A new look at p-values for randomized clinical trials[1]

Erik van Zwet (PhD)[2]*, Andrew Gelman (PhD)[3], Sander Greenland (MD (hon), DrPH)[4],

Guido Imbens (PhD)[5], Simon Schwab (PhD)[6], Steven N. Goodman (MD, PhD)[7]

2 October 2023

Abstract We have examined the primary efficacy results of 23,551 randomized clinical trials (RCTs) from the Cochrane Database of Systematic Reviews (CDSR). We estimate that the great majority of trials have much lower statistical power for actual effects than the 80% or 90% for the effect sizes stated in proposals. 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 re-interpret the *p*-value in terms of a reference population of studies that are, or could have been, in the CDSR. 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. Such an interpretation provides additional insight about the effect under study and can guard medical researchers against naive interpretations of *p*-value and over-optimistic effect sizes. Because many research fields suffer from low power, our results are also relevant outside the medical domain.

1 Background

How should applied researchers interpret the *p*-value for the null hypothesis of no effect from a randomized clinical trial? This *p*-value is commonly defined as the probability,

---

[1] To appear in *NEJM Evidence.*

[2] Department of Biomedical Data Sciences, Leiden University Medical Center, The Netherlands. * Corresponding author: E.W.van_Zwet@lumc.nl

[3] Department of Statistics and Department of Political Science, Columbia University, USA

[4] Department of Epidemiology and Department of Statistics, University of California, Los Angeles, USA

[5] Graduate School of Business, Department of Economics, Stanford University, USA

[6] Swisstransplant, Bern, Switzerland

[7] Department of Epidemiology and Population Health, Stanford University, USA

under the null hypothesis and an assumed data-generating model, that an appropriate test statistic would be as or more extreme than what was observed. Here we will consider the absolute *z*-statistic. We are interested in understanding the resulting two-sided *p*-value without changing its calculation or assuming that the null hypothesis is correct. Instead, we wish to reinterpret it in light of background information about studies with similar statistical properties. The Cochrane Database of Systematic Reviews (CDSR) contains the results of more than 20,000 randomized clinical trials (RCTs) in biomedicine. We have collected the absolute *z*-statistics of the primary efficacy outcome of these RCTs.

Recall that the *z*-statistic is the estimated effect divided by the standard error of the estimate. We will also consider the signal-to-noise ratio (SNR) which is the *true* effect divided by the standard error of the effect estimate. The SNR cannot be observed directly, but there is a very simple relation between the SNR and the *z*-statistic; the SNR is equal to the *z*-statistic plus independent, standard normal "noise". The crux of our approach is that we can estimate the distribution of the absolute *z*-statistics across the CDSR, and then derive the distribution of the absolute SNRs. This allows us to study a number of important statistical properties of the RCTs in the CDSR.

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 look-up table (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 CDSR.

Previous efforts studying these properties have usually relied on Bayesian prior distributions chosen for either theoretical or computational reasons[8 17]. We instead base our inferences on empirical results from large collections of trials, the largest of these being the Cochrane Database of Systematic Reviews.

2 Data, methods, and results

We used 23,551 randomized clinical trials from the Cochrane Database of Systematic Reviews (CDSR), 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 standard error s. Ignoring sampling variability in estimating s, the signal-to-noise ratio is then 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[16]. Under the null hypothesis that the true effect is zero (β=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, Schwab and Senn[21] took the set of *z*-statistics from the CDSR and fitted a mixture of 4 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 1 from the variances of each of the mixture components. This "deconvolution" is a key step in the empirical-Bayes approach[3]. The distributions of the *z*-statistics and the SNRs is given in Table 1 and shown in Figure 1. 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.

The results in the present paper 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[21].

Table 1: Estimated normal mixture distributions of the z-statistics and the signal-to-noise ratios (SNRs) across 23,551 trials of the CDSR[21].

	1	2	3	4
proportion	0.32	0.31	0.30	0.07
mean	0.00	0.00	0.00	0.00
std. dev. *z*-statistic	1.17	1.74	2.38	5.73
std. dev. SNR	0.61	1.42	2.16	5.64

