

Meta-Research: Replication of “null results” – Absence of evidence or evidence of absence?

Samuel Pawel^{1*†}, Rachel Heyard^{1†}, Charlotte Micheloud¹, Leonhard Held¹

*For correspondence:
samuel.pawel@uzh.ch (SP)

†Contributed equally

¹Epidemiology, Biostatistics and Prevention Institute, Center for Reproducible Science, University of Zurich, Switzerland

Abstract In several large-scale replication projects, statistically non-significant results in both the original and the replication study have been interpreted as a “replication success”. Here we discuss the logical problems with this approach. Non-significance in both studies does not ensure that the studies provide evidence for the absence of an effect and “replication success” can virtually always be achieved if the sample sizes of the studies are small enough. In addition, the relevant error rates are not controlled. We show how methods, such as equivalence testing and Bayes factors, can be used to adequately quantify the evidence for the absence of an effect and how they can be applied in the replication setting. Using data from the Reproducibility Project: Cancer Biology we illustrate that many original and replication studies with “null results” are in fact inconclusive. We conclude that it is important to also replicate studies with statistically non-significant results, but that they should be designed, analyzed, and interpreted appropriately.

Introduction

Absence of evidence is not evidence of absence – the title of the 1995 paper by Douglas Altman and Martin Bland has since become a mantra in the statistical and medical literature (Altman and Bland, 1995). Yet, the misconception that a statistically non-significant result indicates evidence for the absence of an effect is unfortunately still widespread (Makin and de Xivry, 2019). Such a “null result” – typically characterized by a p -value of $p > 0.05$ for the null hypothesis of an absent effect – may also occur if an effect is actually present. For example, if the sample size of a study is chosen to detect an assumed effect with a power of 80%, null results will incorrectly occur 20% of the time when the assumed effect is actually present. Conversely, if the power of the study is lower, null results will occur more often. In general, the lower the power of a study, the greater the ambiguity of a null result. To put a null result in context, it is therefore critical to know whether the study was adequately powered and under what assumed effect the power was calculated (Hoenig and Heisey, 2001; Greenland, 2012). However, if the goal of a study is to explicitly quantify the evidence for the absence of an effect, more appropriate methods designed for this task, such as equivalence testing (Wellek, 2010) or Bayes factors (Kass and Raftery, 1995), should be used from the outset.

The contextualization of null results becomes even more complicated in the setting of replication studies. In a replication study, researchers attempt to repeat an original study as closely as possible in order to assess whether similar results can be obtained with new data (National Academies of Sciences, Engineering, and Medicine, 2019). There have been various large-scale replication projects in the biomedical and social sciences in the last decade (Prinz et al., 2011; Begley and Ellis, 2012; Klein et al., 2014; Open Science Collaboration, 2015; Camerer et al., 2016, 2018;

41 *Klein et al., 2018; Cova et al., 2018; Errington et al., 2021*, among others). Most of these projects
42 reported alarmingly low replicability rates across a broad spectrum of criteria for quantifying repli-
43 cability. While most of these projects restricted their focus on original studies with statistically
44 significant results (“positive results”), the *Reproducibility Project: Psychology* (RPP, *Open Science Col-*
45 *laboration, 2015*), the *Reproducibility Project: Experimental Philosophy* (RPEP, *Cova et al., 2018*), and
46 the *Reproducibility Project: Cancer Biology* (RPCB, *Errington et al., 2021*) also attempted to replicate
47 some original studies with null results.

48 The RPP excluded the original null results from its overall assessment of replication success,
49 but the RPCB and the RPEP explicitly defined null results in both the original and the replication
50 study as a criterion for “replication success”. There are several logical problems with this “non-
51 significance” criterion. First, if the original study had low statistical power, a non-significant result
52 is highly inconclusive and does not provide evidence for the absence of an effect. It is then un-
53 clear what exactly the goal of the replication should be – to replicate the inconclusiveness of the
54 original result? On the other hand, if the original study was adequately powered, a non-significant
55 result may indeed provide some evidence for the absence of an effect when analyzed with ap-
56 propriate methods, so that the goal of the replication is clearer. However, the criterion does not
57 distinguish between these two cases. Second, with this criterion researchers can virtually always
58 achieve replication success by conducting two studies with very small sample sizes, such that the
59 p -values are non-significant and the results are inconclusive. This is because the null hypothesis un-
60 der which the p -values are computed is misaligned with the goal of inference, which is to quantify
61 the evidence for the absence of an effect. We will discuss methods that are better aligned with this
62 inferential goal. Third, the criterion does not control the error of falsely claiming the absence of an
63 effect at some predetermined rate. This is in contrast to the standard replication success criterion
64 of requiring significance from both studies (also known as the two-trials rule, see chapter 12.2.8 in
65 *Senn, 2008*), which ensures that the error of falsely claiming the presence of an effect is controlled
66 at a rate equal to the squared significance level (for example, $5\% \times 5\% = 0.25\%$ for a 5% significance
67 level). The non-significance criterion may be intended to complement the two-trials rule for null
68 results, but it fails to do so in this respect, which may be important to regulators, funders, and
69 researchers. We will now demonstrate these issues and potential solutions using the null results
70 from the RPCB.

