Reliable inference: benefits of open raw data may be universal in meta-analysis

Danijela Žanko[1], Sunčana Geček[1], Azra Tafro[2], Livia Puljak[3], Antica Culina[1,*]

[1]Laboratory for Informatics and Environmental Modelling, Division for Marine and Environmental Research, Ruđer Bošković Institute, Bijenička Cesta 54, Zagreb 10000, Croatia
[2]Faculty of Forestry and Wood Technology, University of Zagreb, Svetošimunska cesta 23, Zagreb 10000, Croatia
[3]Center for Evidence-Based Medicine, Catholic University of Croatia, Ilica 244, Zagreb 10000, Croatia
*Corresponding author: Antica Culina

Keywords: meta-analysis, publication bias, raw data, simulation, effect size estimation, open science

Abstract

While the benefits of open data are often discussed, they are rarely quantified. Here, we provide the first evidence of the potential gains from using raw data for evidence synthesis and introduce a tool that helps researchers determine when this approach is most beneficial. Classical meta-analysis (CMA) relies on published results, making it vulnerable to publication bias and p-hacking. We developed a simulation framework comparing CMA with raw data meta-analysis (RDMA) across varying true effect sizes, heterogeneity, and bias levels. Using ecology as a case study, we demonstrate that RDMA outperforms CMA in most scenarios. When true effects are small and bias is severe, RDMA reduces relative mean absolute error by 56–76%. Under moderate bias, reductions reach 31–62%. For medium true effects, reductions were 50–71% and 3–38%, respectively. RDMA maintained reliable confidence interval coverage (>90%) across all scenarios, whereas CMA failed to do so. Crucially, RDMA's errors reflect natural sampling variation, while CMA's reflect systematic bias that persists regardless of the sample size. We provide a decision tool for meta-analysts across disciplines to calculate RDMA's benefits. Our results offer the first quantitative evidence that open data improves meta-analytic accuracy, strengthening the case for open science.

1 Introduction

Meta-analysis is the main method for quantitatively synthesising existing evidence and is often used to inform future research, interventions, and policies [1–3]. Importantly, the statistical power of individual studies is often inadequate to detect true effects and can thus lead to imprecise effect estimates that limit the detection of meaningful effects [4–8]. Meta-analysis aggregates results from independent studies, increasing statistical power and precision [5]. Critically, this approach assumes that the included studies represent an unbiased sample of all research conducted on a given research question and that the included effects capture the true effects well. However, this assumption is often violated because selective publication and p-hacking can lead to overestimated effects and misguided decisions when meta-analysis relies on published results alone. Such selective publication and p-hacking are common in different scientific disciplines [9–13].

Publication bias [14], the preferential publication of statistically significant results, systematically distorts the published literature [12, 15–18]. Because significant results are more likely to be published, researchers might manipulate their analyses to achieve statistical significance (p-hacking) [19]. Data across disciplines demonstrate that these issues could be very widespread. For example, 94% of p-values reported in display items of the top multidisciplinary journals are significant [20], 96% of biomedical abstracts and full-text articles reporting p-values contain at least one significant result [21], while in ecology 98.9% of published articles report at least one significant result (computed from [22, 23]). Further, surveys of researchers show that p-hacking and other questionable research practices may be common: in psychology, self-admission rates for practices such as selective reporting and optional stopping range from 16–63% [24], with comparable rates confirmed across countries [25, 26] and in ecology and evolution [27], engineering [28], and across Dutch academia more broadly [29]. These findings convincingly indicate that researchers might selectively find and report significant effects, while null findings and non-significant results remain inaccessible, leading to a substantial loss of scientific knowledge [30].