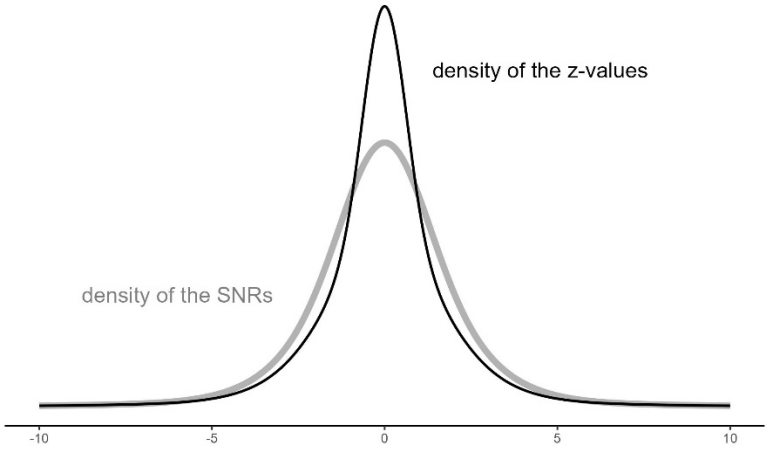

Figure 1: Estimated distributions of the z-statistics (black) and the signal-to-noise ratios (grey) across 23,551 trials of the CDSR[21].

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 CDSR. We use a simple Monte Carlo scheme:

---

1. Generate a sample of size 10^6 from the estimated mixture distribution of the SNR.
2. To each sampled SNR, add independent standard normal noise to obtain z.
3. Compute the two-sided p-value as $p = 2\Phi(-|z|)$, where Φ is the standard normal cumulative distribution function.
4. 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 4 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%, while just 12% of the trials reach 80% power.

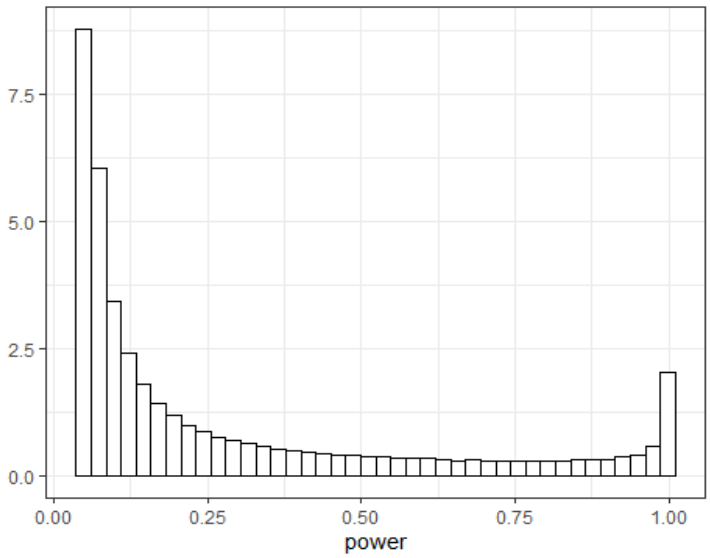

Figure 2: The estimated distribution of the power against the true effect among the trials in the CDSR.

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

1. The three quartiles of the *exaggeration factor*, $\left|\frac{b}{\beta}\right| = \left|\frac{z}{SNR}\right|$. 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."

2. The *coverage*, 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 SNR falls in the range $z \pm 1.96$.

3. The *probability of the estimate having the correct sign*, which is the event $b\beta > 0$ or equivalently, $z \cdot SNR > 0$.

4. The probability that an exact replication study will obtain *a two-sided p-value less than 0.05* with the estimate in the same direction as the original study, that is, the co-occurrence of the events $zz_2 > 0$ and $|z_2| \geq 1.96$.

Table 2 presents these quantities stratified on $p > 0.05$ and $p \leq 0.05$.

Table 2: Some characteristics of the CDSR, stratified by p-value. 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% CI. "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.

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

Among other things, Table 2 shows that, conditionally on $p \leq 0.05$, 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[6 5].

Table 3 presents the same quantities, but stratified on the p-value falling in smaller intervals. We also represent the results of Table 3 graphically in Figure 3.

Table 3: Some characteristics of the CDSR, stratified by p-value in finer intervals. See the caption of Table 2 for details, and see Figure 2 for a graph.