71 **Null results from the Reproducibility Project: Cancer Biology**

72 Figure 1 shows standardized mean difference effect estimates with confidence intervals from two
73 RPCB study pairs. Both are “null results” and meet the non-significance criterion for replication
74 success (the two-sided p -values are greater than 0.05 in both the original and the replication study),
75 but intuition would suggest that these two pairs are very much different.

76 The original study from *Dawson et al. (2011)* and its replication both show large effect estimates
77 in magnitude, but due to the small sample sizes, the uncertainty of these estimates is very large,
78 too. If the sample sizes of the studies were larger and the point estimates remained the same,
79 intuitively both studies would provide evidence for a non-zero effect. However, with the samples
80 sizes that were actually used, the results seem inconclusive. In contrast, the effect estimates from
81 *Goetz et al. (2011)* and its replication are much smaller in magnitude and their uncertainty is also
82 smaller because the studies used larger sample sizes. Intuitively, these studies seem to provide
83 some evidence for a zero (or negligibly small) effect. While these two examples show the qualitative
84 difference between absence of evidence and evidence of absence, we will now discuss how the two
85 can be quantitatively distinguished.

86 **Methods for assessing replicability of null results**

87 There are both frequentist and Bayesian methods that can be used for assessing evidence for the
88 absence of an effect. *Anderson and Maxwell (2016)* provide an excellent summary of both ap-

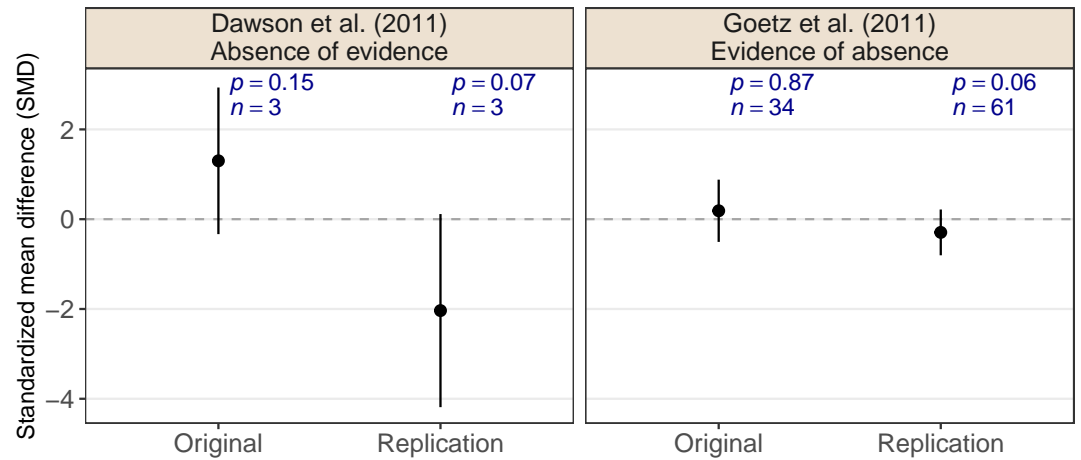


Figure 1. Two examples of original and replication study pairs which meet the non-significance replication success criterion from the Reproducibility Project: Cancer Biology (Errington et al., 2021). Shown are standardized mean difference effect estimates with 95% confidence intervals, sample sizes, and two-sided p -values for the null hypothesis that the standardized mean difference is zero.

proaches in the context of replication studies in psychology. We now briefly discuss two possible approaches – frequentist equivalence testing and Bayesian hypothesis testing – and their application to the RPCB data.

Equivalence testing

Equivalence testing was developed in the context of clinical trials to assess whether a new treatment – typically cheaper or with fewer side effects than the established treatment – is practically equivalent to the established treatment (Westlake, 1972; Schuirmann, 1987). The method can also be used to assess whether an effect is practically equivalent to the value of an absent effect, usually zero. Using equivalence testing as a remedy for non-significant results has been suggested by several authors (Hauck and Anderson, 1986; Campbell and Gustafson, 2018). The main challenge is to specify the margin $\Delta > 0$ that defines an equivalence range $[-\Delta, +\Delta]$ in which an effect is considered as absent for practical purposes. The goal is then to reject the null hypothesis that the true effect is outside the equivalence range. This is in contrast to the usual null hypothesis of a superiority test which states that the effect is zero or smaller than zero, see Figure 2 for an illustration.

