NEJM Evidence

Published December 22, 2023

DOI: 10.1056/EVIDoa2300003

ORIGINAL ARTICLE

A New Look at P Values for Randomized Clinical Trials

Erik van Zwet, Ph.D.,[1] Andrew Gelman, Ph.D.,[2,3] Sander Greenland, M.D., Ph.D.,[4,5] Guido Imbens, Ph.D.,[6] Simon Schwab, Ph.D.,[7] and Steven N. Goodman, M.D., Ph.D.[8]

Abstract

BACKGROUND We have examined the primary efficacy results of 23,551 randomized clinical trials from the Cochrane Database of Systematic Reviews.

METHODS We estimate that the great majority of trials have much lower statistical power for actual effects than the 80 or 90% for the stated effect sizes. Consequently, "statistically significant" estimates tend to seriously overestimate actual treatment effects, "nonsignificant" results often correspond to important effects, and efforts to replicate often fail to achieve "significance" and may even appear to contradict initial results. To address these issues, we reinterpret the P value in terms of a reference population of studies that are, or could have been, in the Cochrane Database.

RESULTS This leads to an empirical guide for the interpretation of an observed P value from a "typical" clinical trial in terms of the degree of overestimation of the reported effect, the probability of the effect's sign being wrong, and the predictive power of the trial.

CONCLUSIONS Such an interpretation provides additional insight about the effect under study and can guard medical researchers against naive interpretations of the P value and overoptimistic effect sizes. Because many research fields suffer from low power, our results are also relevant outside the medical domain. (Funded by the U.S. Office of Naval Research.)

Introduction

How should researchers and clinicians interpret the P value for the null hypothesis of no effect from a randomized clinical trial (RCT)? This P value is commonly defined as the probability, under the null hypothesis and an assumed statistical model, that an appropriate test statistic would be as or more extreme than what was observed. Here, we will consider the absolute z statistic as a test statistic. We wish to reinterpret the resulting two-sided P value in light of background information about studies with similar statistical properties. The Cochrane Database of Systematic Reviews contains

The author affiliations are listed at the end of the article.

Dr. van Zwet can be contacted at E.W.van_Zwet@lumc.nl or at Biomedical Data Sciences, Leiden University Medical Center, Einthovenweg 20, 2333 ZC Leiden, The Netherlands.

the results of more than 20,000 RCTs in biomedicine. We have collected the absolute z statistics of the primary efficacy outcome for all of these RCTs.

Recall that the z statistic is the estimated effect divided by the standard error (SE) of the estimate. We also wish to consider the signal-to-noise ratio (SNR), which is the *true* effect divided by the SE of the effect estimate. The SNR cannot be observed directly, but there is a very simple relation between the SNR and the z statistic. Because the estimated effect is equal to the true effect plus an independent normal error term, the z statistic is equal to the SNR plus an independent, *standard* normal error term.[1] Thus, the distribution of the z statistic is the "convolution" of the distribution of the SNR and the standard normal distribution. The crux of our approach is that we can estimate the distribution of the absolute z statistics across the Cochrane Database and then derive the distribution of the absolute SNRs by "deconvolution," that is, by removing the standard normal component. This allows us to study a number of important statistical properties of the RCTs in the Cochrane Database.

We will focus on three properties of particular interest in this era of reproducibility concerns: the degree of overestimation; the probability that the estimated effect is in the same direction as the true effect; and the "predictive power" of a trial for obtaining $P \leq 0.05$ in the same direction for another study with the same underlying statistical parameters as the original trial, including the same underlying effect size and precision, and thus, the same power (an "exact replication" study in purely statistical terms). We present our results in a lookup table (see Table 3), which can help researchers interpret the two-sided P value of the primary efficacy result of a particular RCT in the context of the other RCTs from the Cochrane Database.

Previous efforts studying these properties have usually relied on Bayesian prior distributions chosen for either theoretical or computational reasons.[2,3] We instead base our inferences on empirical results from large collections of trials, the largest of these being the Cochrane Database.