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

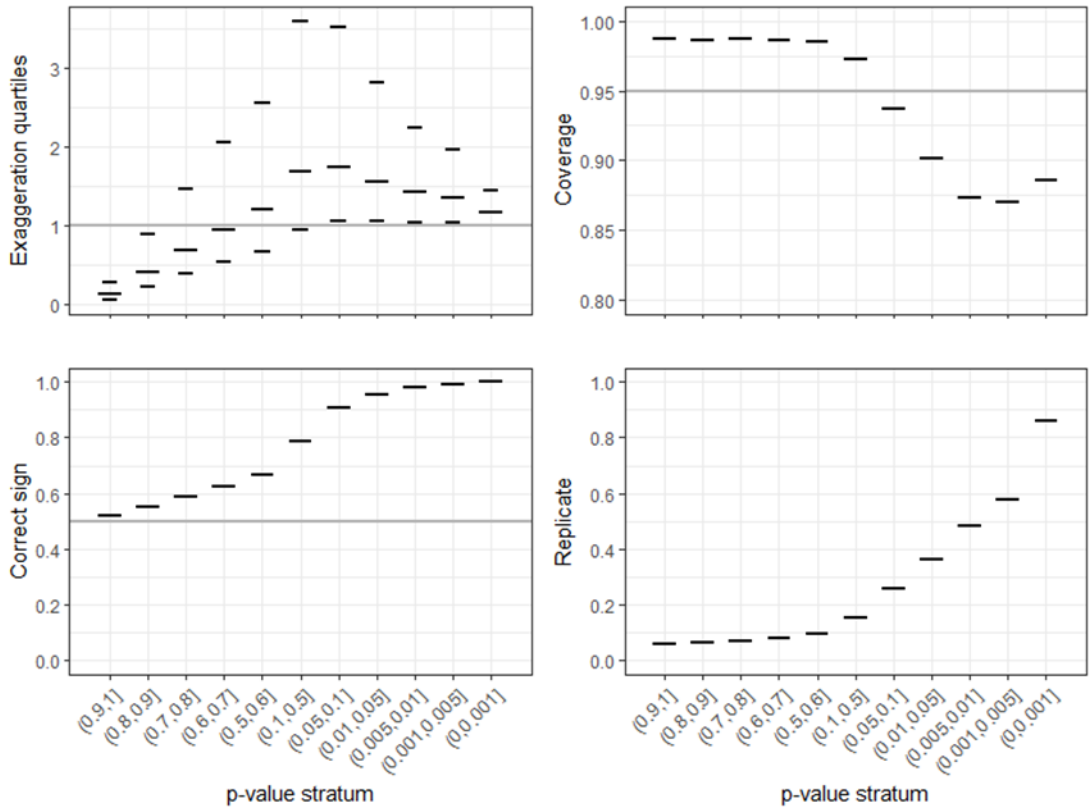

Figure 3: Graphical representation of Table 3. As a function of the p-value range of a study assumed to have been drawn at random from the corpus, these graphs show: (top left) the 25%, 50%, and 75% quantiles of the exaggeration factor; (top right) the actual coverage of the standard 95% confidence interval; (bottom left) the probability of the true effect having the same sign as the estimate; and (bottom right) 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.

3 Interpretation

The interpretation of *p*-values is usually discussed without reference to a particular study design or research area. By referring to the CDSR, 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[18].

The compilation of *z*-statistics from the CDSR 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 does not have a direct biological meaning. However, the *distribution* of *z*-statistics across the CDSR 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 *z*-statistics that the median power across the CDSR for the true effect is in fact only 13%, corresponding to a far lower SNR[21].

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 CDSR at random, and find 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[2]. Its connection to results of randomized trials has been pointed out by several authors[11 5 19]. 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 CDSR. 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 $p \leq 0.05$.
- 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 $p \leq 0.05$ 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 $p \leq 0.05$ 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[21 22 20]. In those papers we conditioned on the exact *z*-statistic instead of stratifying on intervals.

For the most common *p*-values when "statistically 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 $p \leq 0.05$ 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.

4 Discussion

The results of Tables 2 and 3 show for example that an initial *p*-value between 0.001 and 0.005 implies a only a 58% chance of getting $p \leq 0.05$ 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 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, outcomes are often unpredictable, and there are clear incentives to be optimistic about effect sizes when designing a study. If, contrary to Figure 1, studies often did have80% 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[7]. 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 it was estimating.

The results in Tables 2 and 3 hold not only for a randomly selected RCT from the CDSR, but also for a randomly selected trial from the population of all trials that are "exchangeable" with those in the CDSR, i.e., trials that *could have been* in the CDSR. 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 CDSR[13]. The inclusion of a trial in the CDSR largely depends on whether someone happens to be interested in a particular treatment or intervention, so the CDSR is not a random sample from the population of all trials. In practical terms, "exchangeability with the CDSR" 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 CDSR. As such, we think that Tables 2 and 3 provide a useful frame of reference to interpret the result of a randomized clinical trial.

The CDSR 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 CDSR, which means in almost all cases that the study treatment is compared to 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.

While our quantitative results cannot be applied directly to other fields, we think they are qualitatively relevant for fields in which the signal-to-noise ratio 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[20]. 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[23 12 1 14 24 15 10 9].

5 Availability of data and code

The data on which our results are based are available through the Open Science Framework[16]. R code to reproduce our results is provided in the Appendix.