To ensure that the null hypothesis is falsely rejected at most $\alpha \times 100\%$ of the time, one either rejects it if the $(1 - 2\alpha) \times 100\%$ confidence interval for the effect is contained within the equivalence range (for example, a 90% confidence interval for $\alpha = 5\%$), or if two one-sided tests (TOST) for the effect being smaller/greater than $+\Delta$ and $-\Delta$ are significant at level α , respectively. A quantitative measure of evidence for the absence of an effect is then given by the maximum of the two one-sided p -values (the TOST p -value).

Returning to the RPCB data, Figure 3 shows the standardized mean difference effect estimates with 90% confidence intervals for the 20 study pairs with quantitative null results in the original study ($p > 0.05$). The dotted red lines represent an equivalence range for the margin $\Delta = 1$, for which the shown TOST p -values are computed. This margin is rather lax compared to the margins typically used in clinical research; we chose it primarily for illustrative purposes and because effect sizes in preclinical research are typically much larger than in clinical research. In practice, the margin should be determined on a case-by-case basis by researchers who are familiar with the subject matter. However, even with this generous margin, only four of the twenty study pairs – one of them

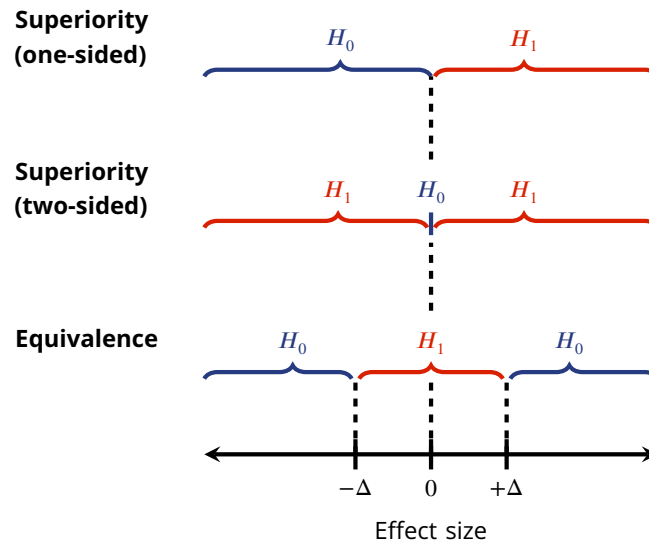


Figure 2. Null hypothesis (H_0) and alternative hypothesis (H_1) for different study designs with equivalence margin Δ .

being the previously discussed example from *Goetz et al. (2011)* – are able to establish equivalence at the 5% level in the sense that both the original and the replication 90% confidence interval fall within the equivalence range (or equivalently that their TOST p -values are smaller than 0.05). For the remaining 16 studies – for instance, the previously discussed example from *Dawson et al. (2011)* – the situation remains inconclusive and there is neither evidence for the absence nor the presence of the effect.

Bayesian hypothesis testing

The distinction between absence of evidence and evidence of absence is naturally built into the Bayesian approach to hypothesis testing. A central measure of evidence is the Bayes factor (*Kass and Raftery, 1995*), which is the updating factor of the prior odds to the posterior odds of the null hypothesis H_0 versus the alternative hypothesis H_1

$$\underbrace{\frac{\Pr(H_0 | \text{data})}{\Pr(H_1 | \text{data})}}_{\text{Posterior odds}} = \underbrace{\frac{\Pr(H_0)}{\Pr(H_1)}}_{\text{Prior odds}} \times \underbrace{\frac{p(\text{data} | H_0)}{p(\text{data} | H_1)}}_{\text{Bayes factor } BF_{01}}.$$

The Bayes factor quantifies how much the observed data have increased or decreased the probability of the null hypothesis H_0 relative to the alternative H_1 . If the null hypothesis states the absence of an effect, a Bayes factor greater than one ($BF_{01} > 1$) indicates evidence for the absence of the effect and a Bayes factor smaller than one indicates evidence for the presence of the effect ($BF_{01} < 1$), whereas a Bayes factor not much different from one indicates absence of evidence for either hypothesis ($BF_{01} \approx 1$).