Data, Methods, and Results

We used 23,551 RCTs from the Cochrane Database, which is arguably the most comprehensive collection of evidence on medical interventions. For simplicity, we represent a clinical trial as a triple (β, b, s), where β is the effect measure (true effect) targeted by the analysis and b is an estimate of β with SE s. Ignoring sampling variability in estimating s, the SNR is then $\text{SNR} = \frac{\beta}{s}$, and the z statistic is $z = \frac{b}{s}$. The effect β is usually a difference in means if the outcome of the trial is a continuous measurement, a log odds ratio if the outcome is binary, and a log hazard ratio if the outcome is time to an event. The precise choice does not matter for our purposes as long as b represents an estimator that is approximately normally distributed with mean β (i.e., is approximately unbiased for the targeted effect).

We collected the z statistics of the primary efficacy outcome of each of these trials.[4] Under the null hypothesis that the true effect is zero ($\beta = 0$) and there is no systematic error (bias), the z statistic has approximately a standard normal distribution. Thus, a z statistic of 1.96 or –1.96 corresponds to a two-sided P value of 0.05, and there is a one-to-one correspondence between the absolute z statistic and the two-sided P value.

van Zwet et al.[1] took the set of z statistics from the Cochrane Database and fitted a mixture of four zero-mean normal distributions to them. The z statistic is the sum of the SNR and standard normal noise, so we can obtain the distribution of the SNR by simply subtracting one from the variances of each of the mixture components. This "deconvolution" is a key step in the empirical Bayes approach.[5,6] The distributions of the z statistics and the SNRs are given in Table 1 and shown in Figure 1. If all the true effects were exactly zero, then the z statistics would approximately have the standard normal distribution. This is, of course, not the case, and so, the estimated distribution of the z statistics is much wider and has heavier tails than the standard normal. It might seem surprising that it is possible to estimate the joint distribution of the z statistics and the SNRs from observing only the z statistics. However, this is just a consequence of the fact that there is a very simple relation between the two distributions.

Table 1. Estimated Normal Mixture Distributions of the z Statistics and the Signal-to-Noise Ratios across 23,551 Trials of the Cochrane Database of Systematic Reviews[1]*

	Component			
Property	**1**	**2**	**3**	**4**
Proportion	0.32	0.31	0.30	0.07
Mean	0.00	0.00	0.00	0.00
SD z statistic	1.17	1.74	2.38	5.73
SD SNR	0.61	1.42	2.16	5.64

* The numbers 1–4 refer to individual components of the mixture distribution. SNR denotes signal-to-noise ratio.

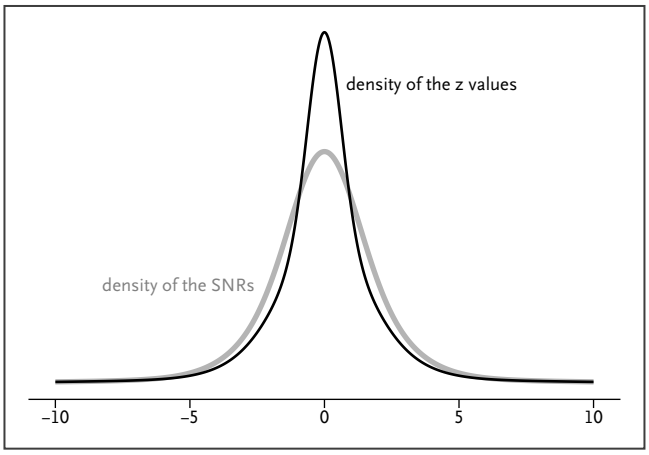

Figure 1. Estimated Distributions of the z Statistics (Black) and the Signal-to-Noise Ratios (Gray) across 23,551 Trials of the Cochrane Database of Systematic Reviews[1].

SNR denotes signal-to-noise ratio.