Classical meta-analysis (CMA) commonly synthesises effect sizes calculated from the results reported in published studies, where each study contributes one or more effect size. Thus, the biases discussed above can lead to overstated meta-analytical means. For example, [31] simulations showed that p-hacking and publication bias interact to severely distort meta-analytic effect size estimates, particularly when true effects are small. Empirically, [15] found that correcting for publication bias across 87 ecological and evolutionary meta-analyses reduced effect sizes by at least 0.12 standard deviations on average, with 33 out

of 50 initially significant meta-analytic means becoming non-significant after correction. These biases can also lead to confidence intervals with sub-nominal coverage (i.e., intervals labelled as 95% may contain the true effect less than 95% of the time) [15]. Statistical corrections for publication bias exist, but often perform poorly in the presence of effect size heterogeneity [16], which is widespread in different research fields [32–34]. Consequently, even well-designed CMAs may systematically overestimate true effect sizes when relying solely on published results.

Raw data meta-analysis (RDMA) provides an alternative approach by calculating effect sizes directly from the original datasets of individual studies rather than extracting summary statistics from published articles. This approach is becoming increasingly viable across the range of disciplines, as open data becomes the norm [35]. Open data is often publicly available in repositories, allowing researchers to recompute the effect size for each study using raw observations. These recalculated study-level effects are then synthesised using standard meta-analytic models. This differs from the individual participant data (IPD) meta-analysis often used in clinical research, where participant-level data from multiple studies are combined into a single dataset and analysed jointly. In RDMA, raw data are used only to recompute the effect sizes within each study, and the meta-analysis still combines study-level estimates rather than pooling observations across studies.

However, while the growing availability of datasets through data-sharing mandates and changes in research culture [36–38] makes RDMA increasingly feasible, when and to what extent it outperforms CMA is unknown. Thus, it remains unclear when researchers should aim to conduct RDMA (or incorporate raw data in their meta-analysis) rather than CMA. Although more data are becoming open [39, 40], open data are often time-consuming to find [39, 40], interpret, and use [41, 42]. Here, we addressed this gap using a simulation-based approach, where the main conditions under which meta-analysis operates can be set to levels that reflect those present in a given research field. These conditions include the magnitude of true effect size, heterogeneity between studies, typical sample size, and the severity of publication bias. To parametrise the simulation and understand the potential benefits of RDMA, we used ecology as a case example. However, the framework was designed so that researchers can evaluate the performance of RDMA versus CMA in their own domain by modifying the input parameters (effect size magnitude, heterogeneity level, sample size, and bias severity). We provide worked examples using parameters from ecology, complete documentation, and annotated scripts accessible in the online repository [43].

2 Methodology

The study was preregistered on the Open Science Framework [44]. Deviations are explained in Supplementary Material S1. All code and data are openly available [43].

2.1 The overarching approach

Our framework simulates RDMA and CMA across different scenarios of true effect size magnitude and heterogeneity levels, and under different levels of presence of significant results in the published literature. For each scenario, the algorithm generates a population of N conducted studies, representing all studies that would be published if no selective publication occurred. Each study is represented by an effect size (Hedges' g) drawn from a distribution with mean θ (the true effect size) and variance τ^2 (the heterogeneity between studies). This population of studies is then subjected to a publication selection process designed to achieve target rates of statistical significance among published studies. We chose the Hedges' g as the effect size measure because standardised mean difference represents one of the two most common effect size measures in meta-analysis across disciplines [2, 3, 45, 46].

To model realistic conditions, the framework incorporates a simulated publication selection process that combines two mechanisms: (1) selective publication that favours significant results and (2) p-hacking that enables achieving the desired significance rates (Figure 1). The relative contribution of each mechanism to the final rates of significant results in the published literature is recorded across simulation scenarios (see Section 3.2). Some primary studies achieve statistical significance naturally (i.e., when the true effect size and the sample size allow for this). When the number of such naturally significant studies is insufficient to meet the target significance rates in published studies, p-hacking is applied. Here, the smallest needed number of non-significant studies, whose natural significance is $0.05 < p \leq 0.30$, is manipulated to reach significance. This approach ensures that p-hacking is applied conservatively—only when needed to reach empirically observed proportions of published significant results.