When the observed data are dichotomized into positive ($p < 0.05$) or null results ($p > 0.05$), the Bayes factor based on a null result is the probability of observing $p > 0.05$ when the effect is indeed absent (which is 95%) divided by the probability of observing $p > 0.05$ when the effect is indeed present (which is one minus the power of the study). For example, if the power is 90%, we have $BF_{01} = 95\%/10\% = 9.5$ indicating almost ten times more evidence for the absence of the effect than for its presence. On the other hand, if the power is only 50%, we have $BF_{01} = 95\%/50\% = 1.9$ indicating only slightly more evidence for the absence of the effect. This example also highlights the main challenge with Bayes factors – the specification of the alternative hypothesis H_1 . The assumed effect under H_1 is directly related to the power of the study, and researchers who assume different

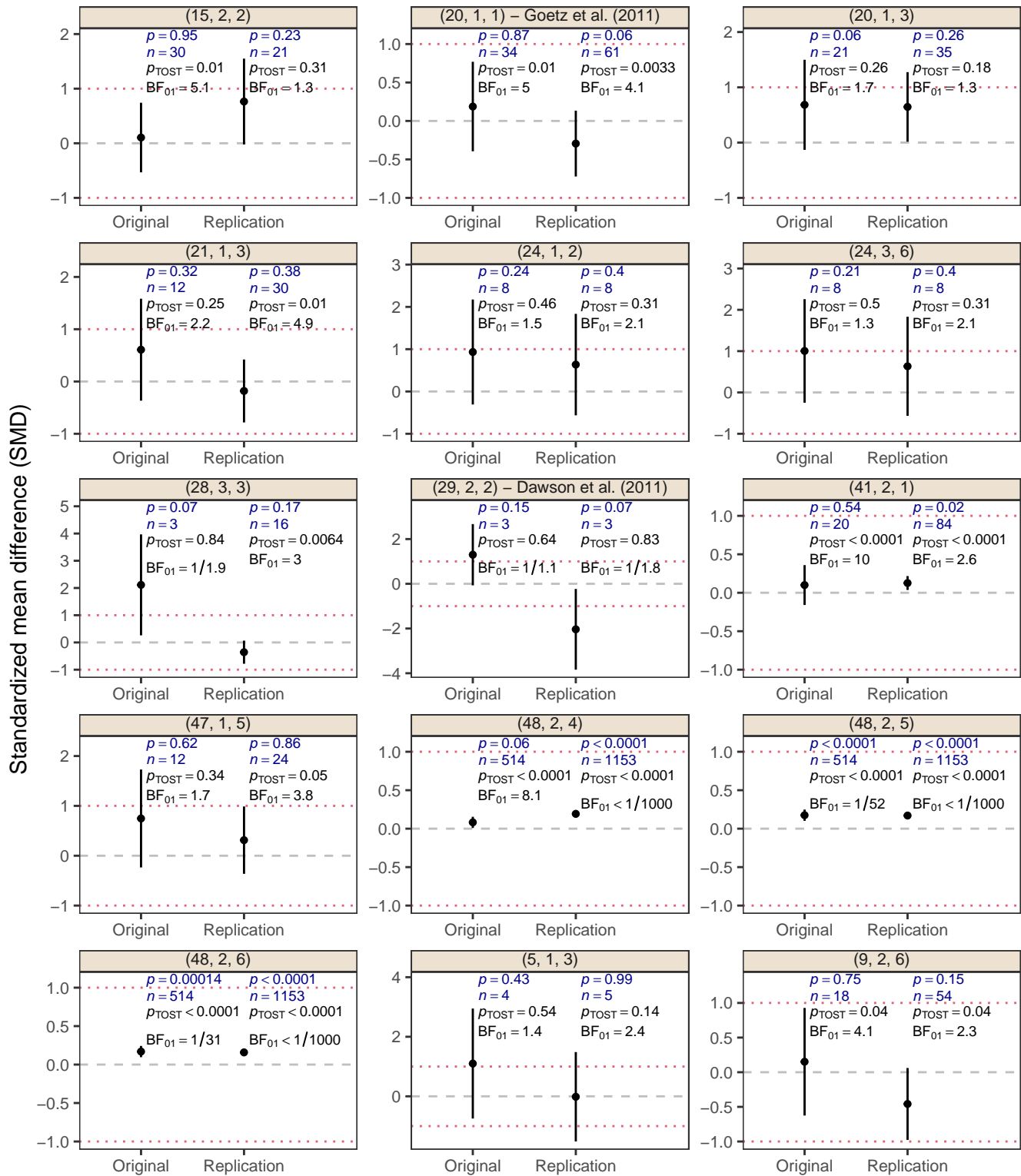


Figure 3. Standardized mean difference (SMD) effect estimates with 90% confidence interval for the “null results” (those with original two-sided p -value $p > 0.05$) and their replication studies from the Reproducibility Project: Cancer Biology (Errington et al., 2021). The identifier above each plot indicates (Original paper number, Experiment number, Effect number). The two examples from Figure 1 are indicated in the plot titles. The dashed grey line depicts the value of no effect ($SMD = 0$) whereas the dotted red lines depict the equivalence range with margin $\Delta = 1$. The p -values p_{TOST} are the maximum of the two one-sided p -values for the effect being smaller or greater than $+\Delta$ or $-\Delta$, respectively. The Bayes factors BF_{01} quantify evidence for the null hypothesis $H_0 : SMD = 0$ against the alternative $H_1 : SMD \neq 0$ with normal unit-information prior assigned to the SMD under H_1 .