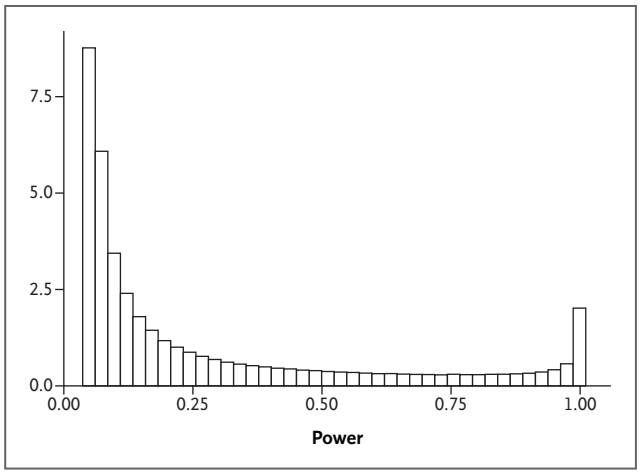

Figure 2. The Estimated Distribution of the Power against the True Effect among the Trials in the Cochrane Database of Systematic Reviews.

The results in the present article depend only on the distribution of the absolute values of the SNRs. Using a mixture of zero-mean normal distributions for the z statistics means that we are assuming a mixture of half-normal distributions for the absolute values of the SNRs. Any mixture of half-normal distributions has a decreasing density, so in practical terms, we are assuming that smaller values are more frequent than larger ones. We refer to our earlier work where we argue that this is a realistic assumption.[1]

We can use the distribution from Table 1 to compute several statistical quantities that should hold on average across the primary efficacy outcomes of trials similar to those in the Cochrane Database. We use a simple Monte Carlo scheme.

- Generate a sample, of size 10^6, from the estimated mixture distribution of the SNR.
- To each sampled SNR, add independent standard normal noise to obtain z.
- Compute the two-sided P value as $P = 2\Phi(-|z|)$, where Φ is the standard normal cumulative distribution function.
- To each sampled SNR, add another independent standard normal to obtain z_2, which represents the z statistic of a hypothetical "exact replication" study.

The result is a sample of size 10^6 of sets of four numbers (SNR, z, P, z_2). Now, the statistical power for the true effect is a transformation of the SNR:

$$power = \Phi(-1.96 - SNR) + 1 - \Phi(1.96 - SNR).$$

We can thus easily transform our sample of the SNRs into a sample of the powers, which we show in Figure 2. We estimate that the median power is only 13%, whereas just 12% of the trials reach 80% power.

By selecting and averaging, we can also compute the following quantities, conditional on P falling in some interval (these computations are provided in the Supplemental Appendix).

- The three quartiles of the exaggeration factor: The exaggeration factor is defined as the ratio of the estimated effect to the true effect, i.e., $|\frac{b}{\beta}|$. Note that this is equal to $|\frac{z}{SNR}|$. If researchers are more likely to report results with P values smaller than some (any) cutoff, this induces an upward bias in the effect estimate, sometimes called the "winner's curse." We quantify this bias in terms of the exaggeration factor.
- The coverage: The coverage is the probability that the 95% confidence interval covers the true effect. The true effect β falls in the range $b \pm 1.96 \cdot s$ if and only if the signal-to-noise ratio SNR falls in the range $z \pm 1.96$.
- The probability of the estimated effect b having the same sign (direction) as the true effect β: This is equivalent to the z and SNR having the same sign.
- The probability of a "successful replication": That is, the probability that an exact replication study will obtain a two-sided P value less than 0.05 with the

Table 2. Some Characteristics of the Cochrane Database Stratified by P Value*							
P-Value Stratum	Proportion	Q25	Q50	Q75	Coverage	Correct Sign	Replicate
(0.05, 1]	0.71	0.60	1.20	2.66	0.97	0.71	0.13
(0, 0.05]	0.29	1.02	1.29	1.92	0.89	0.98	0.60

* We report the proportion of P values in each stratum. Q25, Q50, and Q75 are the quartiles of the exaggeration. "Coverage" is the coverage of the usual 95% confidence interval. "Correct sign" refers to the probability that the sign (direction) of the estimated effect is correct. "Replicate" is the probability that an exact replication study will have a two-sided P value less than 0.05, and the direction of the original and replicated estimate are the same.

estimate in the same direction as the original study. This is the co-occurrence of the events $z \cdot z_2 > 0$ and $|z_2| \geq 1.96$.