CMA then samples $k = N/2$ studies from the sample of studies that have gone through the simulated publication selection process. We have chosen to sample 50% of conducted studies as this reflects the publication rates detected across disciplines [18, 30, 47]. RDMA randomly samples $k = N/2$ studies from the population of all conducted studies (before the publication selection process). Performance of CMA and RDMA is then compared based on the accuracy of the effect size estimate, confidence interval coverage, and heterogeneity es-

timation (see Performance Metrics below for details). Compared to the logic of IPD meta-analysis in clinical research, our process is more relevant to other research areas (i.e. ecology, social sciences, environmental sciences) where effect sizes are calculated per study, rather than calculated for pooled data across studies.

The process is replicated 10,000 times for each combination of effect size, heterogeneity level, and bias scenario ($3 \times 3 \times 2 = 18$ combinations, of which 15 were valid in our case; see Section 2.2.3).

2.2 Empirical parameter estimation

To understand how different conditions affect the performance of CMA and RDMA, we use ecology as a case study. However, many of the parameters we use are realistic across scientific disciplines (see Discussion). To obtain realistic simulation parameters, we used two ecology literature datasets: (1) Costello and Fox [45, 46], which compiled 467 ecological meta-analyses with over 111,000 effect sizes, and (2) Kimmel et al. [22, 23], which contains 26,747 statistical estimates from 350 ecology papers published in Ecology, Ecology Letters, Journal of Ecology, Science and Nature between January 2018 and May 2020. Based on the Costello and Fox [45, 46] dataset, we calculated the typical effect size magnitude, between-study heterogeneity, and the number of studies used in meta-analysis. Based on both datasets [22, 23, 45, 46], we calculated the prevalence of statistically significant results in published studies. Sample sizes of primary studies were derived from Fox [48] and Kimmel et al. [22, 23] (see Section 2.2.2 for details).

2.2.1 Effect sizes and heterogeneity

Based on Costello and Fox [45, 46], ecological meta-analyses included absolute Hedges' g values distributed with quartiles of 0.22 (Q_1), 0.55 (median), and 1.1 (Q_3). These empirical values align closely with Cohen's (1988) conventions for small ($g = 0.2$), medium ($g = 0.5$), and large ($g = 0.8$) effect sizes. We therefore adopted Cohen's standardised thresholds for our simulations.

We computed between-study heterogeneity (τ^2) from the ecological meta-analyses compiled by Costello and Fox [45, 46], finding a distribution with $Q_1 = 0.26$, median $= 0.53$, and $Q_3 = 0.88$. Critically, τ^2 showed a modest positive correlation with the mean meta-analytical effect size magnitude ($r = 0.41$, $p = 0.005$). To capture this empirical pattern, rather than using fixed τ^2 values across all effect sizes, we scaled heterogeneity proportionally to the effect size magnitude:

- Low heterogeneity: $\tau^2 = 0.40\times$ effect size
- Medium heterogeneity: $\tau^2 = 1.00\times$ effect size

