A Statistical Case for Qualified Scientific Optimism

Erik van Zwet* Andrew Gelman† Witold Więcek‡

12 Feb 2026

Abstract

We create an open-source database from nearly 80,000 studies, combining 11 publicly available curated collections of empirical research in medicine, economics, psychology, psychotherapy, ecology, evolution, political science, education, and more, and make it available at `https://github.com/wwiecek/BEAR`. Each result is represented by a z-value: an estimated effect divided by its standard error.

We then provide a modeling framework that allows us to move away from the usual attempts to classify effects as exactly zero or not, and instead focus on identifying the direction of effects and their replicability.

After adjusting for publication bias, we find two patterns that hold across all datasets and their various subsets: (1) statistical significance does not imply a high probability of successful replication, but (2) it does imply a high probability that the direction of the observed effect was correct.

We conclude that replication rate is a poor measure of the quality of a field of research and argue that an attitude of global scientific pessimism based on low replication rates is misguided. Instead, we favor a more optimistic view of scientific research based on the ability to identify the direction of effects. This optimism is not naive about the many challenges science continues to face such as questionable research practices (QRPs), paper mills, fake AI papers and such.

1. Studying the quality of empirical science

The replication crisis has made us aware that there's a wide range of quality in the life and behavioral sciences, even considering subsets such as particular research areas and papers published in top journals. This makes sense; if we roughly measure the quality of a study by its signal-to-noise ratio (the true effect size divided by its standard error), then we have some control over the denominator (by taking more or better measurements) but less control over the numerator. It's the nature of research to be uncertain and it's the duty of researchers to pursue all avenues, studying effects which might turn out to be null or highly context-dependent.

A key turning point in the replication crisis was the recognition and refutation of the false intuition that you can tell whether a study is good or not (tell whether a result is strong or not) using a statistical significance threshold. Beyond the problems with any sharp threshold, even seemingly clear distinctions such as, for example, $p < 0.01$ compared to $p > 0.2$ do not discriminate well between high and low signal-to-noise ratios,[1] and a similar problem arises not just with p-values but also with Bayes factors and other measures of quality of evidence.

The message of the present paper is that the difficulty of assessing the signal-to-noise ratio for any given study should not be taken as a negative verdict on statistically-based science as a whole.

To study this, we have created an extensive open-source database of empirical research results, composed of 15 publicly available datasets of studies in diverse fields (11 curated sets, 2 replication

*Department of Biomedical Data Sciences, Leiden University Medical Center.

†Department of Statistics and Department of Political Science, Columbia University, New York.

‡Development Innovation Lab, University of Chicago.

[1] The two-sided p-values of 0.01 and 0.2 correspond to absolute z-values of 2.58 and 1.28, respectively. These are about equally likely when the signal-to-noise ratio is $(2.58 + 1.28)/2 = 1.93$. This signal-to-noise ratio corresponds to a little under 50% power as $\Pr(|z| \geq 1.96 \mid \text{SNR} = 1.93) = 0.49$.

sets and 2 automated "scrapes"). Our optimistic finding is that, in these corpora, most results seem to be correct in sign. Our remaining pessimism is that you can't learn so much from any individual study, which points toward the need for real replication—not in the hope of confirming significance but to understand effects, and to minimize selection in all phases of the process.

1.1. Going beyond the characterization of scientific claims as true or false

In an influential paper, Ioannidis (2005) modeled empirical research as "a 2 × 2 table in which research findings are compared against the gold standard of true relationships in a scientific field."

This 2 × 2 model is implicitly accepted in the meta-scientific literature, which tends to focus on statistical significance and replication rates (Simmons, Nelson, and Simonsohn 2011; Open Science Collaboration 2015; Camerer et al. 2016). In particular, low replication rates are taken by many as evidence of published empirical research being full of null claims and hence not credible as a whole, a position with political influence; see, for example, Garisto (2025).

We understand that the 2 × 2 model was always intended as an instructive simplification, introduced for the purpose of explaining that in an area of research where the power is low and there are few true relationships, many statistically significant results may not be true. However, far-reaching conclusions from this model have been made about the actual state of scientific research, in particular the claim that more than half of all published research findings are false.

We have two main criticisms of the 2 × 2 model. First, we disagree with the notion that the primary goal of scientific study is to separate "no relationships" or nulls from "true relationships." Effects of interest are rarely exactly zero, and even if they are, they will no longer be zero if one takes leakage, expectation effects, dropout, etc. into account. But apart from that, with finite data, it is simply not possible to distinguish effects that are zero from those that are small.

Our second criticism is that the 2 × 2 model does not comport with data we have collected on published research. We have compiled a database of 15 corpora, each containing thousands of studies from medicine, economics, psychology, psychotherapy, ecology, evolution, political science, education, and more. We have used these data to build more realistic models of research areas.

In this paper we make two contributions. First, we make a large number of results from empirical research easily accessible in a new database, Benchmarks of Empirical Accuracy in Research (BEAR). We do not contribute any new data, but we process the available data to make them ready for meta-scientific research. We provide clear documentation and all data are available under open source license at `https://github.com/wwiecek/BEAR`. We will continue to add new datasets as they become available.

Second, we extend the meta-scientific model of Stephens (2017), van Zwet, Schwab, and Senn (2021), and van Zwet and Gelman (2022) to account for truncation and publication bias. This model allows us to study measures of research quality that are insensitive to whether effects are exactly zero or merely very small, such as the probability of correctly identifying the direction of an effect and the probability of replication (Greenland 2017; Gelman and Carlin 2014; Gelman and Tuerlinckx 2000).

1.2. Summary of findings

We estimate from our analysis that statistical significance usually implies a high probability that the direction of the observed effect is correct, but that it does not imply a high probability that a hypothetical exact replication study would reach statistical significance. In other words, it is expected that true findings (in the sense of correctly identifying the direction of an effect) will often not replicate. Similar conclusions were reached by Bak-Coleman et al. (2022) and Neves, Tan, and Amaral (2022) on the basis of simulation models.

The scientific process is far from perfect, and questionable research practices, paper mills, and outright fraud are serious concerns (Aquarius et al. 2025; Matusz, Abalkina, and Bishop 2025). But the mere fact that many studies have low probability of reaching statistical significance and therefore fail to replicate does not imply that most published research findings are false.

The challenge is internalizing the juxtaposition of two ideas. On one hand, there really is a replication crisis: substantive and statistical theory, empirical evidence, and sociological reasoning all point toward the conclusion that traditional procedures for scientific quality control (peer review, causal identification, and statistical significance) are not enough to ensure replicability or to bring us closer to an understanding of reality. On the other hand, the statistical evidence implies that in many fields, published results that reach $p < 0.05$ generally fall in the correct direction.

2. Data

We have compiled a large number of empirical results from a variety of sources in a new database which we call Benchmarks of Empirical Accuracy in Research (BEAR). The online version of BEAR includes 20 datasets and 11.5 million z-values from empirical research, but for this paper we focus mainly on the z-values found in nearly 80,000 studies across 11 curated datasets. For comparison, we also show two sets of replications and two large-scale reviews which scraped p-values. Characteristics of the presented datasets are given in Table 1 and in Appendix B. We describe our processing steps in Appendix C.

Among the 11 curated, domain-specific collections that are our focus here, we include: a large collection of mainly clinical trials which posted results on `clinicaltrials.gov` or the EU Clinical Trials Repository (National Library of Medicine 2025; European Medicines Agency 2025), a subset of 30,000 trials from the Cochrane Database of Systematic Reviews (CDSR; (Schwab 2024; Cochrane Collaboration 2025)), which is well-curated and large (over 400,000 data rows), and 13,000 studies in ecology and evolution (Costello and Fox 2022; Yang, van Zwet, et al. 2024).

Other datasets (each numbering from hundreds to about 2,000 studies) are in education research, compiled by What Works Clearinghouse (Institute of Education Sciences 2025); psychotherapy (Cuijpers, Harrer, Miguel, et al. 2025), psychology (Rodriguez and Williams 2022), cognitive impacts of exercise (Bartoš et al. 2025; Singh et al. 2025), intelligence research (Nuijten et al. 2020), economics (Askarov et al. 2023; Brodeur et al. 2024), and political science (Arel-Bundock et al. 2022).

Six of the datasets are derived from curated databases used mainly for aggregation and evaluation of interventions. The rest are usually meta-scientific investigations, for example of pre-registration, publication bias, statistical power. Some datasets are themselves sets of meta-analyses, but in this paper we treat them as collections of individual studies. However, the meta-analytic structure is retained in BEAR and may be useful for other meta-scientific research, for example characterizing heterogeneity in treatment effects (van Zwet, Więcek, and Gelman 2025).

For comparison, we also include two replication projects in psychology (Open Science Collaboration 2015; Klein et al. 2018) and two large-scale data-scraping papers which used abstracts or full texts of biomedical literature available via Medline and PubMed: Chavalarias et al. (2016) provide 8 million p-values across 1.8 million studies, while Barnett and Wren (2019) collect effect estimates of binary outcomes (odds ratios, risk ratios and such) and their confidence intervals in over 400,000 studies.