effects under H_1 will end up with different Bayes factors. Instead of specifying a single effect, one therefore typically specifies a “prior distribution” of plausible effects. Importantly, the prior distribution, like the equivalence margin, should be determined by researchers with subject knowledge and before the data are observed.

In practice, the observed data should not be dichotomized into positive or null results, as this leads to a loss of information. Therefore, to compute the Bayes factors for the RPCB null results, we used the observed effect estimates as the data and assumed a normal sampling distribution for them, as in a meta-analysis. The Bayes factors BF_{01} shown in Figure 3 then quantify the evidence for the null hypothesis of no effect ($H_0 : SMD = 0$) against the alternative hypothesis that there is an effect ($H_1 : SMD \neq 0$) using a normal “unit-information” prior distribution (Kass and Wasserman, 1995) for the effect size under the alternative H_1 . There are several more advanced prior distributions that could be used here, and they should ideally be specified for each effect individually based on domain knowledge. The normal unit-information prior (with a standard deviation of 2 for SMDs) is only a reasonable default choice, as it implies that small to large effects are plausible under the alternative. We see that in most cases there is no substantial evidence for either the absence or the presence of an effect, as with the equivalence tests. The Bayes factors for the two previously discussed examples from Goetz et al. (2011) and Dawson et al. (2011) are consistent with our intuitions – there is indeed some evidence for the absence of an effect in Goetz et al. (2011), while there is even slightly more evidence for the presence of an effect in Dawson et al. (2011), though the Bayes factor is very close to one due to the small sample sizes. With a lenient Bayes factor threshold of $BF_{01} > 3$ to define evidence for the absence of the effect, only one of the twenty study pairs meets this criterion in both the original and replication study.

Among the twenty RPCB null results, there is one interesting case (the rightmost plot in the fourth row (48, 2, 4, 1)) where the Bayes factor is qualitatively different from the equivalence test, revealing a fundamental difference between the two approaches. The Bayes factor is concerned with testing whether the effect is *exactly zero*, whereas the equivalence test is concerned with whether the effect is within an *interval around zero*. Due to the very large sample size in the original study ($n = 514$) and the replication ($n = 1153$), the data are incompatible with an exactly zero effect, but compatible with effects within the equivalence range. Apart from this example, however, the approaches lead to the same qualitative conclusion – most RPCB null results are highly ambiguous.

Conclusions

We showed that in most of the RPCB studies with “null results” (those with $p > 0.05$), neither the original nor the replication study provided conclusive evidence for the presence or absence of an effect. It seems logically questionable to declare an inconclusive replication of an inconclusive original study as a replication success. While it is important to replicate original studies with null results, our analysis highlights that they should be analyzed and interpreted appropriately.

For both the equivalence testing and the Bayes factor approach, it is critical that the parameters of the procedure (the equivalence margin and the prior distribution) are specified independently of the data, ideally before the studies are conducted. Typically, however, the original studies were designed to find evidence for the presence of an effect, and the goal of replicating the “null result” was formulated only after failure to do so. Campbell and Gustafson (2021) discuss various approaches to post-hoc specification of equivalence margins, such as motivating it using data from previous studies or using field conventions. Hauck and Anderson (1986) propose a sensitivity analysis approach in the form of plotting the TOST p -value against a range of possible margins (“equivalence curves”). Post-hoc specification of a prior distribution for a Bayes factor may likewise be based on historical data, field conventions, or assessed visually with sensitivity analyses.

While the equivalence test and the Bayes factor are two principled methods for analyzing original and replication studies with null results, they are not the only possible methods for doing so. For instance, the reverse-Bayes approach from Micheloud and Held (2022b) specifically tailored to equivalence testing in the replication setting may lead to more appropriate inferences as it also

takes into account the compatibility of the effect estimates from original and replication studies. In addition, there are various other Bayesian methods which could potentially improve upon the considered Bayes factor approach. For example, Bayes factors based on non-local priors (Johnson and Rossell, 2010) or based on interval null hypotheses (Morey and Rouder, 2011; Liao et al., 2020), methods for equivalence testing based on effect size posterior distributions (Kruschke, 2018), or Bayesian procedures that involve utilities of decisions (Lindley, 1998). Finally, the design of replication studies should align with the planned analysis (Anderson and Maxwell, 2017; Anderson and Kelley, 2022; Micheloud and Held, 2022a; Pawel et al., 2022). If the goal of the study is to find evidence for the absence of an effect, the replication sample size should also be determined so that the study has adequate power to make conclusive inferences regarding the absence of the effect.