Table 2 presents these quantities stratified on P>0.05 and P≤0.05.

Among other things, Table 2 shows that, conditionally on P≤0.05, the median exaggeration factor is 1.3 and the probability of a sign error is 2%. These quantities are closely related to the so-called type M (magnitude) and type S (sign) errors.[7,8]

Table 3 provides a more detailed picture by presenting the same quantities stratified on the P value falling in smaller intervals. Table 3 may be used to complement the usual interpretation of the P value of a particular trial of interest. We represent the results of Table 3 graphically in Figure 3.

Interpretation

The interpretation of P values is usually discussed without reference to a particular study design or research area. By referring to the Cochrane Database, we may state the implications of observing a particular nominal two-sided P value (or equivalently, an absolute z statistic) in a typical clinical trial without conditioning on the usually unreasonable null hypothesis of exactly zero effect. The use of a two-sided P value also means we do not have to worry about the sign of the effect, which has been a contentious issue in past studies of empirical P-value distributions.[9]

The compilation of z statistics from the Cochrane Database allows us to estimate properties of interest that have heretofore been thought possible only if we had a Bayesian prior on the true effect size. The true effect size can be viewed as a property of nature. A z statistic depends on both the true effect size and the trial design — including the sample size — and thus, it does not have a direct biologic meaning. However, the *distribution* of z statistics across the Cochrane Database does reflect the effect sizes that are being investigated and the designs of the clinical trials that are used in practice. RCTs are expensive, and investigators typically limit their planned size to what is needed to detect plausible or important anticipated effects. A standard sample size calculation with two-sided alpha = 5% and power = 90% sets the "effect of interest" as equal to 3.2 SEs. We can derive from the distribution of

Table 3. Some Characteristics of the Cochrane Database of Systematic Reviews Stratified by P Value in Finer Intervals*							
P-Value Stratum	Proportion	Q25	Q50	Q75	Coverage	Correct Sign	Replicate
(0.9, 1]	0.06	0.06	0.12	0.29	0.99	0.52	0.06
(0.8, 0.9]	0.06	0.23	0.41	0.89	0.99	0.55	0.07
(0.7, 0.8]	0.06	0.39	0.68	1.47	0.99	0.59	0.07
(0.6, 0.7]	0.06	0.54	0.95	2.05	0.99	0.63	0.08
(0.5, 0.6]	0.06	0.67	1.19	2.55	0.99	0.66	0.10
(0.1, 0.5]	0.33	0.95	1.67	3.59	0.97	0.79	0.15
(0.05, 0.1]	0.07	1.06	1.74	3.51	0.94	0.91	0.26
(0.01, 0.05]	0.10	1.05	1.56	2.81	0.90	0.95	0.37
(0.005, 0.01]	0.03	1.04	1.41	2.25	0.87	0.98	0.48
(0.001, 0.005]	0.04	1.04	1.35	1.95	0.87	0.99	0.58
(0, 0.001]	0.11	1.00	1.16	1.44	0.89	1.00	0.86

* The caption of Table 2 has details, and Figure 3 shows a graph. Q denotes quartile.