While we do some data processing of our own, all of data collection and the vast majority of data processing was already done by the creators of the datasets we reference. Most of them already include z-values or effect sizes and standard errors, but in some cases we make calculations ourselves

Corpus	Purpose of dataset	studies	$\bar{k}$	% signif.
Cochrane Collaboration (2025)	Database of systematic reviews in health and medicine	30,306	1.3	31%
ClinicalTrials.gov + EU CTR	Registries of clinical trials in the United States and European Union	24,468	2.0	52%
Costello and Fox (2022)	Do effects in ecology decline over time?	12,927	6.8	42%
Arel-Bundock et al. (2022)	Assess statistical power in political science research	2,252	7.4	47%
Bartoš et al. (2025)	Effect of physical exercise on cognition, memory, and executive function	2,239	1.0	26%
Askarov et al. (2023)	Impact of mandatory data sharing on statistical significance in economics papers	1,913	11.2	52%
Nuijten et al. (2020)	Meta-meta-analysis of studies in intelligence research	1,913	1.3	53%
Metapsy (2025)	Database and meta-analyses of psychotherapy RCTs	1,494	2.9	48%
What Works Clearinghouse	Database of education intervention effect sizes (Institute of Education Sciences 2025)	1,408	8.6	34%
psymetadata	Curated psychology meta-analysis datasets (Rodriguez and Williams 2022)	721	11.8	36%
Brodeur et al. (2024)	Pre-registration and pre-analysis plans in economics journals?	176	47.9	37%
Many Labs 2 (Klein et al. 2018)	Replication of classic and contemporary psychology experiments across labs	1,414	1.0	44%
Open Science Collaboration (2015)	Replications of a quasi-random sample of 100 experiments in psychology	99	1.0	35%
Chavalarias et al. (2016)	Reporting of p-values over time	1,887,178	4.2	84%
Barnett and Wren (2019)	Bias for statistical significance in health and medical journals	416,027	3.1	83%

Table 1: Datasets currently included in BEAR, divided into three groups: curated datasets, meta-analyses, and datasets based on large-scale scraping of published results; $\bar{k}$ is the mean number of observations per study, and "% signif." is the percentage of absolute z-values exceeding 1.96.

(for example, from p-values or counts for binary data); calculation details are in *Appendix: Datasets included in BEAR*. The average number of estimates per study for each dataset is given in Table 1; it can be as high as 50 z-values per study in the case of Arel-Bundock et al. (2022), which aimed to collect all tests reported by economics papers, but in most cases we have either one or several estimates per study.

3. Statistical model

We extend the model of Stephens (2017), van Zwet, Schwab, and Senn (2021), and van Zwet and Gelman (2022) to account for the truncation and publication bias which are conspicuously present in some of our datasets. The main goal is to estimate a latent distribution of the signal-to-noise ratios across a field of research and then use this distribution to derive various aspects of the sampling distribution of replication results.

3.1. Signal-to-noise ratios

Suppose that we have a collection of unbiased effect estimates which are normally distributed with known standard errors. The z-value is the ratio of the effect estimate to its standard error. We define the signal-to-noise ratio (SNR) as the ratio of the (unobserved) true effect to the standard error of its estimate. Our assumptions imply that the z-value is equal to the SNR plus an independent standard normal error.

The null hypothesis that there is no effect is equivalent to the hypothesis that the SNR is zero. This hypothesis is commonly rejected when the absolute value of the z-value exceeds 1.96 or, equivalently, when the associated two-sided p-value is less than 0.05. The type I error probability of this test is $\alpha = 0.05$. In our notation,

$$\Pr(|z| \geq 1.96 \mid \text{SNR} = 0) = 0.05.$$

The probability of statistical significance (PoS) depends on the true effect and the standard error through the magnitude of the SNR,

$$\text{PoS}(\text{SNR}) = \Pr(|z| \geq 1.96 \mid \text{SNR}) = \Phi(-1.96 - |\text{SNR}|) + 1 - \Phi(1.96 - |\text{SNR}|). \tag{1}$$

For example, if the SNR $= 2.8$ (or -2.8), then the PoS is 80%. Averaging the PoS over the distribution of the SNR in a field of research, we obtain the "assurance" which we denote by $\overline{\text{PoS}}$. It is the proportion of studies for that field that reaches the 5% level of statistical significance. We can also consider $\Pr(|\text{SNR}| \geq 2.8) = \Pr(\text{PoS}(\text{SNR}) \geq 0.8)$, which is the proportion of studies for which the power relative to the true effect size is 80% or more.

The inherent variability of p-values can be most easily understood by considering SNRs and z-values. Suppose a study has 80% power. In that case, the sampling distribution of the z-value is normal with mean 2.8 and unit variance. So, both $z = 0.8 (p = 0.42)$ and $z = 4.8 (p = 1.6 \cdot 10^{-6})$ can easily occur.

3.2. A model for the distribution of signal-to-noise ratios in a corpus

We will use maximum likelihood to estimate the distribution of SNRs given a sample of absolute z-values. Our assumptions imply that z-values are equal to SNRs plus independent standard normal errors. We can therefore obtain the distribution of the absolute SNRs by deconvolution of the standard normal error component from the distribution of the absolute z-values (Efron 2016; Stephens 2017).

We construct the likelihood in four stages. First, we introduce a mixture distribution for the absolute z-values. Then, we add a parameter to account for publication bias. Next, we account for the fact that reported values are sometimes truncated. Finally, we weight the observations according to the number of z-values per study.

Our main assumption is that the distribution of the absolute z-values is well represented by a mixture of half-normal distributions. We will informally verify this assumption by inspection of the observed histograms. Noting that this assumption is equivalent to modeling the signed z-values as a mixture of zero-mean normal distributions, we have the mixture density,

$$f(z \mid p, \sigma) = \sum_{i=1}^{k} p_i \, \frac{1}{\sigma_i} \, \phi\left(\frac{z}{\sigma_i}\right), \tag{2}$$

where $\phi(\cdot)$ is the standard normal density, $p = (p_1, \ldots, p_k)$ are the non-negative mixture weights that sum to unity and $\sigma = (\sigma_1, \ldots, \sigma_k)$ are the standard deviations. These standard deviations are at least 1 because we know that z-values have a standard normal noise component. We will use $k = 4$, as we have verified that larger values of k yield nearly identical results.

The deconvolution to obtain the distribution of the SNRs is easy. We simply subtract 1 from the variances of the mixture components. So,

$$f_{\text{SNR}}(x \mid p, \sigma) = f(x \mid p, \sqrt{\sigma^2 - 1}). \tag{3}$$

To account for publication bias, we introduce a simple selection component in the likelihood. We follow Hedges (1984, 1992) and define ω as the relative risk that a result with $|z| < 1.96$ is observed compared to a result with $|z| \geq 1.96$

$$\omega \; = \; \frac{\Pr(\text{result is published} \mid |z| < 1.96)}{\Pr(\text{result is published} \mid |z| \geq 1.96)}. \tag{4}$$

The probability of publication is

$$C(p, \sigma, \omega) = \omega \Pr(|z| < 1.96 \mid p, \sigma) + \Pr(|z| \geq 1.96 \mid p, \sigma) \tag{5}$$

and the density of the z-values conditional on publication becomes

$$f(z \mid p, \sigma, \omega, \text{pub}) = \begin{cases} \omega \, f(z \mid p, \sigma)/C(p, \sigma, \omega) & |z| \leq 1.96 \\ f(z \mid p, \sigma)/C(p, \sigma, \omega) & \text{otherwise.} \end{cases} \tag{6}$$

The z-value is sometimes left or right truncated, most often at 1.96, but also at, for example, 2.57 (when papers report $p < 0.01$), 3.29 ($p < 0.001$), or 1.64 ($p < 0.1$). To take this into account, we introduce a censoring indicator δ taking the values $<$, $>$, or $=$ for left censoring, right censoring, or no censoring, respectively. We denote the (possibly) censored value of the z-value by $\tilde{z}$. The distribution of the pair $(\delta, \tilde{z})$ conditional on publication is

$$f(\tilde{z}, \delta \mid p, \sigma, \omega, \text{pub}) =$$
$$\begin{cases} \omega \, f(\tilde{z} \mid p, \sigma)/C(p, \sigma, \omega) & \delta = \texttt{"="}, \; \tilde{z} < 1.96, \\ f(\tilde{z} \mid p, \sigma)/C(p, \sigma, \omega) & \delta = \texttt{"="}, \; \tilde{z} \geq 1.96, \\ \omega \, F(\tilde{z} \mid p, \sigma)/C(p, \sigma, \omega) & \delta = \texttt{"<"}, \; \tilde{z} < 1.96, \\ (\omega \, F(1.96 \mid p, \sigma) + F(\tilde{z} \mid p, \sigma) - F(1.96 \mid p, \sigma))/C(p, \sigma, \omega) & \delta = \texttt{"<"}, \; \tilde{z} \geq 1.96, \\ (1 - F(\tilde{z} \mid p, \sigma))/C(p, \sigma, \omega) & \delta = \texttt{">"}, \; \tilde{z} \geq 1.96, \\ ((1 - F(1.96 \mid p, \sigma)) + \omega \, (F(1.96 \mid p, \sigma) - F(\tilde{z} \mid p, \sigma)))/C(p, \sigma, \omega) & \delta = \texttt{">"}, \; \tilde{z} < 1.96. \end{cases}$$

The first two rows cover the case without truncation. Rows three and four are left censoring (statements such as $p > 0.05$), and rows five and six deal with right censoring (e.g., $p < 0.001$). We also introduce artificial censoring when z-values are reported to be exactly zero, by treating them as $|z| < 0.5$.

Most of the datasets have multiple z-values per study. To prevent studies with many z-values from dominating the likelihood, we weight observations by w_j, the inverse of the number of z-values originating from study j. This does not account for the dependence among z-values from the same study but that does not concern us because we do not attempt to quantify the uncertainty of our estimates. The reason is that our datasets are so large that the sampling uncertainty which determines standard errors and confidence intervals, is negligible compared to model uncertainty. The full weighted log likelihood is

$$\ell(p, \sigma, \omega) = \sum_{j=1}^{n} w_j \log f(\tilde{z}_j, \delta_j \mid p, \sigma, \omega, \text{pub}). \tag{7}$$

Even for large datasets, numerical optimization of this log likelihood is feasible, taking just minutes. Large datasets with hundreds of thousands or even millions of observations could be randomly subsampled to, say, 50,000 observations. However, we would see no reason to do this, because if such a large corpus were available, it would make more sense to break it up and estimate the distribution of SNRs separately for each sub-corpus.

