Comment on "The Role of p-Values in Judging the Strength of Evidence and Realistic Replication Expectations"

Held, Leonhard ; Pawel, Samuel

DOI: https://doi.org/10.1080/19466315.2020.1828161

Comment on "The Role of p-values in Judging the Strength of Evidence and Realistic Replication Expectations"

Leonhard Held*, Samuel Pawel†

Department of Biostatistics, EBPI, and

Center for Reproducible Science,

University of Zurich, Hirschengraben 84, 8001 Zurich,

Switzerland

21st September 2020

Eric Gibson is to be congratulated for a thoughtful review of the role of *p*-values in the assessment of the strength of evidence of research findings in pharmaceutical drug development. This perspective highlights important issues in a highly regulated environment, where study planning, protocol writing and pre-registration have been the standard for many years. It gives important insights to other disciplines, where similar standards are currently being implemented (Chambers, 2019a,b).

*Email: leonhard.held@uzh.ch
†Email: samuel.pawel@uzh.ch

Gibson (2020) mentions the "reproducibility probability" as a way to quantify the strength of evidence measured by p-values, his results are reproduced in Table 1. What is the probability that an identically designed second study (with the same sample size) will be significant, given the result from the first study? This is perhaps better referred to as the replication probability, following the distinction between reproducibility and replicability as suggested by National Academies of Sciences, Engineering, and Medicine (2019), see also Goodman et al. (2016). We would like to comment on how this quantity can be further adjusted to give a more realistic estimate of how likely it is that a replication will again be significant.

	p-value					
Calibration method	0.10	0.05	0.01	0.001	0.0001	0.00001
BFB	1.6	2.5	8	53	399	3195
$-\log_{10}(p\text{-value})$	1	1.3	2	3	4	5
$\widehat{\text{RP}}$ unadjusted	0.38	0.50	0.73	0.91	0.97	0.99
$\widehat{\text{RP}}$ adjusted for uncertainty	0.41	0.50	0.67	0.83	0.91	0.96
+ regression to the mean	0.23	0.35	0.57	0.77	0.89	0.94
+ heterogeneity	0.19	0.30	0.49	0.68	0.80	0.88

Table 1: Comparison of Bayes factor bound, $-\log_{10}(p\text{-value})$, and replication probability calibration of p-values. Replication probabilities are either unadjusted or adjusted for uncertainty of original effect estimate, regression to the mean, between-study heterogeneity of the effect.

In a seminal contribution, Goodman (1992) showed that the replication probability solely depends on the original p-value and that it is only 50% for borderline significant studies ($p \approx 0.05$). In the best-case scenario the observed effect estimate is the true effect, which is also assumed for the computation of the probabilities shown in Gibson (2020). In practice, however, there is still uncertainty about the effect, and we may want to adjust the replication probability by averaging it over the distribution of the effect estimate, as also considered in Goodman (1992). Incorporation of the uncertainty

about the effect leads also to larger uncertainty about whether the replication will be significant. Specifically, the replication probability further decreases for significant p-values, while it increases for non-significant ones, see Table 1.

Although taking into account the uncertainty of the estimate may improve the calibration of the replication probability, taking a study result at face value might still not be good idea since effect estimates are often exaggerated due to publication bias and regression to the mean (as Gibson also mentions in Section 2.4). This problem is particularly severe for low powered studies, where significant findings are likely to be false positive. Copas (1997) suggested a method to address this issue, shrinking the effect estimate towards zero. In short, the amount of shrinkage is $1/z^2$ where z is the standard z-statistic associated with p. The corresponding replication probabilities then decrease further, as shown in Table 1. For example, for $p = 0.05$, the amount of shrinkage is $1/1.96^2 = 0.26$ and the replication probability decreases from 0.50 (without shrinkage) to 0.35, so only one in three borderline significant studies will achieve significance in a replication study.