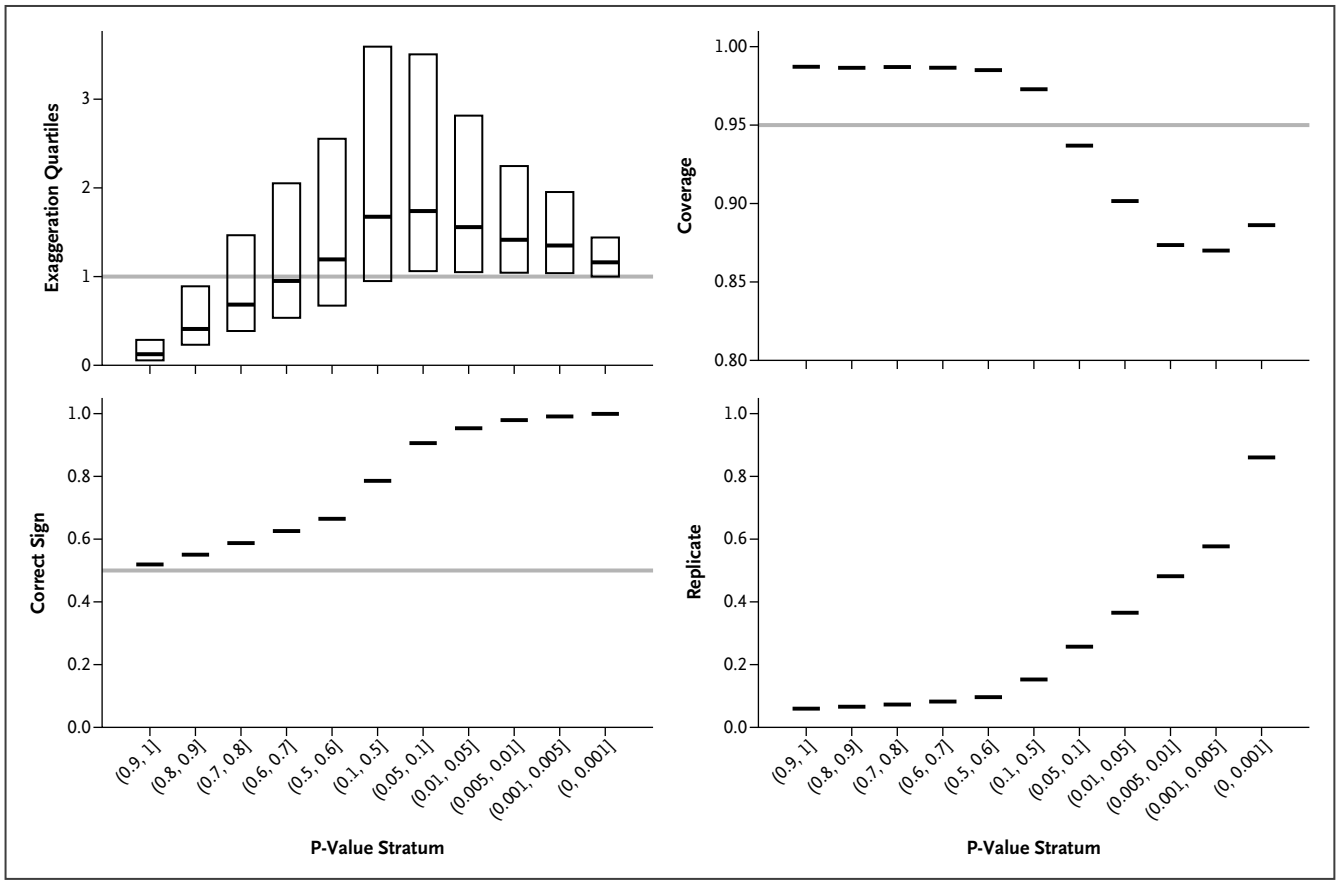

Figure 3. Graphical Representation of Table 3.