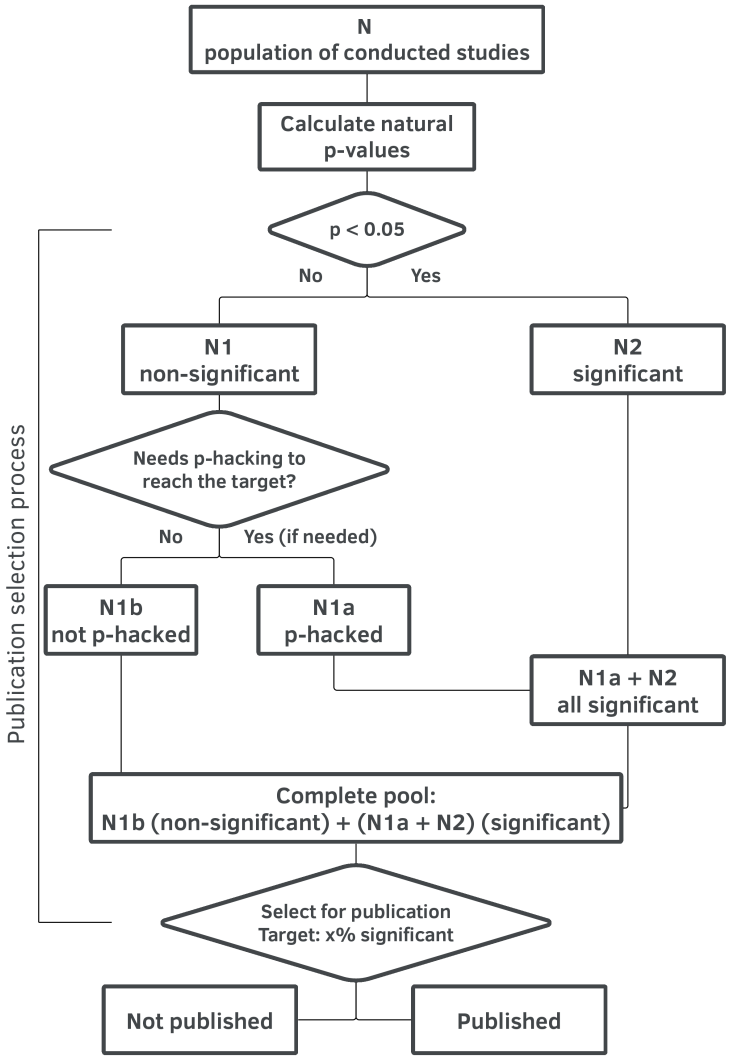

Figure 1: Flowchart illustrating the Classical meta-analysis (CMA) simulation process. The process begins with setting parameters: true effect size magnitude (θ), between-study heterogeneity (τ^2), sample sizes of primary studies (n_i), and population size of conducted studies (N). We then generate a population of N studies with known effect sizes and calculate natural p-values. Studies are then subjected to publication selection process with the aim to reach a target rate x% of $k = N/2$ published studies being significant. This process acts through selective publication and p-hacking.

- High heterogeneity: $\tau^2 = 1.50\times$ effect size

This parametrisation produced τ^2 values that span from below the empirical Q_1 (0.08 for small effects with low heterogeneity) to above Q_3 (1.20 for large effects with high heterogeneity), covering the full empirical range while maintaining consistent relative heterogeneity across effect size magnitudes.

2.2.2 Meta-analysis and primary studies sample size

The median ecological meta-analysis comprises 64 effect sizes from 23 studies [45, 46]. Given that our simulation assumes one effect size per study, each meta-analysis in our simulation included 23 published studies (and thus effects).

For determining the realistic sample size of the primary studies included in the meta-analysis, we used two existing large datasets, both of which have some limitations. The first dataset, Fox [48], provides sample sizes only for studies that report correlation coefficients (median $n = 30$), and not for standardised mean differences (Hedges' g), which we use in our simulation. The second dataset, [22, 23] reports sample size for various effect sizes, including group comparisons suitable for Hedges' g (median $n = 79$). However, these studies come from the general primary literature rather than specifically from studies included in meta-analyses. To reconcile these sources, we used a sampling approach that reflects the variation in the sample sizes of ecological studies: 25% of studies with $n = 20 - 30$ (small samples), 50% with $n = 30 - 70$ (moderate samples), and 25% with $n = 70 - 100$ (larger samples). This stratified sampling approach was applied to all $N = 46$ conducted primary studies in each simulation iteration, with sample sizes randomly drawn from the specified ranges. This yielded a median of around $n = 40$, a value that is intermediate between the median sample sizes of the primary studies included in the ecological meta-analysis ($n = 30$, [48]) and the median of the primary literature ($n = 79$, [22, 23]). The total sample size was split equally between the treatment and control groups.