Finally, the assumption that the true effect is exactly the same in original and replication is often inappropriate. While in theory we can think about an identically designed replication, in practice there will always be deviations from the original study, *e. g.* the study population may differ in some characteristics. It is more reasonable to assume between-study heterogeneity of effects, as is also often done in drug development (see e.g. Neuenschwander et al., 2018). Table 1 also shows replication probabilities that were adjusted for between-study heterogeneity on top of the other adjustments. The heterogeneity parameter was chosen based on the upper limit of "negligible" heterogeneity ($I^2 = 40\%$) according to the Cochrane guidelines for systematic reviews (Deeks et al., 2019). We can see that the replication probabilities decrease further. For example,

for $p = 0.0001$ it decreases from unadjusted 0.97 to adjusted 0.80, the convention for a reasonable power in many fields.

Gibson (2020) argues that for p-values below 0.001 replication probabilities do not calibrate as well as $-\log_{10}(p)$ or Bayes factor bounds. However, this is not the case anymore after adjusting for uncertainty, regression to the mean, and heterogeneity. In an empirical investigation we attempted to predict replication effect estimates using data from four different replication projects (Pawel and Held, 2020). With the adjustments mentioned above, we were able to substantially improve predictive performance upon previous attempts. In fact, taking into account both regression to the mean and heterogeneity led to well calibrated predictions in two of the four datasets.

	Example		
	1	2	3
original p-value	0.049	0.051	0.049
replication p-value	0.049	0.001	0.001
relative sample size	1	1	8
harmonic mean p-value	0.003	0.0004	0.0004
BFB	6	129	133
relative effect size	1	1.69	0.59
one-sided sceptical p-value	0.082	0.047	0.10

Table 2: Three examples with different original and replication studies. Harmonic mean p-value, Bayes factor bound, relative effect size, and one-sided sceptical p-value are shown for each.

The case studies described in Gibson (2020, Section 3) are clear failures with replication effect estimates even in the wrong direction. However, quite often the effect estimates go in the same direction, but it is not clear whether the observed result can be regarded as replication success. The "two-trials rule" (Senn, 2007) requires both studies to be significant, but can produce anomalies which do not reflect the available

evidence. For example, two trials both with (two-sided) $p = 0.049$ (example 1 in Table 2) will then lead to drug approval but carry less evidence for a treatment effect than one trial with $p = 0.051$, say, and the other one with $p = 0.001$. The latter, however, would not pass the two-trials rule, although its Bayes factor bound is much larger than for example 1.

An alternative to the two-trials rule with better properties, the harmonic mean χ^2-test, was recently proposed (Held, 2020b). This method produces a meta-analytic p-value p_H and can be extended to more than two studies, but differs substantially from more standard meta-analytic approaches, as it requires all individual studies to be convincing to a certain degree. Using the p-value threshold $2 \times (1/40) \times (1/40) = 0.00125$ suggested by Gibson (2020, Section 2.5), the first example would not lead to approval ($p_H = 0.003$), whereas the second would ($p_H = 0.0004$).

Low powered original studies (with small sample size n_o) are not the only problem. Replication studies with relatively large sample sizes n_r can also be misleading, as they may lead to significance even if the replication effect estimate $\hat{\theta}_r$ is much smaller than the original one $\hat{\theta}_o$. Let $c = n_r/n_o$ and $d = \hat{\theta}_r/\hat{\theta}_o$ denote the relative sample size and the relative effect size of replication to original study, respectively. Assume the two studies have the same primary endpoint. Under the usual normality assumption for the effect estimate combined with the standard $\sqrt{n}$ law for the standard error we obtain the relative effect size

$$d = c^{-1/2} \frac{z_o}{z_r}, \tag{1}$$

here z_o and z_r are the z-statistics of the original and replication study, respectively.

Consider now the third example with $p_o = 0.049$ and $p_r = 0.001$ and assume the sample size of the replication study has been eight times as large compared to the original study, so $c = 8$. This sounds exaggerated, but is roughly the sample size

needed to achieve 80% power to detect the effect observed in the first study accounting for the necessary shrinkage implied by regression to the mean (Pawel and Held, 2020, Appendix S2). Then $d = 0.59$, so there is substantial shrinkage of the replication effect estimate. Common sense suggests that this result should be treated with more suspicion than example 2, say, where the effect estimate even increases ($d = 1.69$), but the p-values are virtually the same. These considerations suggest that the two-trials rule is a poor indicator of replication success (Simonsohn, 2015).