The estimated distribution of the z-values without publication bias is $f(z \mid \hat{p}, \hat{\sigma})$ as defined in (2) and consequently the estimated distribution of the SNRs is

$$f_{\text{SNR}}(x \mid \hat{p}, \hat{\sigma}) = f(x \mid \hat{p}, \sqrt{\hat{\sigma}^2 - 1}). \tag{8}$$

We can derive many important quantities from the distribution of the SNRs. For example, the PoS is a function of the SNR, so we can obtain its distribution by transforming the distribution of the SNRs. In particular, the proportion of studies that have at least 80% probability of reaching statistical significance is

$$\Pr(\text{PoS} \geq 0.8) = \Pr(|\text{SNR}| \geq 2.8). \tag{9}$$

The marginal density of the SNRs, together with the conditional density of the z-values given the SNRs, determines the joint distribution. And from the joint distribution we can derive the conditional distribution of the SNR given the z-value. We provide the formulas in the Appendix A. In turn, this allows us to derive the conditional probability of the correct sign,

$$\text{Probability of the correct sign} = \Pr(z \cdot \text{SNR} > 0 \,\big|\, |z|), \tag{10}$$

Finally, we imagine that a second replication study with exactly the same SNR yields another z-value z_{repl}. This z_{repl} is conditionally independent of the original z given SNR, and has the same distribution. Thus van Zwet and Goodman (2022) were able to derive the conditional distribution of $|z_{\text{repl}}|$ given $|z|$. In particular, we obtain

$$\text{Probability of “successful replication”} = \Pr(z_{\text{repl}} \cdot z > 0 \text{ and } |z_{\text{repl}}| > 1.96 \,\big|\, |z|) \tag{11}$$

All the computations that yielded the results reported in the next section are documented in the online supplement to this paper.

4. Results

Figure 1 displays the histograms of the observed absolute z-statistics of our 15 corpora. The solid curves are the fitted mixtures of half-normal distributions conditional on publication, and the dashed curves are the "corrected" unconditional distributions. We find that the estimated distributions generally track the histogram well and conclude that our model is appropriate for our purpose of quantifying large-scale properties of research areas. Our main focus is the 11 curated, domain-specific sets (sets of single studies or collections of meta-analyses), which comprise the first 3 rows, shown in blue and arranged by their estimated assurances $\overline{\text{PoS}}$.

Mixtures of half-normal densities are necessarily decreasing. Apart from the effects of publication bias, we also see this feature in the histograms. We claim that this is typical of large collections of z-values. To further support this claim, we have broken down the Cochrane data by 19 medical specialties and the data from `clinicaltrials.gov` by trial phase, see Figure S2. Again, we see decreasing distributions of absolute z-values.

The discontinuities ("jumps") of the densities at $|z| = 1.96$ are a feature of Hedges' selection model. It is tempting to focus on the extent of publication bias, which may in principle be quantified by the estimated relative risk of publication $\hat{\omega}$ as reported in Table S2. However, there are several reasons why we must resist that temptation. Many studies in our datasets have multiple outcomes, and we do not know which of them drive publication decisions. Moreover, while Hedges' single-parameter selection model is useful for adjusting our estimates of the distributions of the z-statistics, it seems too limited for reliably quantifying and comparing the extent of publication bias across fields of research. This is also not the goal of our study.

The main difference between the distributions of the z-values is their widths. A wider distribution of z-values implies a wider distribution of SNRs. In other words, the studies in that research area tend to have larger absolute SNRs. Since the PoS is an increasing function of the absolute SNR, this also means that a higher proportion of results will reach statistical significance.

The two columns in Table S2 that are labeled "Significance" report the observed proportions of significant results and the estimated ("corrected") assurances in the 15 corpora. Recall that the assurance $\overline{\text{PoS}}$ is the proportion of significant results, adjusted for publication bias; see (1).

The assurance is strictly less than the raw proportion of significant results, unless $\hat{\omega} = 1$ so there is no discernible publication bias. Among the 11 curated datasets, the estimated assurances range from 0.23 to 0.48. The estimated assurance of phase 3 clinical trials in Figure S2 matches the estimated success rates (probability of moving to regulatory approval following trial) in phase 3 trials in drug development reported by Thomas et al. (2021).

By transforming the distribution of the signal-to-noise ratios, we obtain the full distribution of the PoS within each dataset. The left panel of Figure 2 displays the complementary cumulative distribution functions (survival functions) of the PoS in the 15 datasets. In other words, it shows what proportions of studies in each dataset "survive" to achieve a given level of power. We find that between 8% and 30% of studies achieve 80% PoS, see Table S2.

Since the great majority of studies are targeted at reaching 80% power, the proportion of studies that reach 80% PoS may be tentatively interpreted as the proportions of studies where the assumptions of the sample size calculation were true. In general, this proportion is low. However, sample size calculations are not supposed to be aimed at the true effect (which is never known) but at the effect one would not want to miss (Senn 2002). In that sense, a low proportion of studies with 80% PoS just means that it is hard to come up with interventions that have large effects. Pressures of time, money, and logistics also can drive down sample sizes.

The three columns in Table S2 that are labeled "Successful replication" report what would happen if a randomly selected study from a particular corpus would be replicated exactly. We show the

		Significance		“Successful” replication			Correct sign		
Corpus	$\hat{\omega}$	signif.	$\overline{PoS}$	uncond.	$\lvert z\rvert = 1.96$	$\lvert z\rvert \geq 1.96$	uncond.	$\lvert z\rvert = 1.96$	$\lvert z\rvert \geq 1.96$
Cochrane	0.70	0.31	0.23	0.21	0.26	0.60	0.69	0.82	0.94
ctgov / EU CTR	0.95	0.52	0.47	0.46	0.38	0.75	0.87	0.96	0.99
Costello and Fox	0.79	0.42	0.39	0.37	0.34	0.74	0.79	0.89	0.98
Arel-Bundock et al.	0.65	0.47	0.40	0.38	0.36	0.69	0.85	0.96	0.99
Bartos et al.	0.82	0.26	0.23	0.20	0.23	0.54	0.76	0.91	0.97
Askarov et al.	0.71	0.52	0.48	0.47	0.40	0.75	0.88	0.96	0.99
Nuijten et al.	0.85	0.53	0.48	0.47	0.41	0.75	0.88	0.96	0.99
Metapsy	0.83	0.48	0.46	0.45	0.41	0.73	0.88	0.96	0.99
What Works	0.88	0.34	0.32	0.30	0.31	0.69	0.76	0.88	0.97
psymetadata	0.67	0.36	0.33	0.31	0.29	0.75	0.72	0.81	0.96
Brodeur et al.	0.79	0.37	0.36	0.34	0.33	0.66	0.83	0.95	0.99
Many Labs 2	0.99	0.44	0.44	0.42	0.30	0.85	0.75	0.79	0.97
OpenSciCollab	1.00	0.35	0.36	0.34	0.31	0.74	0.78	0.90	0.98
Chavalarias et al.	0.20	0.84	0.51	0.50	0.44	0.76	0.89	0.97	1.00
Barnett and Wren	0.08	0.83	0.32	0.30	0.32	0.61	0.82	0.95	0.98

Table 2: Summary of signal-to-noise ratio modeling results for the 15 corpora of studies in our data. Calculations of power, replication, and direction of effects use the distributions of z-values and signal-to-noise ratios by fitted mixture models without selection.

probability of obtaining a statistically significant result in the same direction both unconditionally and conditionally on the original study having $|z| = 1.96$ or $|z| \geq 1.96$, cf equation (11). We find that if a result is marginally statistically significant at the 5% level (i.e., $|z| = 1.96$) then the probability of “successful replication” is never much more than the base rate, and often even less. This will be a surprise to those who think of statistical significance as a stand-in for replication. This phenomenon is *not* a consequence of many studies having low SNRs (low PoS). If all the studies had higher SNRs, then the assurance would be higher but observing $|z| = 1.96$ would be even less impressive.

The three columns that are labeled “Correct sign” report the estimated proportion of studies where the observed effect has the correct direction (sign). Again, we also condition on $|z| = 1.96$ or $|z| \geq 1.96$, cf as defined in equation (10).

If we focus on the case where $|z| = 1.96$ then we find a large difference between the probabilities of successful replication and the correct sign. Among the 11 curated datasets, the estimated probability of replication ranges from 26% to 41% while the estimated probability of correct sign ranges from 82% to 96%.

The right panel of Figure 2 displays the probabilities of successful replication and correct sign conditional on the observed absolute z-value, see also Figure S1. The probability of correct sign is always much higher than the probability of successful replication.

The bottom row of Figure 1 shows the two replication datasets (yellow) and the two data scrapes (green). We find that for the replication studies the estimated relative risk of publication is equal to one, which means that—as expected—there is no evidence of publication bias among the replications. The two data scrapes, on the other hand, show signs of massive selection on statistical significance: while for the 11 curated datasets $\hat{\omega}$ ranges from 0.65 to 0.95, the two scraped datasets read out $\hat{\omega} = 0.08$ and $\hat{\omega} = 0.20$. However, we must be careful not to over-interpret this result. The corpus of Barnett and Wren (2019) is based on confidence intervals, and it is possible that non-significant