Acknowledgements

We thank the contributors of the RPCB for their tremendous efforts and for making their data publicly available. We thank Maya Mathur for helpful advice with the data preparation. This work was supported by the Swiss National Science Foundation (grant #189295).

Conflict of interest

We declare no conflict of interest.

Software and data

The code and data to reproduce our analyses is openly available at <https://gitlab.uzh.ch/samuel.pawel/rsAbsence>. A snapshot of the repository at the time of writing is available at <https://doi.org/10.5281/zenodo.XXXXXX>. We used the statistical programming language R version 4.2.3 (R Core Team, 2022) for analyses. The R packages ggplot2 (Wickham, 2016), dplyr (Wickham et al., 2022), knitr (Xie, 2022), and reporttools (Rufibach, 2009) were used for plotting, data preparation, dynamic reporting, and formatting, respectively. The data from the RPCB were obtained by downloading the files from <https://github.com/mayamathur/rpcb> (commit a1e0c63) and extracting the relevant variables as indicated in the R script preprocess-rpcb-data.R which is available in our git repository.

References

- Altman DG, Bland JM. Statistics notes: Absence of evidence is not evidence of absence. BMJ. 1995; 311(7003):485–485. doi: 10.1136/bmj.311.7003.485.
- Anderson SF, Kelley K. Sample size planning for replication studies: The devil is in the design. Psychological Methods. 2022; doi: 10.1037/met0000520.
- Anderson SF, Maxwell SE. There's more than one way to conduct a replication study: Beyond statistical significance. Psychological Methods. 2016; 21(1):1–12. doi: 10.1037/met0000051.
- Anderson SF, Maxwell SE. Addressing the “replication crisis”: Using original studies to design replication studies with appropriate statistical power. Multivariate Behavioral Research. 2017; 52(3):305–324. doi: 10.1080/00273171.2017.1289361.
- Begley CG, Ellis LM. Raise standards for preclinical cancer research. Nature. 2012; 483(7391):531–533. doi: 10.1038/483531a.
- Camerer CF, Dreber A, Forsell E, Ho T, Huber J, Johannesson M, Kirchler M, Almenberg J, Altmeld A, et al. Evaluating replicability of laboratory experiments in economics. Science. 2016; 351:1433–1436. doi: 10.1126/science.aaf0918.
- Camerer CF, Dreber A, Holzmeister F, Ho T, Huber J, Johannesson M, Kirchler M, Nave G, Nosek B, et al. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nature Human Behavior. 2018; 2:637–644. doi: 10.1038/s41562-018-0399-z.

234 **Campbell H**, Gustafson P. Conditional equivalence testing: An alternative remedy for publication bias. *PLOS*
235 *ONE*. 2018; 13(4):e0195145. doi: [10.1371/journal.pone.0195145](https://doi.org/10.1371/journal.pone.0195145).

236 **Campbell H**, Gustafson P. What to make of equivalence testing with a post-specified margin? *Meta-Psychology*.
237 2021; 5. doi: [10.15626/mp.2020.2506](https://doi.org/10.15626/mp.2020.2506).

238 **Cova F**, Strickland B, Abatista A, Allard A, Andow J, Attie M, Beebe J, Berniūnas R, Boudesseul J, Colombo M, et al.
239 Estimating the Reproducibility of Experimental Philosophy. *Review of Philosophy and Psychology*. 2018; doi:
240 10.1007/s13164-018-0400-9.

241 **Dawson MA**, Prinjha RK, Dittmann A, Giotopoulos G, Bantscheff M, Chan WI, Robson SC, wa Chung C, Hopf
242 C, Savitski MM, Huthmacher C, Gudgin E, Lugo D, Beinke S, Chapman TD, Roberts EJ, Soden PE, Auger KR,
243 Mirguet O, Doehner K, et al. Inhibition of BET recruitment to chromatin as an effective treatment for MLL-
244 fusion leukaemia. *Nature*. 2011; 478(7370):529–533. doi: [10.1038/nature10509](https://doi.org/10.1038/nature10509).