A reverse-Bayes approach for the assessment of replication success was proposed in Held (2020a), which penalizes shrinkage of the replication estimate compared to the original estimate, while ensuring that both effect estimates are statistically significant to some extent. The method takes into account not only the p-values from the two studies, but also the relative sample size c and therefore the relative effect size d via (1). A quantitative measure of the degree of replication success is proposed, the sceptical p-value p_S. It quantifies the degree of conflict between the replication experiment and a sceptical prior that would make the original experiment no longer significant. Table 2 gives the one-sided version of the sceptical p-value. While the interpretation of the actual value of p_S requires a recalibration (Held et al., 2020), it can be easily used to compare the degree of replication success of different study pairs (the smaller, the better). Interestingly, the first example with $p_o = p_r = 0.049$ and $c = 1$ (and hence $d = 1$) is then more trustworthy (with $p_S = 0.082$) than the seemingly more convincing third example with $p_o = 0.049$, $p_r = 0.001$ and $c = 8$ (with $p_S = 0.10$). This shows how p_S takes into account sample and effect sizes when assessing replication success.

We want to add a few final comments on the interpretation of the 5% level for statistical significance. It is now well accepted that $p < 0.05$ is a too lax criterion for a scientific discovery. Indeed, even in the absence of multiplicity issues, selective re-

porting, etc, $p \approx 0.05$ gives only weak evidence against the null as quantified by the corresponding Bayes factor bound. This is why Benjamin et al. (2018) have suggested the more stringent 0.005 significance threshold for claims of new discoveries. Studies with $0.005 < p < 0.05$ are called "suggestive", calling for confirmation through replication. It is worth noting that it was Fisher who said that a significant observation (at the 0.05 threshold) indicates that it is merely worth to repeat the experiment (Goodman, 2016). This view underlines the central role of replication and has to be contrasted to the misleading, but still prevailing view that a single significant result gives "statistical proof" of a scientific claim.

Acknowledgments Support by the Swiss National Science Foundation (Project # 189295) is gratefully acknowledged.

References

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., et al. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2:6–10.

Chambers, C. (2019a). The registered reports revolution – lessons in cultural reform. *Significance*, 16(4):23–27.

Chambers, C. (2019b). What's next for registered reports. *Nature*, 573(7773):187–189.

Copas, J. B. (1997). Using regression models for prediction: shrinkage and regression to the mean. *Stat. Methods Med. Res.*, 6(2):167–183.

Deeks, J. J., Higgins, J. P., and Altman, D. G. (2019). Analysing data and undertaking meta-analyses. In *Cochrane Handbook for Systematic Reviews of Interventions*, chapter 10, pages 241–284. John Wiley & Sons, Ltd.

Gibson, E. W. (2020). The role of p-values in judging the strength of evidence and realistic replication expectations. *Statistics in Biopharmaceutical Research*, 0(0):1–13.

Goodman, S. N. (1992). A comment on replication, p-values and evidence. *Statistics in medicine*, 11(7):875–879.

Goodman, S. N. (2016). Aligning statistical and scientific reasoning. *Science*, 352(6290):1180–1181.

Goodman, S. N., Fanelli, D., and Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, 8(341):341ps12–341ps12.

Held, L. (2020a). A new standard for the analysis and design of replication studies (with discussion). *Journal of the Royal Statistical Society, Series A*, 183:431–469.

Held, L. (2020b). The harmonic mean χ^2 test to substantiate scientific findings. *Journal of the Royal Statistical Society, Series C*, 69:697–708.

Held, L., Micheloud, C., and Pawel, S. (2020). The assessment of replication success based on relative effect size. Technical report. `http://arxiv.org/abs/2009.07782`.

National Academies of Sciences, Engineering, and Medicine (2019). *Reproducibility and Replicability in Science*. The National Academies Press. `https://doi.org/10.17226/25303`.

Neuenschwander, B., Roychoudhury, S., and Branson, M. (2018). Predictive evidence threshold scaling: does the evidence meet a confirmatory standard? *Statistics in Biopharmaceutical Research*.

Pawel, S. and Held, L. (2020). Probabilistic forecasting of replication studies. *PLOS ONE*, 15(4):e0231416. `https://doi.org/10.1371/journal.pone.0231416`.

Senn, S. (2007). *Statistical Issues in Drug Development*. John Wiley & Sons, Chichester, U.K., second edition.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, 26(5):559–569.