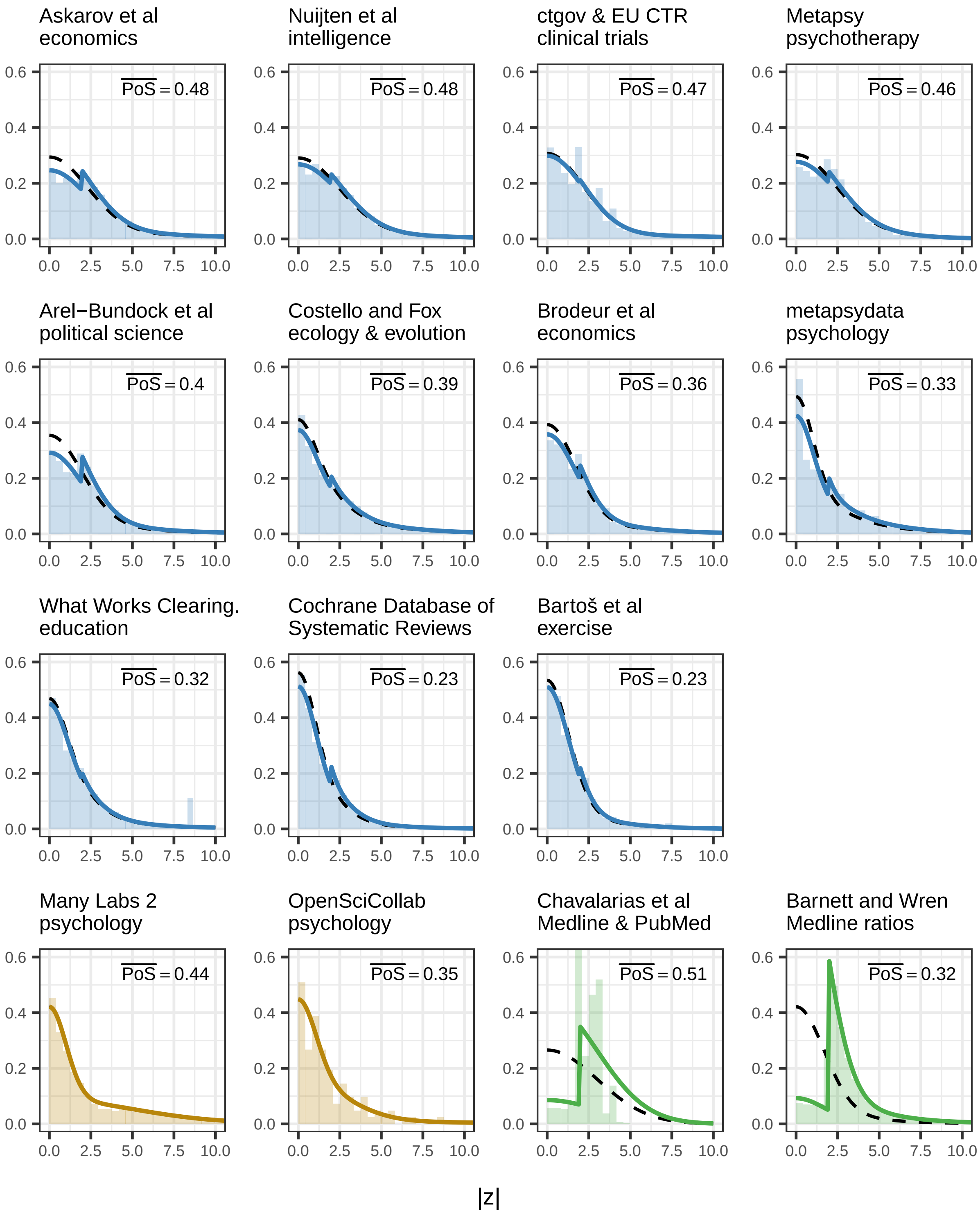

Figure 1: Histograms of absolute z-values across 15 corpora with fitted curves overlaid on top. The solid curves are the fitted mixtures of half-normal distributions conditional on publication. Dashed curves are the unconditional "corrected" distributions. Blue: curated datasets; yellow: replications; green: large-scale scraped datasets.

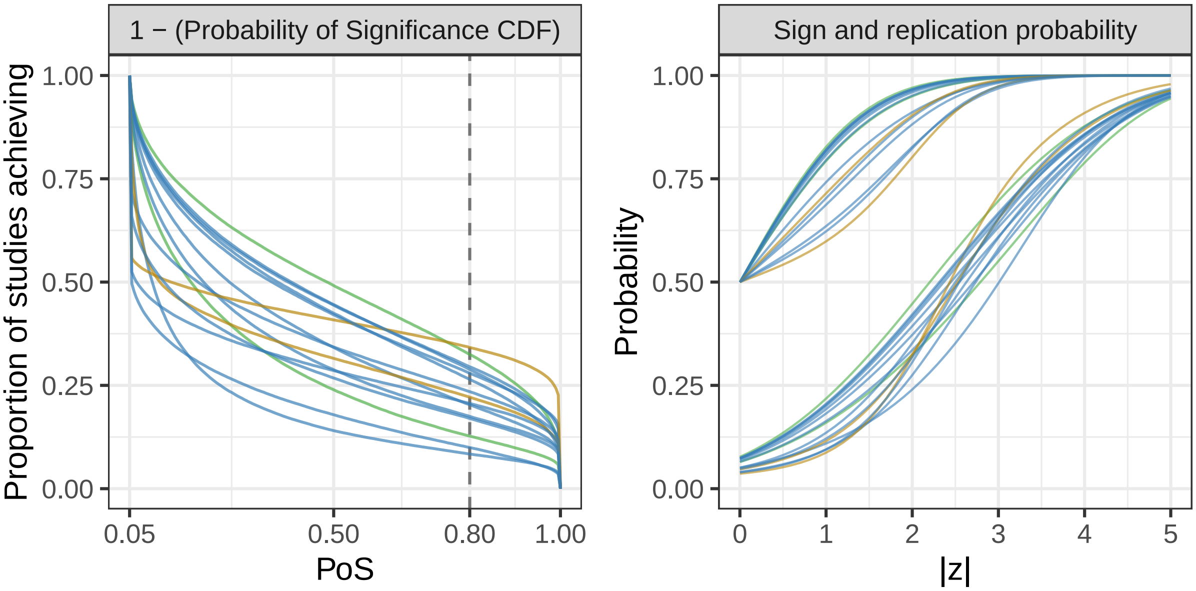

Figure 2: Left panel: "survival curves" of the probability significance: $1-(\text{CDF of PoS})$ for each of the 15 corpora. Right panel: conditional probabilities that the observed direction of the effect matches the true direction (top set of curves) and the conditional probability of "successful" replication (bottom set). Colors here are the same as in Figure 1.

results were actually reported, just not with the associated confidence intervals. The noticeable spikes in the histogram of the data from Chavalarias et al. (2016) are due to severe truncation. We do not consider our results based on these two datasets to be very reliable, but it does seem clear that the significance filter has left a mark on the scientific literature.

5. Discussion

5.1. An alternative to the binary model of scientific discovery

The 2×2 model (Ioannidis 2005) is often taken to be a reasonable, if simplified, description of scientific research. According to this model, the probabilities of significance (PoS) of the "true relationships" and "no relationships" are well separated. The PoS of "no relationships" is 5% by the design of the statistical test, and the PoS of "relationships" is much larger by the design of the experiment. If the ratio of "true relationships" to "no relationships" is not too low, then we can reliably separate them. That is, the positive predictive value (the conditional probability of a true relationship given statistical significance) and the negative predictive value (the conditional probability of no relationship given non-significance) are both high. Moreover, the conditional probability of successful replication (reaching statistical significance again in an exact replication study) is also high.

We have developed an alternative statistical model that goes beyond the 2×2 framework by allowing the PoS to range continuously from 0.05 to 1 and uses Hedges' model to adjust for selective reporting. We also allow for truncation. We fit our model to corpora from about 80,000 studies across diverse fields collected from publicly available sources.

The model fits well; see Figures 1 and S2. In contrast, the 2×2 model does not fit our data, as we observe no clear separation between the probability of statistical significance for “true relationships” and “no relationships.”

5.2. What our meta-analysis says about effect sizes and replication rates

There are four notable features that hold across all the datasets of empirical research we collected. First, the absolute z-values have a decreasing density, which implies that the absolute SNRs also have a decreasing density. In other words, studies with low SNRs are more common than those with high SNRs. We do find that there are considerable differences between research domains in the widths of these distributions which means that the studies in some areas tend to have higher SNRs than in other areas.

Second, the probability of significance (PoS) tends to be low and only a small minority of studies achieve 80% power. Third, reaching statistical significance does not imply a high probability of replication, see Figure 2 and Figure S1. Even at $p = 0.001$ ($|z| = 3.29$) replication is far from assured. Fourth, statistical significance *does* imply that the sign of the effect is likely correct.

5.3. Reasons for qualified optimism about empirical research

Instead of focusing on the paradigm of “no relationships” versus “true relationships” and associated type I and type II errors, our model allows us to examine replication and identifying the direction of effects which are both insensitive to whether effects are zero or merely small.

We find that statistical significance usually implies a high probability that the direction of the observed effect is correct, but that it does not imply a high probability that an exact replication study will be significant again. In other words, empirical research often leads to true findings—in the sense of correctly identifying the direction of an effect—that do not replicate. Therefore, failure to replicate should not be taken to mean that the original result was a fluke or a fake.

This should not come as a surprise, as the difference between “significant” and “not significant” is not itself statistically significant (Gelman and Stern 2006). And yet, it can be viewed as scandalous when a statistically significant published result does not replicate (Tversky and Kahneman 1971). This suggests a widespread discrepancy between expectations and reality which we believe to be due to the unrealistic 2 × 2 model that people have in mind.

The pessimism in much of the meta-scientific literature reflects a one-sided interpretation of Figure 2 and Figure S1 focusing only on the low replication rates. This skewed view has serious consequences as it may lead to dangerous nihilism among policy makers (Garisto 2025). We argue for a more balanced view where we also take into account the high probabilities of getting the sign right.

5.4. Understanding our findings using the anthropic principle

The current paper is meta-scientific in that it is based on a global analysis of large corpora of published results. Why do these different corpora, coming from many different experiments in different fields, show similar distributions of the SNR? We tentatively ascribe this to a sort of anthropic principle (Gelman 2018) whereby researchers respond to the pressures of time, money, and logistics on one side and the goal to reach statistical significance on the other, by making their studies just about large enough to assess the direction of the effect of interest, but not larger.

That said, these incentives do not apply in all areas, and there are subfields of science that are plagued by questionable research practices and increasingly influenced by paper mills and fraudulent research. If we have good prior reasons to doubt the study design or integrity of the analysis, we should use that information. In other words, not every result should be treated as an typical draw from the broad reference class. Our qualified optimism extends only to research for which we do not have decisive reasons to distrust.

5.5. Relevance to statistical practice

We conclude by comparing to standard practice, which is that if a result reaches a statistical significance threshold, it is treated as real and the corresponding estimate is taken at face value, while statistically non-significant results are presented as null findings or, at best, scenarios where there is not enough information to draw any conclusions.