6 Author contributions

SS collected the CDSR data[16]. EvZ did the calculations for Tables 2 and 3, and wrote the first draft. All authors verified the calculations, discussed the results and their implications, and commented on the manuscript at all stages.

7 Acknowledgements

We thank Eric-Jan Wagenmakers for his comments on the manuscript. Andrew Gelman thanks the U.S. Office of Naval Research for partial support of this work.

8 References

1. Amrhein, V., Trafimow, D. and Greenland, S., 2019. Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *American Statistician*, *73* (sup1), 262-270.
2. Bazerman M.H., Samuelson W.F. 1983 I won the auction but don't want the prize. *Journal of Conflict Resolution* 27 618-634.
3. Efron, B. 2016. Empirical Bayes deconvolution estimates. *Biometrika*, *103*(1), pp.1-20.
4. Gelman, A. 2018. The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. *Personality and Social Psychology Bulletin* 44 (1), 16-23.
5. Gelman, A, and Carlin, J. 2014. Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science* 9 (6): 641–651.
6. Gelman, A, and Tuerlinckx, F. 2000. Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics* 15 (3): 373–390.
7. Gelman, A. 2017. The 80% power lie. Statistical Modeling, Causal Inference, and Social Science, 4 Dec. https://statmodeling.stat.columbia.edu/2017/12/04/80-power-lie/
8. Goodman, S.N. 1992. A comment on replication, p-values and evidence. *Statistics in Medicine* 11 (7): 875–79.
9. Greenland, S., Mansournia, M., and Joffe, M. 2022. To curb research misreporting, replace significance and confidence by compatibility. *Preventive Medicine*, 164, https://www.sciencedirect.com/science/article/pii/S0091743522001761.
10. Imbens, G. W. 2021 Statistical significance, p-values, and the reporting of uncertainty. *Journal of Economic Perspectives* 35 (3): 157-74.
11. Ioannidis, J.P. A. 2008. Why most discovered true associations are inflated. *Epidemiology* 19 (5): 640–648.
12. Lakens, D., Adolfi, F.G., Albers, C.J., Anvari, F., Apps, M.A., Argamon, S.E., Baguley, T., Becker, R.B., Benning, S.D., Bradford, D.E. and Buchanan, E.M., 2018. Justify your alpha. *Nature Human Behaviour*, *2* (3), 168-171.
13. McKenzie, J.E., Brennan, S.E., Ryan, R.E., Thomson, H.J., Johnston, R.V. and Thomas, J., 2019. Defining the criteria for including studies and how they will be grouped for the synthesis. *Cochrane Handbook for Systematic Reviews of Interventions*, 33-65.
14. McShane, B.B., Gal, D., Gelman, A., Robert, C. and Tackett, J.L., 2019. Abandon statistical significance. *American Statistician*, *73*(sup1), 235-245.
15. Rafi, Z., and Greenland, S., 2020. Semantic and cognitive tools to aid statistical science: Replace confidence and significance by compatibility and surprise. *BMC Medical Research Methodology*, 20, 244. doi: 10.1186/s12874-020-01105-9,
16. Schwab, S. 2020. Re-estimating 400,000 treatment effects from intervention studies in the Cochrane Database of Systematic Reviews [data set]. *Open Science Framework*. https://doi.org/10.17605/OSF.IO/XJV9G
17. Spiegelhalter, D.J., Freedman, L.S. and Parmar, M.K.B. 1994. Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society: Series A* 157 (3): 357–387.
18. Sterne, J. 2018. 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

19. van Zwet, E.W. and Cator, E.A. 2021. The significance filter, the winner's curse and the need to shrink. *Statistica Neerlandica* 75: 437–452.
20. van Zwet, E.W., and Goodman, S.N. 2022. How large should the next study be? Predictive power and sample size requirements for replication studies. *Statistics in Medicine* 41: 3090–3101.
21. van Zwet, E.W., Schwab, S. and Senn S.J. 2021. The statistical properties of RCTs and a proposal for shrinkage. *Statistics in Medicine* 40 (27): 6107–6017.
22. van Zwet, E.W., Schwab, S. and Greenland, S. 2021. Addressing exaggeration of effects from single RCTs. *Significance* 18 (6): 16–21.
23. Wasserstein R.L., and Lazar N.A. 2016. The ASA statement on p-values: context, process, and purpose. *The American Statistician*. 70(2):129–133.
24. Wasserstein R.L., Schirm, A., and Lazar N.A. 2019. Moving to a world beyond "p < 0.05". *The American Statistician* 73:1–19.