the analysis, if the study aims to be published. (23)

In conclusion, the ability of Congresses and Publications to put forward original ideas, even in the process of development, and at the same time to be more demanding so that when a finding is declared it can be supported by facts, are roads to be trodden. In the Argentine Journal of Cardiology we make our best effort to achieve it.

REFERENCES

1. Fanelli D. "Positive" results increase down the Hierarchy of the Sciences. PLoS One 2010;5:e10068. https://doi.org/10.1371/journal.pone.0010068

2. Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nat Rev Drug Discov 2011;10:712. https://doi.org/10.1371/journal.pone.0010068

3. Open Science C. PSYCHOLOGY. Estimating the reproducibility of psychological science. Science. 2015;349:aac4716. https://doi.org/10.1126/science.aac4716

4. Lew MJ. A Reckless Guide to P-values: Local Evidence, Global Errors. Handb Exp Pharmacol. 2020;257:223-56. https://doi.org/10.1007/164_2019_286

5. Akobeng AK. Understanding type I and type II errors, statistical power and sample size. Acta Paediatr 2016;105:605-9. https://doi.org/10.1111/apa.13384

6. Forstmeier W, Wagenmakers EJ, Parker TH. Detecting and avoiding likely false-positive findings - a practical guide. Biol Rev Camb Philos Soc 2017;92:1941-68.
https://doi.org/10.1111/brv.12315

7. Ioannidis JPA. The Proposal to Lower P Value Thresholds to .005. JAMA 2018;319:1429-30. https://doi.org/10.1001/jama.2018.1536

8. Goodman S. A dirty dozen: twelve p-value misconceptions. Semin Hematol 2008;45:135-40. https://doi.org/10.1053/j.seminhematol.2008.04.003

9. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature 2019;567:305-7. https://doi.org/10.1038/d41586-019-00857-9

10. Ioannidis JPA. Retiring statistical significance would give bias a free pass. Nature 2019;567:461. https://doi.org/10.1038/d41586-019-00969-2

11. Ioannidis JP. Why most published research findings are false. PLoS Med 2005;2:e124. https://doi.org/10.1371/journal.pmed.0020124

12. Goodman SN. Multiple comparisons, explained. Am J Epidemiol 1998;147:807-12; discussion 15. https://doi.org/10.1093/oxfordjournals.aje.a009531

13. Wicherts JM, Veldkamp CL, Augusteijn HE, Bakker M, van Aert RC, van Assen MA. Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. Front Psychol 2016;7:1832. https://doi.org/10.3389/fpsyg.2016.01832

14. Fraser H, Parker T, Nakagawa S, Barnett A, Fidler F. Questionable research practices in ecology and evolution. PLoS One. 2018;13(7):e0200303. https://doi.org/10.1371/journal.pone.0200303

15. Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci 2011;22:1359-66.https://doi.org/10.1177/0956797611417632

16. Ulrich R, Miller J. p-hacking by post hoc selection with multiple opportunities: Detectability by skewness test?: Comment on Simonsohn, Nelson, and Simmons (2014). J Exp Psychol Gen 2015;144:1137-45. https://doi.org/10.1037/xge0000086

17. Kerr NL. HARKing: hypothesizing after the results are known. Pers Soc Psychol Rev 1998;2:196-217.https://doi.org/10.1207/s15327957pspr0203_4

18. Wicherts JM. The Weak Spots in Contemporary Science (and How to Fix Them). Animals (Basel) 2017;7(12).https://doi.org/10.3390/ani7120090

19. Pepys MB. Science and serendipity. Clin Med (Lond) 2007;7:562-78. https://doi.org/10.7861/clinmedicine.7-6-562

20. Blumenthal-Barby JS, Krieger H. Cognitive biases and heuristics in medical decision making: a critical review using a systematic search strategy. Med Decis Making 2015;35:539-57. https://doi.org/10.1177/0272989X14547740

21. Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, et al. SCIENTIFIC STANDARDS. Promoting an open research culture. Science 2015;348:1422-5. https://doi.org/10.1126/science.aab2374

22. Yamada Y. How to Crack Pre-registration: Toward Transparent and Open Science. Front Psychol 2018;9:1831. https://doi.org/10.3389/fpsyg.2018.01831

23. Miyakawa T. No raw data, no science: another possible source of the reproducibility crisis. Mol Brain 2020;13:24. https://doi.org/10.1186/s13041-020-0552-2