Along with our qualified optimism about the direction of effects, our analysis suggests that a low observed rate of statistically significant replications should not be taken as evidence that most published claims are false. The bad news is that most studies have less than 80% probability of significance (power relative to the true effect size); the good news is that, even though average probabilities of significance range from 20% to 50%, replications are likely to go in the right direction, with estimated average type S error rates in our corpora in the range of 1%–3%.

6. Acknowledgments

We thank the compilers of the corpora and the authors of the original studies for their great effort and the U.S. Office of Naval Research for partial support of this work.

References

Aquarius, René, Elisabeth M. Bik, David Bimler, Morten P. Oksvold, and Kevin Patrick (2025). "Tackling paper mills requires us to prevent future contamination and clean up the past—the case of the journal Bioengineered." *Bioengineered* 16.1, p. 2542668.

Arel-Bundock, Vincent, Ryan C. Briggs, Hristos Doucouliagos, Marco Mendoza Aviña, and Tom D. Stanley (2022). "Quantitative political science research is greatly underpowered." *I4R Discussion Paper Series* 6.

Askarov, Zohid, Anthony Doucouliagos, Hristos Doucouliagos, and T. D. Stanley (2023). "The significance of data-sharing policy." *Journal of the European Economic Association* 21.3, pp. 1191–1226.

Bak-Coleman, Joseph, Richard Mann, Carl Bergstrom, Kevin Gross, and Jevin West (2022). "Revisiting the replication crisis without false positives." *Center for Open Science*. https://osf.io/preprints/socarxiv/rkyf7_v1.

Barnett, Adrian Gerard and Jonathan D. Wren (2019). "Examination of CIs in health and medical journals from 1976 to 2019: an observational study." *BMJ Open* 9.11, e032506.

Bartoš, František, Michaela Lušková, Katerina Bortnikova, Karolína Hozová, Kristína Kantova, Zuzana Irsova, and Tomas Havranek (2025). "Effect of exercise on cognition, memory, and executive function: A study-level meta-meta-analysis across populations and exercise categories." *PsyArXiv preprint*. DOI: 10.31234/osf.io/qr8e2_v1.

Brodeur, Abel, Nikolai M. Cook, Jonathan S. Hartley, and Anthony Heyes (2024). "Do preregistration and preanalysis plans reduce p-hacking and publication bias? Evidence from 15,992 test statistics and suggestions for improvement." *Journal of Political Economy Microeconomics* 2.3, pp. 527–561.

Camerer, Colin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, et al. (2016). "Evaluating replicability of laboratory experiments in economics." *Science* 351.6280, pp. 1433–1436.

Chavalarias, David, Joshua David Wallach, Alvin Ho Ting Li, and John P. A. Ioannidis (2016). "Evolution of reporting P values in the biomedical literature, 1990-2015." *Journal of the American Medical Association* 315.11, pp. 1141–1148.

Cochrane Collaboration (2025). *Cochrane Database of Systematic Reviews.* URL: `https://www.cochranelibrary.com/cdsr`.

Costello, Laura and Jeremy W. Fox (2022). "Decline effects are rare in ecology." *Ecology* 103.6, e3680.

Cuijpers, Pim, Mathias Harrer, Clara Miguel, et al. (2025). "Cognitive behavior therapy for mental disorders in adults: A unified series of meta-analyses." *JAMA Psychiatry* 82.6, pp. 563–571.

Efron, Bradley (2016). "Empirical Bayes deconvolution estimates." *Biometrika* 103.1, pp. 1–20.

European Medicines Agency (2025). *EU Clinical Trials Register.* URL: `https://www.clinicaltrialsregister.eu/`.

Garisto, Dan (2025). "Trump order gives political appointees vast powers over research grants." *Nature* 644.8077, pp. 585–586.

Gelman, Andrew (2018). "The anthropic principle in statistics." *Statistical Modeling, Causal Inference, and Social Science.* `https://statmodeling.stat.columbia.edu/2018/05/23/anthropic-principle-statistics/`.

Gelman, Andrew and John Carlin (2014). "Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors." *Perspectives on Psychological Science* 9.6, pp. 641–651.

Gelman, Andrew and Hal Stern (2006). "The difference between "significant" and "not significant" is not itself statistically significant." *American Statistician* 60.4, pp. 328–331.

Gelman, Andrew and Francis Tuerlinckx (2000). "Type S error rates for classical and Bayesian single and multiple comparison procedures." *Computational Statistics* 15.3, pp. 373–390.

Greenland, Sander (2017). "The need for cognitive science in methodology." *American Journal of Epidemiology* 186.6, pp. 639–645.

Head, Megan L., Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions (2015). "The extent and consequences of p-hacking in science." *PLoS Biology* 13.3, e1002106.

Hedges, Larry V. (1984). "Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences." *Journal of Educational Statistics* 9.1, pp. 61–85.

Hedges, Larry V. (1992). "Modeling publication selection effects in meta-analysis." *Statistical Science* 7.2, pp. 246–255.

Institute of Education Sciences (2025). *What Works Clearinghouse Study Findings.* URL: `https://ies.ed.gov/ncee/wwc/studyfindings`.

Ioannidis, John P. A. (2005). "Why most published research findings are false." *PLoS Medicine* 2.8, e124.

Ioannidis, John P. A., T. D. Stanley, and Hristos Doucouliagos (2017). "The power of bias in economics research." *Economic Journal* 127, F236–F265.

Jager, Leah R. and Jeffrey T. Leek (2014). "An estimate of the science-wise false discovery rate and application to the top medical literature." *Biostatistics* 15.1, pp. 1–12.

Klein, Richard A., Michelangelo Vianello, Fred Hasselman, Byron G. Adams, Regan B. Adams, Sinan Alper, Mark Aveyard, Jordan R. Axt, Mayowa T. Babalola, et al. (2018). "Many Labs 2: Investigating variation in replicability across samples and settings." *Advances in Methods and Practices in Psychological Science* 1.4, pp. 443–490.

Matusz, Pawel J., Anna Abalkina, and Dorothy V. M. Bishop (2025). "The threat of paper mills to biomedical and social science journals: The case of the Tanu.pro paper mill in *Mind, Brain, and Education*." *Mind, Brain, and Education* 19.2, pp. 90–100.

Metapssy (2026). *Meta-Analytic Psychotherapy Databases.* https://www.metapsy.org/.

National Library of Medicine (2025). *ClinicalTrials.gov.* URL: `https://clinicaltrials.gov/`.

Neves, Kleber, Pedro B. Tan, and Olavo B. Amaral (2022). "Are most published research findings false in a continuous universe?" *Plos One* 17.12, e0277935.

Nuijten, Michèle B., Marcel A. L. M. Van Assen, Hilde E. M. Augusteijn, Elise A. V. Crompvoets, and Jelte M. Wicherts (2020). "Effect sizes, power, and biases in intelligence research: A meta-meta-analysis." *Journal of Intelligence* 8.4, p. 36.

Open Science Collaboration (2015). "Estimating the reproducibility of psychological science." *Science* 349.6251, aac4716.

Rodriguez, Josue E. and Donald R. Williams (2022). "psymetadata: An R package containing open datasets from meta-analyses in psychology." *Journal of Open Psychology Data*. DOI: 10.5334/jopd.61.

Schwab, Simon (2024). *cochrane: Import data from the Cochrane database of systematic reviews (CDSR)*. R package. URL: https://github.com/schw4b/cochrane.

Senn, Stephen (2002). *Cross-over Trials in Clinical Research*. Wiley.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn (2011). "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant." *Psychological Science* 22.11, pp. 1359–1366.

Singh, B. et al. (2025). "Effectiveness of exercise for improving cognition, memory and executive function: A systematic umbrella review and meta-meta-analysis." *British Journal of Sports Medicine* 59.12, pp. 866–876. DOI: 10.1136/bjsports-2024-108589.

Stephens, Matthew (2017). "False discovery rates: a new deal." *Biostatistics* 18.2, pp. 275–294.

Thomas, David, Daniel Chancellor, Amanda Micklus, Sara LaFever, Michael Hay, Shomesh Chaudhuri, Robert Bowden, and Andrew W. Lo (2021). *Clinical Development Success Rates and Contributing Factors, 2011–2020*. Report. Biotechnology Innovation Organization (BIO). URL: https://go.bio.org/rs/490-EHZ-999/images/ClinicalDevelopmentSuccessRates2011_2020.pdf.

Tversky, Amos and Daniel Kahneman (1971). "Belief in the law of small numbers." *Psychological Bulletin* 76, pp. 105–110.

van Zwet, Erik and Andrew Gelman (2022). "A proposal for informative default priors scaled by the standard error of estimates." *American Statistician* 76.1, pp. 1–9.

van Zwet, Erik and Steven N. Goodman (2022). "How large should the next study be? Predictive power and sample size requirements for replication studies." *Statistics in Medicine* 41.16, pp. 3090–3101.

van Zwet, Erik, Simon Schwab, and Stephen Senn (2021). "The statistical properties of RCTs and a proposal for shrinkage." *Statistics in Medicine* 40.27, pp. 6107–6117.

van Zwet, Erik, Witold Więcek, and Andrew Gelman (2025). "Meta-analysis with a single study." *Statistical Methods in Medical Research*.

Yang, Yefeng, Alfredo Sánchez-Tójar, Rose E. O'Dea, Daniel W. A. Noble, Julia Koricheva, Michael D. Jennions, Timothy H. Parker, Malgorzata Lagisz, and Shinichi Nakagawa (2023). "Publication bias impacts on effect size, statistical power, and magnitude (type M) and sign (type S) errors in ecology and evolutionary biology." *BMC Biology* 21.1, p. 71.

Yang, Yefeng, Erik van Zwet, Nikolaos Ignatiadis, and Shinichi Nakagawa (2024). "A large-scale in silico replication of ecological and evolutionary studies." *Nature Ecology & Evolution*, pp. 1–5.