As a function of the P value range of a study assumed to have been drawn at random from the corpus, these graphs show the 25, 50, and 75% quantiles of the exaggeration factor (upper left panel); the actual coverage of the standard 95% confidence interval (upper right panel); the probability of the true effect having the same sign as the estimate (lower left panel); and the probability that a replicated study of the same size will get $P \leq 0.05$ ("statistically significant at the 5% level") and have the same sign as the estimate from the original study (lower right panel).

z statistics that the median power across the Cochrane Database for the true effect is in fact only 13%, corresponding to a far lower SNR.[1]

Tables 2 and 3 contain several other quantities that can be derived directly from the observed distribution of the z statistics. These results may be interpreted as follows. Suppose we choose a trial from the Cochrane Database at random and find that its two-sided P value for the primary efficacy outcome is between 0.01 and 0.05. Then, we estimate that there is a 75% probability that the magnitude of the effect is overestimated by at least 5%, a 50% probability that it is overestimated by at least 56%, and a 25% probability that it is overestimated by at least 181%. This phenomenon is like the infamous "winner's curse" in auctions.[10] Its connection to results of randomized trials has been pointed out by several authors.[7,11,12] Moreover, conditionally on $0.01<P<0.05$, the probability that the 95% confidence interval covers the true effect is only 90%. Also, the probability that an exact replication study will yield a P value less than 0.05 is only 37%. Fortunately, the probability that the direction (or "sign") of the estimated effect is correct is 95%.

Under our assumptions, Tables 2 and 3 tell us what it means, on average, to observe a particular P value in a study drawn at random from the population represented by the Cochrane Database. Here are a few striking features of Table 3.

- The overestimation of the effect is already severe in the stratum from 0.5 to 0.05. Thus, the "winner's

curse" is something of a misnomer in the sense that the overestimation is not tied to getting P≤0.05.
- The coverage of the 95% confidence interval is greater than 95% for large P values and less than 95% for small P values.
- The probability of the correct sign is already high in the stratum from 0.05 to 0.01.
- The probability of a replication study yielding P≤0.05 in the same direction is small, even in the stratum from 0.005 to 0.001. Thus, a replication with P>0.05 does not imply that the original finding was a fluke — at least not in the context of historical clinical trials — just as P≤0.05 in the original study does not imply that the initial conclusion was correct, especially when P is near 0.05.

Elsewhere, we have studied the same quantities as in Tables 2 and 3 and found similar results, despite using an entirely different method of computation.[1,13,14] In those articles, we conditioned on the exact z statistic instead of stratifying on intervals.

For the most common P values when "statistical significance" is declared (P values from 0.001 to 0.05), we expect high exaggeration factors (overestimating effect sizes by around 50% on average); mediocre coverage (nominal 95% intervals containing the true value approximately 90% of the time; that is, double the nominal error rate); and a probability of successful replication of P≤0.05 in the same direction, using the same sample size, of only a little over 40%. Given that applied researchers still commonly interpret results in terms of "statistical significance," we believe that this sort of empirical calibration can yield a helpful grounding in reality, either as a corrective to naive beliefs about 95% coverage and replicability or as a starting point for a more targeted Bayesian analysis (i.e., with a context-specific prior).

Discussion

The results of Tables 2 and 3 show, for example, that an initial P value between 0.001 and 0.005 implies only a 58% chance of getting P≤0.05 upon attempted replication. Some may find such a result surprising. We suspect that this surprise stems from the mistaken idea that a small P value confirms that the original trial had high power (80% or even 90%) and that it is, therefore, likely to be confirmed in a subsequent trial. Nonetheless, our results show that a P value between 0.001 and 0.005 indicates that the estimated effect is probably a substantial exaggeration of the actual effect, making the actual power much lower than it would seem.

Figure 1 shows that most trials have low power against the true effect. This should not be a surprise given that medical studies can be expensive and difficult to run, that outcomes are often unpredictable, and that there are clear incentives to be optimistic about effect sizes when designing a study. If, contrary to Figure 1, studies often did have 80% power, then we would routinely see P values ranging from 0.42 to 0.0000016, and we would see P values less than 0.0005 at least a quarter of the time.[15] As it is, a P value between 0.001 and 0.005 should not be taken as confirmation that a study was highly powered relative to the true effect that it was estimating.