This approach captures the low statistical power typical of many research fields. For example, [4] found only 13–16% power to detect small effects and 40–47% for medium effects in behavioural ecology, reflecting the prevalent small sample sizes. Similarly, [6] reported mean power of just 48% for medium effects in psychology, [7] estimated a median power of 21% across neuroscience, and [8] found a median power of 18% or less across 159 areas of economics research.

2.2.3 Publication selection process rates

For the purpose of this study, we define bias as the percentage of published studies that contain a significant result achieved through both selective publication and p-hacking. Because each study contributes one effect size in our simulation, the percentage of significant studies equals the percentage of significant effects. Based on Kimmel et al. [22, 23] data, we calculated that 98.9% of published ecological studies contain at least one p-value below 0.05. From Costello and Fox [45, 46], we calculated that 64% of studies included in meta-analyses contain at least one significant result (amongst the results included in the meta-analysis). Consequently, we simulated two scenarios of bias severity, i.e., the percentage of statistically significant effects included in meta-analyses: 95% (near primary literature conditions, representing severe distortion) and 65% (matching typical meta-analytic data, representing moderate distortion). Throughout this paper, we refer to these scenarios as 'severe bias' and 'moderate bias'. For large effect sizes ($g = 0.8$), the natural significance rate exceeded 65% due to high statistical power (studies naturally achieved significance without bias), making the 65% bias scenario inapplicable. This resulted in 15 valid simulation scenarios.

2.3 Data generation process

Step 1: Generate population of all conducted studies. We generated 46 studies representing all conducted research before the publication selection process. This number is twice the number of studies used in meta-analysis in ecology (k=23, i.e. published studies in our case) and reflects empirical data on how much of conducted research is published in ecology (55.3% [30]), but also in other disciplines (approximately 50% of clinical trials [47], 25–29% of doctoral dissertations in psychology [49]) and selective non-publication of null results is pervasive across the social sciences (only 10 out of 48 null results were published, compared to 56 out of 91 strong results; [18]). This 2-to-1 ratio of conducted vs published studies determines the need for p-hacking.

For each study, we simultaneously: (1) drew one true effect size from $\theta_i = \mu_\theta + \delta_i$ where $\delta_i \sim N(0, \tau^2)$, using $\mu_\theta \in \{0.2, 0.5, 0.8\}$ and scaling τ^2 proportionally as described above; and (2) assigned a total sample size using stratified sampling (25% small: 20–30, 50% medium: 30–70, 25% large: 70–100), which were then split equally between the treatment and the control group.

Step 2: Simulate data and calculate effect sizes. For each study, we simulated individual-level data for treatment and control groups to generate the observed effect size (Hedges' g), its sampling

variance, and p-value, following standard simulation approaches for meta-analysis [50, 51]. In this process, each study is also assigned its natural significance.

Step 3: Create published literature through selective publication process. We constructed a published studies pool targeting 23 studies (our target meta-analysis sample size) where the target percentage (95% or 65%) of studies contained significant results. Some studies from the population of 46 conducted studies achieved statistical significance ($p < 0.05$) naturally in Step 2, based on their true effects and sample sizes. When the number of such naturally significant studies was insufficient to meet the target significance rate, we applied p-hacking to the minimum number of non-significant studies necessary to reach the target. We used only those studies with the original significance $0.05 < p \leq 0.30$, and ordered them by statistical significance. If the study was amongst the first 50% of studies (those with the lowest p-values), it underwent mild manipulation: the p-value was replaced with a random value drawn from a uniform distribution $U(0.01, 0.049)$ while retaining the original effect size. If it was in the remaining 50% (with higher p-values), it underwent aggressive manipulation: a new p-value was drawn from $U(0.01, 0.049)$, a corresponding t-statistic was back-calculated from the new p-value and the study's degrees of freedom, and the effect size was recalculated as $g_i = t_i\sqrt{v_i}$, where v_i is the original sampling variance. Starting from those originally closest to statistical significance, only the minimum number of studies needed to reach the target significance rate were manipulated; all remaining studies retained their original p-values and effect sizes. This approach reflects realistic questionable research practices ranging from rounding of p-values to optional stopping or manipulation of data [27]. Recent evidence suggests that more than half of statistically significant findings in environmental sciences may result from such practices [52]. The resulting published literature of 23 studies combined naturally significant studies (randomly sampled from available naturally significant studies), p-hacked studies, and a small proportion of non-significant studies (randomly sampled from the remaining non-significant studies) that were not excluded by the simulated publication selection process.