A. Details of calculations

Suppose the marginal distribution of the z-value is a mixture of zero-mean normal distributions,

$$f(z \mid p, \sigma) = \sum_{i=1}^{k} p_i \, \frac{1}{\sigma_i} \, \phi\left(\frac{z}{\sigma_i}\right).$$

Suppose also that the conditional distribution of the z-value, given the SNR, is normal with mean SNR and variance 1. Together, the marginal and conditional distributions determine the joint distribution of the z-value and the SNR. Therefore, we also have the conditional distribution of the SNR given the z-value. It is a mixture of normal distributions with weights $q_i(z)$, means $\mu_i(z)$, standard deviations τ_i given by

$$q_i(z) = \frac{p_i \phi(z/\sigma_i)/\sigma_i}{f(z \mid p, \sigma)}, \quad \mu_i(z) = \frac{\sigma_i^2 - 1}{\sigma_i^2} \, z \quad \text{and} \quad \tau_i = \frac{\sqrt{\sigma_i^2 - 1}}{\sigma_i}. \tag{12}$$

The conditional probability of the correct sign is

$$\Pr(z * \mathrm{SNR} > 0 \mid |z|) = \sum_{i=1}^{k} q_i(z) \Phi\left(\frac{\mu_i(|z|)}{\tau_i}\right). \tag{13}$$

If z_{repl} is the z-value of an exact replication study, then the conditional distribution of z_{repl} given the z-value of the original study is a mixture of normal distributions with weights $q_i(z)$, means $\mu_i(z)$ and variances $\tau_i^2 + 1$. The conditional probability of "successful replication" is

$$\Pr(z_{\mathrm{repl}} * z > 0 \text{ and } |z_{\mathrm{repl}}| > 1.96 \mid |z|) = \sum_{i=1}^{k} q_i(z) \Phi\left(\frac{\mu_i(|z|) - 1.96}{\sqrt{\tau_i^2 + 1}}\right). \tag{14}$$

		Significance		PoS at least 80%		
Corpus	$\hat{\omega}$	signif.	$\overline{PoS}$	uncond.	$\lvert z\rvert = 1.96$	$\lvert z\rvert \geq 1.96$
Cochrane	0.70	0.31	0.23	0.10	0.05	0.40
ctgov / EU CTR	0.95	0.52	0.47	0.28	0.09	0.58
Costello and Fox	0.79	0.42	0.39	0.23	0.08	0.58
Arel-Bundock et al	0.65	0.47	0.40	0.20	0.07	0.48
Bartos et al	0.82	0.26	0.22	0.08	0.03	0.35
Askarov et al	0.71	0.52	0.48	0.30	0.10	0.59
Nuijten et al	0.85	0.53	0.48	0.29	0.11	0.57
Metapsy	0.83	0.48	0.46	0.26	0.10	0.54
What Works	0.88	0.34	0.32	0.17	0.06	0.51
psymetadata	0.67	0.36	0.33	0.21	0.08	0.62
Brodeur et al	0.79	0.37	0.36	0.17	0.05	0.46
Many Labs 2	0.99	0.44	0.44	0.35	0.11	0.78
OpenSciCollab	1.00	0.35	0.36	0.22	0.09	0.60
Chavalarias et al	0.20	0.84	0.51	0.32	0.13	0.60
Barnett and Wren	0.08	0.83	0.32	0.12	0.04	0.37

Table S2: Summary of signal-to-noise ratio modeling results for the 15 corpora of studies in our data. Calculations of power, replication, and direction of effects use the distributions of z-values and signal-to-noise ratios by fitted mixture models without selection.

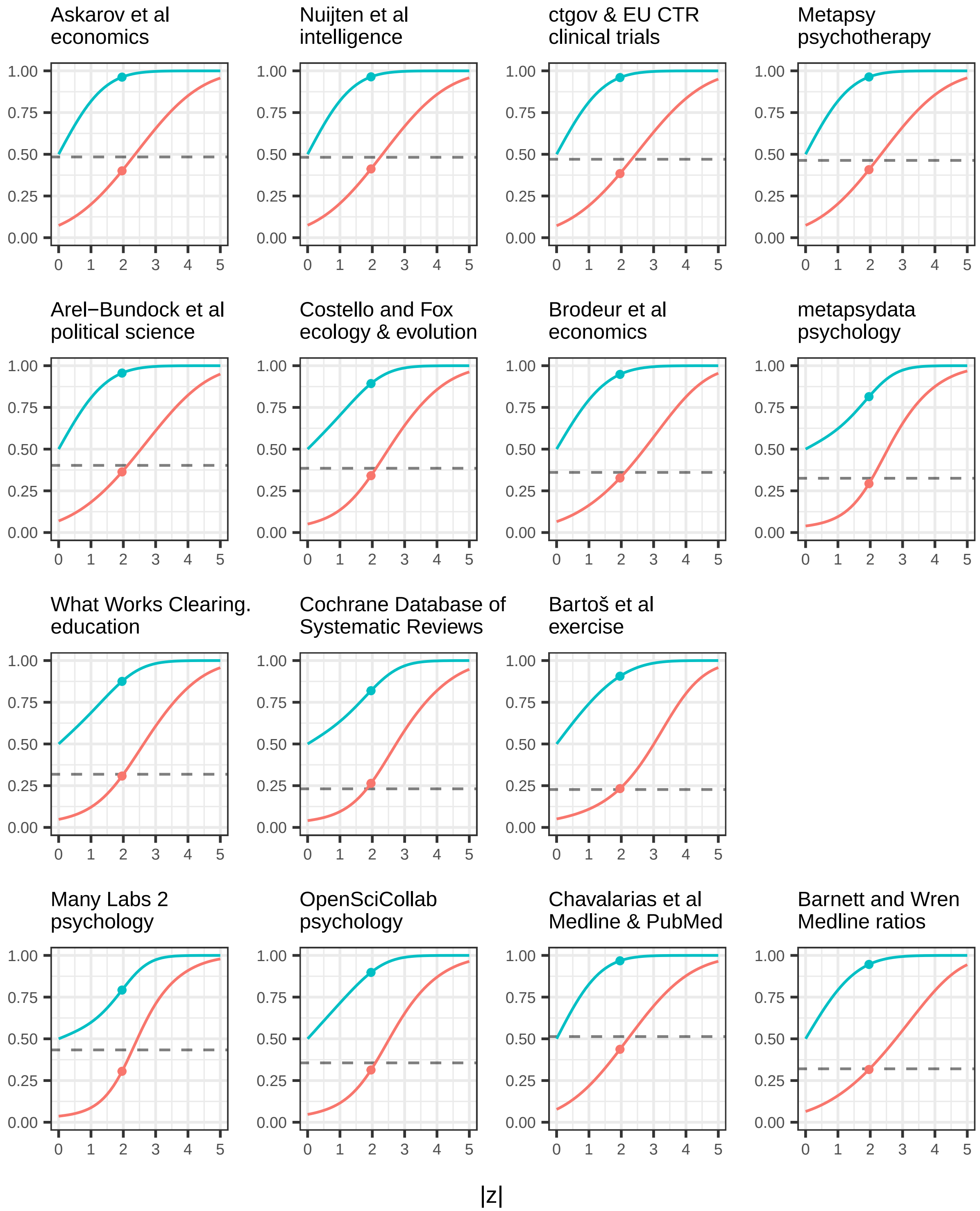

Figure S1: Relationship between absolute z-values and the probability that the direction of the observed effect matches the true direction (blue curves) and the probability of "successful replication" (red curves). The vertical line is $\overline{\text{PoS}}$.

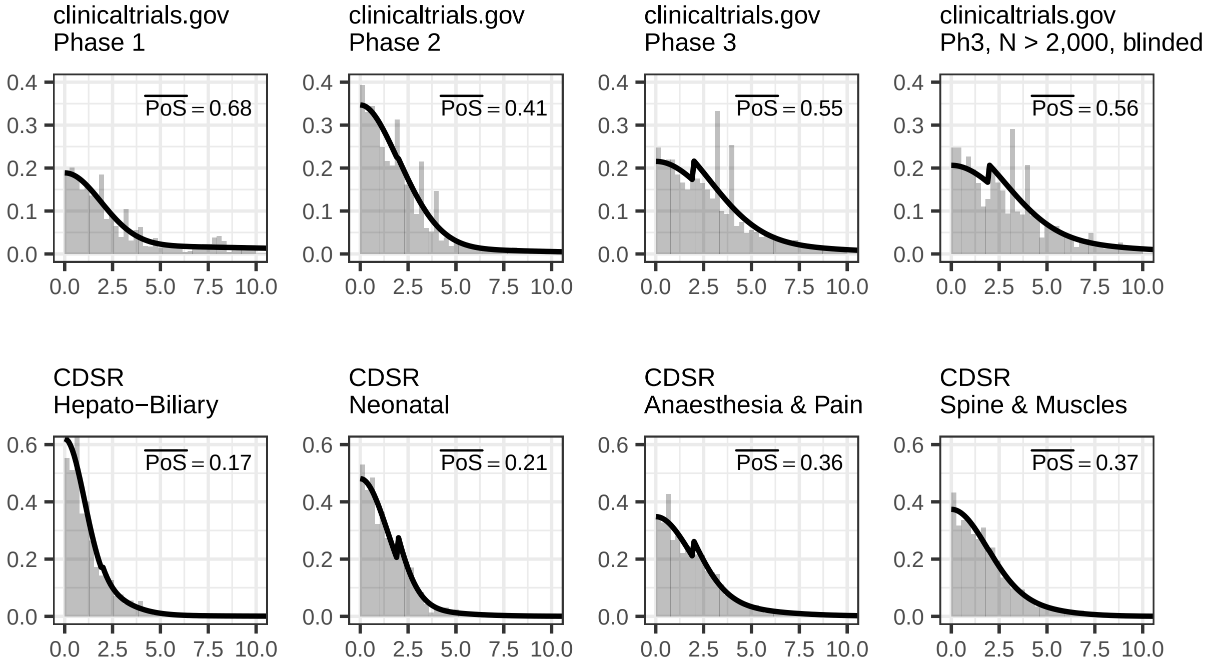

Figure S2: Subsets of interventional treatment studies from `clinicaltrials.gov`, broken down by phase (top row) and from Cochrane database (bottom row), grouped by specialty (out of 19 included in the dataset): two with lowest $\overline{\text{PoS}}$ on the left and two with the highest on the right. The peaks in the histograms at $|z| = 1.96, 2.58, 3.29$ are due to trials reporting (two-sided) $p < 0.05, p < 0.01, p < 0.001$.