The results in Tables 2 and 3 hold not only for a randomly selected RCT from the Cochrane Database but also for a randomly selected trial from the population of all trials that are "exchangeable" with those in the database (i.e., trials that could have been in the Cochrane Database). Although authors of systematic reviews are encouraged to use only studies that are sufficiently rigorous, there are no specific inclusion or exclusion criteria for the Cochrane Database.[16] The inclusion of a trial in the Cochrane Database largely depends on whether someone happens to be interested in a particular treatment or intervention, so the database is not a random sample from the population of all trials. In practical terms, "exchangeability with the Cochrane Database" means that a priori, we have no reason to expect the statistical properties of a particular trial of interest to differ from a randomly selected trial from the database. As such, we think that Tables 2 and 3 provide a useful frame of reference to interpret the result of an RCT.

The Cochrane Database represents common properties of trials, in particular the tendency to have low power against the true effect. This background information is important when interpreting the result of a particular trial. However, we will always have information about a particular trial that sets it apart from all other trials: the disease, treatment, population, trial design, sponsor, etc. We may choose to ignore that information or as an alternative, incorporate it into a prior distribution derived from other available studies on the topic and do a fully Bayesian analysis.

We used the z statistics as we found them in the Cochrane Database, which means in almost all cases, that the study treatment is compared with some control. However, we

do not know if a particular outcome or event is good or bad for the patient. So, we do not know which direction of the effect favors the study treatment, which means that we do not have access to the one-sided P values. We thus used the two-sided P values or equivalently, the absolute values of the z statistics. As a result, Tables 2 and 3 do not depend on the direction (sign) of the effect or whether the study treatment tends to be superior to the control condition.

Although our quantitative results cannot be applied directly to other fields, we think they are qualitatively relevant for fields in which the SNR tends to be low. For example, in those fields, P values between 0.05 and 0.001 will be associated with exaggerated effect estimates and low replication of estimate size and "statistical significance" — manifestations of the familiar phenomenon of regression to the mean.

We have assumed that the sample size in the replication study is the same as that in the original study. Similar calculations show that to have a reasonable chance of replicating (say $P \leq 0.05$), follow-up clinical trials must be many times larger than the original study.[13] A large number of scientific fields suffer from studies with inadequate sample sizes, and use $P \leq 0.05$ as an arbiter of claims. It is thus no surprise that "replication failure" is commonly reported. Our results thus reinforce the many objections to equating $P \leq 0.05$ or "statistical significance" with effect discovery or replication or using them as publication criteria.[17-24]

Disclosures

Supported by the U.S. Office of Naval Research.

Author disclosures and other supplementary materials are available at evidence.nejm.org.

A data sharing statement provided by the authors is available with the full text of this article at NEJM.org.

We thank Eric-Jan Wagenmakers for his comments on the manuscript.

Author Affiliations

[1] Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands

[2] Department of Statistics, Columbia University, New York

[3] Department of Political Science, Columbia University, New York

[4] Department of Epidemiology, University of California, Los Angeles, Los Angeles

[5] Department of Statistics, University of California, Los Angeles, Los Angeles

[6] Graduate School of Business, Department of Economics, Stanford University, Stanford, CA

[7] Swisstransplant, Bern, Switzerland

[8] Department of Epidemiology and Population Health, Stanford University, Stanford, CA