Step 4: Sample for meta-analysis. CMA used all 23 studies from the published (biased) literature created in Step 3. RDMA randomly sampled 23 studies from the original unbiased population of 46 conducted studies (Step 1, before p-hacking and publication selection). This choice makes two simplified assumptions: that CMA includes all available published studies (which is likely not true) and that raw data meta-analysis has access to the same number of studies. However, depending on the research question, RDMA might have access to more studies (i.e. where data are published, though the study was not) or to fewer studies (i.e. where datasets are not available). Our choice reflects

a state between the ideal one (all raw data are available) and the worst one (no data available), while maintaining a balanced comparison by keeping the number of studies equal between the two methods.

Step 5: Meta-analytical estimates. Both CMA and RDMA used random-effects meta-analysis with REML estimation to obtain effect size estimates ($\hat{\theta}$), heterogeneity estimates ($\hat{\tau}^2$), and 95% confidence intervals.

A complete mathematical formalisation of the simulation algorithm—including the exact conditioning structure, the construction of the published sample, and the expectation and bias of the CMA estimator—is provided in Supplementary Material S2.

2.4 Performance metrics

We evaluated the performance of CMA and RDMA using:

Mean absolute error (MAE): $\frac{1}{n}\sum_{i=1}^{n}|\hat{\theta}_i - \mu_\theta|$ is the average magnitude of the estimation error across all iterations, where $n = 10{,}000$ is the number of simulation iterations.

Relative mean absolute error (RelMAE): $\frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{\theta}_i-\mu_\theta|}{\mu_\theta}$ is the estimation error expressed as the proportion of the true effect size, enabling comparison across effect size magnitudes. Throughout this manuscript, when reporting percentage improvements or reductions between CMA and RDMA, we use this metrics to enable fair comparison across different effect size magnitudes.

Confidence interval coverage: is the proportion of 95% CIs that contain the true effect size. Well-calibrated methods should achieve 95% or higher coverage.

Heterogeneity estimation accuracy: the mean absolute deviation $E[|\hat{\tau}-\tau|]$ and relative deviation $E[(\hat{\tau}-\tau)/\tau]$, measuring how well the between-study standard deviation ($\tau = \sqrt{\tau^2}$) is estimated.

2.5 Software

All analyses were conducted in R version 4.4.2 [53], using metafor (v4.6-0) [54], dplyr (v1.1.4) [55], tidyr (v1.3.1) [56], ggplot2 (v4.0.1) [57], and patchwork (v1.2.0) [58].

3 Results

3.1 Overview

Table 1 presents performance metrics across simulation scenarios. Overall, under severe bias (95%), RDMA substantially outperformed CMA

across the performance metrics. RDMA's advantages were most pronounced under low to medium heterogeneity, and increased with decreasing sample size of primary studies used in meta-analysis. Under moderate bias (65%), RDMA maintained advantages primarily for small to medium effects with low heterogeneity. Conceptually, these gains reflect the benefit of bypassing the publication filter, an advantage also relevant to IPD meta-analysis in clinical research.