B. Datasets included in BEAR

B.1. Askarov et al. (2023): economics

Research question: Impact of mandatory data sharing on (excessive) statistical significance in economics papers

Data availability: file `Mandatory data-sharing 30 Aug 2022.dta` may be downloaded from `https://github.com/anthonydouc/Datasharing/blob/master/Stata/`; reproducibility package at `https://dataverse.harvard.edu/file.xhtml?fileId=7884702&version=1.0`

Data description and source: The paper includes 64,000 economic estimates surveyed by Ioannidis, Stanley, and Doucouliagos (2017) plus results of "searching through numerous databases, journal websites, and also an email survey of 109 authors known to have produced meta-analyses of economic topics. We cast the widest net possible to capture empirical economic estimates. Thus, some of the papers included in these meta-analyses were originally published in political science and psychology journals. Nevertheless, these meta-studies deal with economics topics." Later they note: "we collect 166,924 parameter estimates from 345 distinct research areas, of which 20,121 were published in 24 of the leading general interest and field journals" The paper mentions 12,521 observations before and 2,537 after data sharing policies in 24 journals; reproducibility dataset has 32 journals (authors removed some smaller ones) and 22,172 rows.

Data processing: We use the same file `Mandatory data-sharing 30 Aug 2022.dta` (11.6 MB) as in the reproducibility package. No additional manipulation of data was done when creating BEAR dataset.

B.2. Rodriguez and Williams (2022), Klein et al. (2018), and Nuijten et al. (2020): psychology

Snapshot date: Jan 2026

Research question: *psymetadata* is an R package that curates multiple open datasets from published meta-analyses in psychology for teaching and demonstrations. Klein et al. (2018) study how replication results vary across labs and settings for a set of classic and contemporary effects in psychology. Nuijten et al. (2020) is a large meta-meta-analysis of intelligence research, which focused on power and small study effects.

Data availability: 22 datasets are distributed with the `psymetadata` R package (CRAN); after disaggregating Many Labs 2 and intelligence meta-meta-analysis, remaining 20 datasets vary from 65 rows of data to over 1,000. The package is GPL-3; individual datasets originate from the cited source papers.

Data processing: minimal. We used regular expressions to extract year labels from titles. We dropped some invalid rows (missing, non-finite, or non-positive standard errors). For Many Labs we can group by experiment—we use these as `metaid` column.

Additional grouping variable: name of the original psymetadata dataset.

B.3. European Medicines Agency (2025): EU Clinical Trials Register

Snapshot date: 18 January 2026

Data collection: we used the R package `ctrdata` (Herold 2025) to create a snapshot of EU CTR (`collection="euctr"`); we downloaded trial records with results only. We stored the downloaded records in a local SQLite database and extracted a fixed set of protocol and results fields plus several fields computed by `ctrdata` (trial phase, sample size, and a "first primary endpoint" p-value/size derived by the package). We saved the extracted flat dataset as `data_euctr_ctgov.rds`, but we only use the EUCTR portion of this combined dataset in BEAR.

Data availability: EUCTR is publicly accessible. The EUCTR legal notice notes that publication on EUCTR does not constitute an endorsement of the information reported.

Data processing: extensive. We use only primary endpoints. We defaulted missing CI levels to 95% and where truncation of p-values was not stated we assumed equality (the latter was the case for about 15% of data). Rest of the derivation of z-values followed the *Standard procedure for dealing with p-values and confidence intervals* described below. We set `year` to record *entry* date (since "completion date" field was not accessible), dropped rows only when z was not finite, and harmonized phase labels to follow the same naming convention as ClinicalTrials.gov.

Additional variables used: we retained phase labels, a measure class (derived from the estimand label), primary endpoint sample size

B.4. Arel-Bundock et al. (2022): political science

Research question: assess statistical power in political science research

Data availability: replication package (zip file with date 20241010) can be found at `https://osf.io/fgdet`. All data are in public domain. We notified one of the authors, Ryan Briggs, who was not aware of any licensing issues.

Data description and source: Authors searched for (and emailed authors of) meta-analysis articles across 141 journals in political science; this led to "a dataset of 16,649 hypothesis tests, grouped in 351 meta-analyses, reported in 46 peer-reviewed meta-analytic articles."

Notes: There is possibly a considerable overlap with Askarov et al. (2023) dataset, but these two datasets are worth analyzing separately given their different focus.

Data processing: using the same code as original authors (replication package runs on makefiles), but skipping some additional processing that they've done, we write a single `estimates.csv` file (3.3 MB). No manipulation of data was done when creating the BEAR dataset.

B.5. Costello and Fox (2022), Yang, Sánchez-Tójar, et al. (2023), and Yang, van Zwet, et al. (2024): ecology and evolution

These two datasets come from the same replication package and shared data processing pipeline.

Research question: decline effects in ecology Costello and Fox (2022) and publication bias and performance of empirical research

Data collection: Costello and Fox (2022)looked at 466 meta-analyses obtained through searching Web of Science. Reproducibility dataset includes 88,000 rows among 466 analyses in 232 papers. Subsequently, Yang, Sánchez-Tójar, et al. (2023) removed duplicated studies and zeroes, which meant discarding about 20% of rows. Yang, van Zwet, et al. (2024) collected 102 meta-analyses published 2010–2019 in relevant journals, de-duplicated against the Costello and Fox (2022) dataset.

Data availability: two datasets may be downloaded from `https://github.com/Yefeng0920/replication_EcoEvo_git/`. Available under CC BY 4.0 licence.

Data processing: no additional data processing other than cleaning study years. The replication package for Yang, Sánchez-Tójar, et al. (2023) does not include meta-analysis indicators.

Additional variables used: we retained a measure variable ("lnRR," "SMD," "Zr," or "uncommon")

B.6. Brodeur et al. (2024): economics

Research question: are preregistration and pre-analysis plans associated with reduced p-hacking and publication bias in economics?

Data collection: manually extracted test statistics from RCTs in leading economics journals published 2018-2021. Extraction was performed table by table. Each row corresponds to a single reported test statistic, linked to a paper-level identifier. Unlike most other sources in BEAR, this means we have dozens of estimates per paper: 314 articles with 16,390 estimates

Data availability: replication data file `merged.dta` may be downloaded from `https://dataverse.harvard.edu/file.xhtml?fileId=7884702&version=1.0`

Data processing: the authors' reported z-statistics were used directly and the reported coefficients and standard errors were retained. No transformation of test statistics or data was required.

B.7. Metapssy (2026): psychotherapy

A set of living meta-analytic databases of randomized trials of psychotherapeutic interventions.

Snapshot date: Jan 2026

Data availability: Metapsy is maintained by an international collaboration led by Vrije Universiteit Amsterdam. Databases are available at `https://www.metapsy.org/` and via R package `metapsyData` (Harrer, Karyotaki, and Cuijpers 2022). Metapsy FAQ states that all data provided as part of the project are open-access and requires attribution.

Data processing: we merged 20 individual datasets in Metapsy with some cleaning (e.g., to extract study dates from titles). Almost all estimates are Hedges' g, but for one of the databases (`total-response`) we converted log-odds ratios to SMD (dividing by $\Pi/\sqrt{3}$ to scale SD of logistic regression).

B.8. Institute of Education Sciences (2025): What Works Clearinghouse

Snapshot date: 17 May 2025

Data availability: flat files are available at `https://ies.ed.gov/ncee/wwc/studyfindings`. Per permissions and disclaimers section on that website: "Unless stated otherwise, all information on the U.S. Department of Education's IES website . . . is in the public domain and may be reproduced, published, linked to, or otherwise used without permission from IES."

Data collection: WWC datasets have 13,054 findings (data-rows) across 1,908 reviews. The dataset includes effect sizes, but not standard errors, as well as p-values reported by the studies and p-values re-calculated by the WWC reviewers.

Data processing: We retained only RCTs and quasi-experimental studies (98% of data altogether). Where available, we preferred effect sizes and p-values calculated by WWC to the ones reported by the studies. We replaced p-values that were equal to zero (this occurred in 3% of the dataset) with truncated p-value (i.e. assumed $p < 10^{-16}$). We calculated z-values from these p-values under assumption of two-sided tests.

Additional grouping variable: "outcome domain" variable (e.g., "academic achievement," "alphabetics")

B.9. Cochrane Collaboration (2025)

Snapshot date: 20 November 2025

Data collection: using the R package `cochrane` (Schwab 2024), we merged data from 8,726 records of reviews in CDSR and merged them into a single file `cdsr_interventions_19nov2025.csv` (63 MB).

Data processing: minimal processing with extensive filtering. Our processed dataset has 39,768 rows, compared to over 800,000 rows of data in the unprocessed dataset.

On processing side, we cleaned up study years, categorized measures (risk ratio, odds ratio, mean difference etc.) and experimental designs, especially an "RCT" flag (based on scanning of abstracts of each review for inclusion criteria). Most studies in CDSR are RCTs but some reviews also allowed quasi-experimental studies, which prevented us from categorizing many studies.

Unlike in most of the other datasets in BEAR, we filter CDSR data heavily. First, we only use the outcome and comparison of each review that are coded as "1," as that is most likely the primary outcome and most relevant comparison. This removes over 90% of all rows of data. Secondly, we only use studies with continuous and dichotomous outcomes, removing about 10% of data where effect estimates are based on instrumental variables or based on individual patient data. Lastly, we remove about 5% of data where measure of effect is unknown (retaining: OR, RR, Peto OR, mean difference, standardized mean difference, and risk difference). We also remove some rows where there are zero subjects.

Additional grouping variable: we retained a "source data type" variable to distinguish estimates from published, unpublished, "sought," and mixed data sources.

B.10. Bartoš et al. (2025) and singh2025exerciseBartoÅą et al.~2025:)

Research question: What is the effect of physical exercise on cognition, memory, and executive function; with extra focus on selective reporting and heterogeneity.