References

1. van Zwet E, Schwab S, Senn S. The statistical properties of RCTs and a proposal for shrinkage. Stat Med 2021;40:6107-6117. DOI: 10.1002/sim.9173.
2. Goodman SN. A comment on replication, p-values and evidence. Stat Med 1992;11:875-879. DOI: 10.1002/sim.4780110705.
3. Spiegelhalter DJ, Freedman LS, Parmar MKB. Bayesian approaches to randomized trials. J R Stat Soc Ser A Stat Soc 1994;157:357-387. DOI: 10.2307/2983527.
4. Schwab S. Re-estimating 400,000 treatment effects from intervention studies in the Cochrane Database of Systematic Reviews. Data set. Open Science Framework. December 11, 2020 (https://osf.io/xjv9g/).
5. Efron B. Empirical Bayes deconvolution estimates. Biometrika 2016;103:1-20. DOI: 10.1093/biomet/asv068.
6. Stephens M. False discovery rates: a new deal. Biostatistics 2017; 18:275-294.
7. Gelman A, Carlin J. Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. Perspect Psychol Sci 2014;9: 641-651. DOI: 10.1177/1745691614551642.
8. Gelman A, Tuerlinckx F. Type S error rates for classical and Bayesian single and multiple comparison procedures. Comput Stat 2000; 15:373-390. DOI: 10.1007/s001800000040.
9. Sterne J. Does the selective inversion approach demonstrate bias in the results of studies using routinely collected data? BMJ 2018;362: k3259. DOI: 10.1136/bmj.k3259.
10. Bazerman MH, Samuelson WF. I won the auction but don't want the prize. J Conflict Resolut 1983;27:618-634. DOI: 10.1177/0022002783027004003.
11. Ioannidis JPA. Why most discovered true associations are inflated. Epidemiology 2008;19:640-648. DOI: 10.1097/EDE.0b013e31818131e7.
12. van Zwet EW, Cator EA. The significance filter, the winner's curse and the need to shrink. Stat Neerl 2021;75:437-452. DOI: 10.1111/stan.12241.
13. van Zwet EW, Goodman SN. How large should the next study be? Predictive power and sample size requirements for replication studies. Stat Med 2022;41:3090-3101. DOI: 10.1002/sim.9406.
14. van Zwet EW, Schwab S, Greenland S. Addressing exaggeration of effects from single RCTs. Significance 2021;18:16-21. DOI: 10.1111/1740-9713.01587.
15. Gelman A. The 80% power lie. Statistical Modeling, Causal Inference, and Social Science. December 4, 2017 (https://statmodeling.stat.columbia.edu/2017/12/04/80-power-lie/).
16. McKenzie JE, Brennan SE, Ryan RE, Thomson HJ, Johnston RV, Thomas J. Defining the criteria for including studies and how they will be grouped for the synthesis. In: Higgins JPT, et al., eds. Cochrane handbook for systematic reviews of interventions. Hoboken, NJ: Wiley, 2019:33-65.

17. Amrhein V, Trafimow D, Greenland S. Inferential statistics as descriptive statistics: there is no replication crisis if we don't expect replication. Am Stat 2019;73(Suppl 1):262-270.

18. Greenland S, Mansournia MA, Joffe M. To curb research misreporting, replace significance and confidence by compatibility: a Preventive Medicine Golden Jubilee article. Prev Med 2022;164:107127. DOI: 10.1016/j.ypmed.2022.107127.

19. Imbens GW. Statistical significance, p-values, and the reporting of uncertainty. J Econ Perspect 2021;35:157-174. DOI: 10.1257/jep.35.3.157.

20. Lakens D, Adolfi FG, Albers CJ, et al. Justify your alpha. Nat Hum Behav 2018;2:168-171. DOI: 10.1038/s41562-018-0311-x.

21. McShane BB, Gal D, Gelman A, Robert C, Tackett JL. Abandon statistical significance. Am Stat 2019;73(Suppl 1):235-245.

22. Rafi Z, Greenland S. Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. BMC Med Res Methodol 2020;20:244. DOI: 10.1186/s12874-020-01105-9.

23. Wasserstein RL, Lazar NA. The ASA statement on p-values: context, process, and purpose. Am Stat 2016;70:129-133. DOI: 10.1080/00031305.2016.1154108.

24. Wasserstein RL, Schirm A, Lazar NA. Moving to a world beyond "p < 0.05." Am Stat 2019;73(Suppl 1):1-19. DOI: 10.1080/00031305.2019.1583913.