245 **Errington TM**, Mathur M, Soderberg CK, Denis A, Perfito N, Iorns E, Nosek BA. Investigating the replicability of
246 preclinical cancer biology. *eLife*. 2021; 10. doi: [10.7554/elife.71601](https://doi.org/10.7554/elife.71601).

247 **Goetz JG**, Minguet S, Navarro-Lérida I, Lazcano JJ, Samaniego R, Calvo E, Tello M, Osteso-Ibáñez T, Pellinen T,
248 Echarri A, Cerezo A, Klein-Szanto AJP, Garcia R, Keely PJ, Sánchez-Mateos P, Cukierman E, Pozo MAD. Biome-
249 chanical Remodeling of the Microenvironment by Stromal Caveolin-1 Favors Tumor Invasion and Metastasis.
250 *Cell*. 2011; 146(1):148–163. doi: [10.1016/j.cell.2011.05.040](https://doi.org/10.1016/j.cell.2011.05.040).

251 **Greenland S**. Nonsignificance Plus High Power Does Not Imply Support for the Null Over the Alternative. *Annals*
252 *of Epidemiology*. 2012; 22(5):364–368. doi: [10.1016/j.annepidem.2012.02.007](https://doi.org/10.1016/j.annepidem.2012.02.007).

253 **Hauck WW**, Anderson S. A proposal for interpreting and reporting negative studies. *Statistics in Medicine*.
254 1986; 5(3):203–209. doi: [10.1002/sim.4780050302](https://doi.org/10.1002/sim.4780050302).

255 **Hoening JM**, Heisey DM. The Abuse of Power. *The American Statistician*. 2001; 55(1):19–24. doi:
256 10.1198/000313001300339897.

257 **Johnson VE**, Rossell D. On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of*
258 *the Royal Statistical Society: Series B (Statistical Methodology)*. 2010; 72(2):143–170. doi: [10.1111/j.1467-](https://doi.org/10.1111/j.1467-9868.2009.00730.x)
259 [9868.2009.00730.x](https://doi.org/10.1111/j.1467-9868.2009.00730.x).

260 **Kass RE**, Raftery AE. Bayes Factors. *Journal of the American Statistical Association*. 1995; 90(430):773–795. doi:
261 [10.1080/01621459.1995.10476572](https://doi.org/10.1080/01621459.1995.10476572).

262 **Kass RE**, Wasserman L. A Reference Bayesian Test for Nested Hypotheses and its Relationship to
263 the Schwarz Criterion. *Journal of the American Statistical Association*. 1995; 90(431):928–934. doi:
264 [10.1080/01621459.1995.10476592](https://doi.org/10.1080/01621459.1995.10476592).

265 **Klein RA**, Ratliff KA, Vianello M, Adams RB, Bahník v, Bernstein MJ, Bocian K, Brandt MJ, Brooks B, et al. Investi-
266 gating variation in replicability: A “many labs” replication project. *Social Psychology*. 2014; 45:142–152. doi:
267 10.1027/1864-9335/a000178.

268 **Klein RA**, Vianello M, Hasselman F, Adams BG, Reginald B Adams J, Alper S, Aveyard M, Axt JR, Babalola MT,
269 et al. Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods*
270 *and Practices in Psychological Science*. 2018; 1(4):443–490. doi: [10.1177/2515245918810225](https://doi.org/10.1177/2515245918810225).

271 **Kruschke JK**. Rejecting or Accepting Parameter Values in Bayesian Estimation. *Advances in Methods and*
272 *Practices in Psychological Science*. 2018; 1(2):270–280. doi: [10.1177/2515245918771304](https://doi.org/10.1177/2515245918771304).

273 **Liao JG**, Midya V, Berg A. Connecting and Contrasting the Bayes Factor and a Modified ROPE Pro-
274 cedure for Testing Interval Null Hypotheses. *The American Statistician*. 2020; 75(3):256–264. doi:
275 [10.1080/00031305.2019.1701550](https://doi.org/10.1080/00031305.2019.1701550).

276 **Lindley DV**. Decision analysis and bioequivalence trials. *Statistical Science*. 1998; 13(2). doi:
277 10.1214/ss/1028905932.

278 **Makin TR**, de Xivry JJO. Ten common statistical mistakes to watch out for when writing or reviewing a
279 manuscript. *eLife*. 2019; 8. doi: [10.7554/elife.48175](https://doi.org/10.7554/elife.48175).

280 **Micheloud C**, Held L. Power Calculations for Replication Studies. *Statistical Science*. 2022; 37(3):369–379. doi:
281 10.1214/21-sts828.

282 **Micheloud C**, Held L, The replication of non-inferiority and equivalence studies. arXiv; 2022. doi:
 283 [10.48550/ARXIV.2204.06960](https://doi.org/10.48550/ARXIV.2204.06960), arXiv preprint.