Data availability: Replication data are publicly available via PsyArXiv `https://osf.io/preprints/psyarxiv/qr8e2_v1` under Creative Commons By Attribution 4.0 license.

Data description and source: "2,239 effect-size estimates from 215 meta-analyses of randomized controlled trials" was done by Bartoš et al. (2025) based on extension of work by Singh et al. (2025).

Data processing: z-values were computed as effect size divided by its reported standard error. No filtering or recoding beyond variable renaming was required.

Additional variables used: None.

B.11. Open Science Collaboration (2015)

Research question: replications of a quasi-random sample of 100 experiments in psychology.

Data collection: authors of the paper make a comprehensive dataset of results available with their paper. In BEAR we did not use the original studies (since all studies in that set have $p < 0.05$), only replications.

Data availability: the data file `rpp_data.csv` is available as part of repository for the paper `https://osf.io/ytpuq/overview` under CC0 1.0 Universal license.

Data processing: we derived z values from the replication p-values assuming two-sided p-values (we use the replication p-value column, not the original-study p-value). Because the replication dataset does not carry direction in a way we use here, z values are unsigned. We treated p-values recorded as exactly 0 as truncated and recorded this as a lower bound on z (z-operator `>`); otherwise we treated p-values as exact (z-operator `=`). We retained the replication sample size as `ss`.

B.12. Chavalarias et al. (2016): Medline/Pubmed

Research question: reporting of p-values in the biomedical literature

Data collection: the authors used large-scale text mining of MEDLINE abstracts and PubMed full texts: 4.5 million p-values in 1.6 million MEDLINE abstracts; 3.4 million p-values in 385,000 PubMed full-text articles.

Data availability: the extracted p-value dataset is publicly available `https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/6FMTT3` (about 6GB of data) As per the hosting platform: this dataset is released under the Creative Commons Attribution-NonCommercial 4.0 International License. We extracted the relevant columns into `chavalarias.rds` dataset (25 MB).

Data processing: z-values were derived from p-values assuming two-sided tests. P-value truncation operators were processed by collapsing variants such as `<<`, `<<<`, `<=`, `less than`, and `=<` into `<`. We dropped 0.7% of rows where p-values did not have a "plain" format and 0.08% of rows where truncation could not be unambiguously classified.

Additional variables used: source indicator (abstract versus full text).

B.13. Barnett and Wren (2019)

Research question: Bias for statistical significance in health and medical journals

Data availability: The dataset Georgescu.Wren.RData may be downloaded from `https://github.com/agbarnett/intervals/tree/master/data` and also `https://github.com/jdwren/ASEC`. Data file is `Georgescu.Wren.RData`. Availability: no clear license given in the repo but the article is CC BY 4.0 in BMJ Open.

Data description and source: The dataset is a collection of confidence intervals for ratio estimates from Medline papers from 1976-2019. Authors scraped (via regular expressions, followed by an independent check using another data mining algorithm, manual checks for 10,000 intervals) 968,000 CIs from abstracts and 350,000 from full texts.

Notes: These are "ratio estimates" like odds ratios, hazard ratios and risk ratios. Binary outcomes tend to have much lower information than continuous outcomes.

Data processing: We excluded a small fraction of data. First we restricted to 95% confidence intervals only: where `ci.level` was missing we assumed it was 0.95; among intervals with a known `ci.level`, about 0.3% were not 95% and were dropped. We dropped intervals with non-positive width (i.e. `lower >= upper`). We used log scale (we replaced zero or negative lower bounds with a small positive constant; this affected about 0.1% of the sample) and backed out an approximate standard error and point estimate under the usual normal approximation for a 95% CI, setting point estimate to half-point of the interval.

B.14. clinicaltrials.gov snapshot

Trials which reported results at clinicaltrials.gov (National Library of Medicine 2025)

Snapshot date: 12 August 2025

Data source: This is a set of flat file of studies obtained via `https://aact.ctti-clinicaltrials.org/` (over 2GBs of data for relevant tables) with relevant columns extracted and joined by us to produce a single table `clinicaltrials.gov_aug2025.rds` (12 MB of data).

Data availability: terms and conditions for ClinicalTrials.gov are at `https://clinicaltrials.gov/about-site/terms-conditions`. We make distributions of derived effect sizes available in accordance with ToC's. We only use publicly available data and do not imply any endorsement of this work by National Library of Medicine, ClinicalTrials.gov or AACT.

Data processing: extensive. We kept only studies where status was "Completed" and only rows of data where outcome type was "Primary." We set the year to the year of `completion_date`.

Both p-values and confidence intervals were reported. See *Standard procedure for dealing with p-values and confidence intervals* below for how we created and chose the z-values. Briefly, we calculated z-values using two separate procedures (based on p-values and confidence intervals) and, if both were available, we made a judgment on which was more appropriate using a pre-defined rule. We dropped rows only when `z` was missing. Notably, infinite z-values can arise when `p_value` is recorded as exactly 0, and these are retained because they are not missing.

The result of our data processing is `data_cut_with_z.rds`, 2.1MB file.

B.15. Head et al. (2015): PubMed papers)

Research question: extent of p-hacking and its impact on meta-analyses.

Data collection: PubMed papers that are open access, up to 2014. Authors used regular expressions to extract p-values from text of Abstract and Results, but not tables. Downloaded data has about 220,000 studies and about 2 million rows.

Data availability: dataset available at `https://datadryad.org/dataset/doi:10.5061/dryad.79d43` under CC0 1.0 Universal licence.

Data file: `head.rds` (derived from the Dryad files `p.values.csv` and `journal.categories.csv`, with an additional DOI-to-PMID matching table created by us).

Data processing: we followed the same minimal clean-up steps (e.g., removing values found in large supplementary sections) as the original paper and also attached PubMed IDs to the dataset using DOI look-up. We saved a reduced analysis file `head.rds` (9 MB).

When constructing the BEAR dataset, we treated p-values recorded as exactly 0 as truncated and set the z-operator to `>`. We recoded all inequalities to sharp inequalities.

Additional variables used: we use the "Abstract vs Results" variable for grouping

B.16. Jager and Leek (2014): PubMed abstracts

Research question: estimating science-wise false discovery rate.

Data collection: authors used a custom program to extract p-values from scraped PubMed abstracts for papers published in 5 main medical journals 2000-2010. 15,653 p-values are available in 5,322 articles.

Data availability: the data file `pvalueData.rda` is available `https://github.com/jtleek/swfdr` License for the programs in that repository is GNU GPL, although a license for the dataset is not stated, as far as we are aware.

Data processing: there was only minimal processing. We derived z values from p-values assuming they were two-sided.

Notably, large proportion of p-values is truncated, almost always at 0.0001, 0.001, 0.01, or 0.05. As in all datasets, we retained information on truncation. We also treated p-values recorded as exactly 0 as truncated (z-operator `>`). We created a crude flag for RCTs by searching the paper titles for "randomised," "randomized," and "controlled."

C. Standard procedure for dealing with p-values and confidence intervals

Most datasets come with pre-computed effect sizes and standard errors. Some (especially databases of clinical trials and large sets of scraped studies) report p-values. Reporting in clinical trial database is done by investigators and data entry fields are not standardized, requiring more extensive clean-up and a procedure for deriving z-values.

C.1. Cleaning data

For confidence interval levels, we clean them up by setting values above 100 or below 0 to missing, and treating values below 1 as proportions rather than percentages by multiplying by 100. This means that, for example, 0.95 is treated as a 95% confidence interval. For missing values in the EU CTR database, we assume 95%.

In clinical trial datasets we classify the effect measure using text matching to a small set of classes (mean difference, odds ratio, risk ratio, hazard ratio, geometric ratio, risk difference, difference in percentages, ratio/other ratio, and other). This string-based mapping is an arbitrary choice. Unstandardized data in these datasets have many hundreds of values.

C.2. Confidence intervals

When working with CIs, ratio measures were analyzed on the log scale (if and only if lower bound, upper bound, and point estimate were all strictly positive) and non-ratio measures were always kept on the raw scale.

For z-statistics, we computed the critical value using stated CI percent and the reported number of sides (if missing, we defaulted to a two-sided interval). To avoid undefined critical values (e.g., when at 100% or extremely close to it), we bounded the implied α at a small value. We then computed two one-sided standard error estimates from the upper and lower bounds and used their average when both were finite.

We computed a simple symmetry diagnostic for the confidence interval on the chosen scale, defined as the smaller of the two half-widths divided by the larger. We treated intervals as "symmetric enough" only if this ratio exceeded 0.8. This threshold is arbitrary and affects when we trust the CI-derived z-statistic.

We also flagged potentially non-Wald intervals using text matching on type of parameter column, if available, searching for keywords such as median, Hodges, posterior/Bayes, exact, Fieller, bootstrap, and permutation. Any row flagged this way was treated as unreliable for CI-based inference, regardless of its numeric properties.

C.3. p-values

The default approach, used for several datasets which report p-values, especially those scraping data from PubMed/Medline, is to treat p-values as two-sided (`z <- qnorm(1 - p/2)`).

For clinical trials databases (clinicaltrials.gov and EU CTR) we sometimes found ambiguity in how p-values were recorded. Sometimes significant results report p-values close to 1, suggesting that authors reported cumulative distribution, $\Phi(z)$. Therefore for each row we also calculated $2\min(p, 1-p)$. When a CI-derived z-statistic was available, we chose whichever p-value led to a z-value that was closest to the CI-derived z.

To avoid dropping rows, we retained unsigned z-statistics when the sign could not be determined. If the sign of effect was known, we applied it to the "p-derived" z; otherwise we used the absolute value.

C.4. Choice of a single z-value when multiple candidates available

In situations where z can be derived from both CIs and p-values. We choose a single "best" z-statistic per row as follows. We prefer the CI-derived z when it looked like a plausible Wald interval on the chosen scale, meaning: (1) finite z and standard error, (2) positive standard error, (3) not flagged as non-Wald by keyword search in measure label, and (4) either missing symmetry information or symmetry above 0.8. Otherwise we used the p-derived z.