284 **Morey RD**, Rouder JN. Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*.
 285 2011; 16(4):406–419. doi: 10.1037/a0024377.

286 **National Academies of Sciences, Engineering, and Medicine**. Reproducibility and Replicability in Science.
 287 National Academies Press; 2019. doi: 10.17226/25303.

288 **Open Science Collaboration**. Estimating the reproducibility of psychological science. *Science*. 2015;
 289 349(6251):aac4716. doi: [10.1126/science.aac4716](https://doi.org/10.1126/science.aac4716).

290 **Pawel S**, Consonni G, Held L, Bayesian approaches to designing replication studies. arXiv; 2022. doi:
 291 [10.48550/ARXIV.2211.02552](https://doi.org/10.48550/ARXIV.2211.02552), arXiv preprint.

292 **Prinz F**, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug
 293 targets? *Nature Reviews Drug Discovery*. 2011; 10(9):712–712. doi: 10.1038/nrd3439-c1.

294 **R Core Team**. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing,
 295 Vienna, Austria; 2022, <https://www.R-project.org/>.

296 **Rufibach K**. reporttools: R Functions to Generate \LaTeX Tables of Descriptive Statistics. *Journal of Statistical*
 297 *Software, Code Snippets*. 2009; 31(1). doi: [10.18637/jss.v031.c01](https://doi.org/10.18637/jss.v031.c01).

298 **Schuurmann DJ**. A comparison of the Two One-Sided Tests Procedure and the Power Approach for assess-
 299 ing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*. 1987;
 300 15(6):657–680. doi: 10.1007/bf01068419.

301 **Senn S**. *Statistical Issues in Drug Development*, vol. 69. John Wiley & Sons; 2008.

302 **Wellek S**. *Testing statistical hypotheses of equivalence and noninferiority*. CRC press; 2010.

303 **Westlake WJ**. Use of Confidence Intervals in Analysis of Comparative Bioavailability Trials. *Journal of Pharma-*
 304 *ceutical Sciences*. 1972; 61(8):1340–1341. doi: [10.1002/jps.2600610845](https://doi.org/10.1002/jps.2600610845).

305 **Wickham H**. *ggplot2: Elegant Graphics for Data Analysis*. Springer International Publishing; 2016. doi:
 306 [10.1007/978-3-319-24277-4](https://doi.org/10.1007/978-3-319-24277-4).

307 **Wickham H**, François R, Henry L, Müller K. *dplyr: A Grammar of Data Manipulation*; 2022, [https://CRAN.](https://CRAN.R-project.org/package=dplyr)
 308 [R-project.org/package=dplyr](https://CRAN.R-project.org/package=dplyr), r package version 1.0.10.

309 **Xie Y**. *knitr: A General-Purpose Package for Dynamic Report Generation in R*; 2022, <https://yihui.org/knitr/>, r
 310 package version 1.40.

Computational details

```
cat(paste(Sys.time(), Sys.timezone(), "\n"))

## 2023-03-29 17:52:02 Europe/Zurich

sessionInfo()

## R version 4.2.3 (2023-03-15)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 20.04.6 LTS
##
## Matrix products: default
## BLAS: /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.9.0
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.9.0
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8      LC_NUMERIC=C
##  [3] LC_TIME=de_CH.UTF-8      LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=de_CH.UTF-8  LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=de_CH.UTF-8     LC_NAME=C
##  [9] LC_ADDRESS=C             LC_TELEPHONE=C
## [11] LC_MEASUREMENT=de_CH.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats      graphics  grDevices  utils      datasets  methods   base
##
## other attached packages:
## [1] reporttools_1.1.3 xtable_1.8-4      dplyr_1.0.10      ggplot2_3.4.0
## [5] knitr_1.41
##
## loaded via a namespace (and not attached):
##  [1] magrittr_2.0.3    tidyselect_1.2.0  munsell_0.5.0     colorspace_2.1-0
##  [5] R6_2.5.1          rlang_1.0.6       fansi_1.0.3       highr_0.10
##  [9] stringr_1.5.0     tools_4.2.3       grid_4.2.3        gtable_0.3.1
## [13] xfun_0.36         utf8_1.2.2        cli_3.6.0         DBI_1.1.3
## [17] withr_2.5.0       assertthat_0.2.1  tibble_3.1.8      lifecycle_1.0.3
## [21] farver_2.1.1      vctrs_0.5.1       glue_1.6.2        evaluate_0.20
## [25] labeling_0.4.2    stringi_1.7.12    compiler_4.2.3    pillar_1.8.1
## [29] generics_0.1.3    scales_1.2.1      pkgconfig_2.0.3
```