3.2 P-hacking patterns

The simulation revealed that the contribution of p-hacking to bias considerably varied with the effect size magnitude. For small effects (ES=0.2) under severe bias (95%), p-hacking rates also varied substantially across heterogeneity levels: 52.6% under low heterogeneity, 34.4% under medium heterogeneity, and 23.2% under high heterogeneity. For medium effects (ES=0.5), p-hacking was considerably lower (9.8%, 2.7%, and 0.9% for low, medium, and high heterogeneity, respectively), and for large effects (ES=0.8), essentially no p-hacking was needed (0.1%, 0.1%, and 0.0%), suggesting that selective nonpublication dominated over questionable research practices in these scenarios. Under moderate bias (65%), p-hacking was substantially reduced: small effects required 25.1%, 8.4%, and 2.7% p-hacking (low, medium, and high heterogeneity), while medium effects required minimal manipulation (0.3%, 0.0%, and 0.0%).

3.3 Effect size estimation across the scenarios

RDMA's estimates of the mean effect size are consistently centred on the true effect size (dashed lines), with distribution widths reflecting the inherent sampling variability from $k = 23$ studies and heterogeneity (Figure 2). In contrast, CMA's estimates are systematically shifted upward under severe bias (Figure 2, Panel B), with the degree of the shift varying by effect size magnitude and heterogeneity level. RelMAE values (Figure 2, Panels C and D) demonstrate how these patterns translate to proportional estimation error.

The magnitude of CMA's overestimation directionally varied with the p-hacking rates (Section 3.2, Table 1). For small effects with low heterogeneity, and under severe bias, where p-hacking was the highest (52.6%), CMA deviated the most from the true effect, and RDMA achieved a 76% reduction in mean absolute error (MAE). This mechanistic link explained a counter-intuitive pattern visible in Table 1: for small effects, CMA's MAE decreases as heterogeneity increases (0.29 $\rightarrow$ 0.25 $\rightarrow$ 0.23). This occurred because increased heterogeneity decreased the need for p-hacking (52.6% $\rightarrow$ 34.4% $\rightarrow$ 23.2%), thus reducing CMA's systematic bias, more than it increased the random error.

Table 1: Performance Metrics

Heterogeneity Level	Method	65% Bias ME est.	65% Bias MAE	65% Bias RelMAE	65% Bias CI Cov.	65% Bias Mean p-hack %	95% Bias ME est.	95% Bias MAE	95% Bias RelMAE	95% Bias CI Cov.	95% Bias Mean p-hack %
True ES = 0.2											
Low	CMA	0.37	0.17	0.86	0.57	25.14	0.49	0.29	1.43	0.20	52.63
	RDMA	0.20	0.07	0.33	0.94	—	0.21	0.07	0.34	0.94	—
Medium	CMA	0.33	0.15	0.77	0.81	8.36	0.44	0.25	1.23	0.56	34.42
	RDMA	0.20	0.09	0.45	0.93	—	0.20	0.09	0.45	0.94	—
High	CMA	0.30	0.15	0.74	0.88	2.73	0.42	0.23	1.16	0.70	23.23
	RDMA	0.19	0.10	0.51	0.94	—	0.20	0.10	0.51	0.94	—
True ES = 0.5											
Low	CMA	0.64	0.14	0.29	0.78	0.32	0.81	0.31	0.62	0.20	9.85
	RDMA	0.49	0.09	0.18	0.94	—	0.49	0.09	0.18	0.94	—
Medium	CMA	0.59	0.14	0.29	0.92	0.02	0.81	0.31	0.63	0.56	2.73
	RDMA	0.49	0.13	0.25	0.93	—	0.49	0.13	0.25	0.94	—
High	CMA	0.56	0.16	0.31	0.94	0.00	0.78	0.30	0.60	0.72	0.92
	RDMA	0.49	0.15	0.30	0.94	—	0.49	0.15	0.30	0.94	—
True ES = 0.8											
Low	CMA			—			1.07	0.27	0.34	0.31	0.12
	RDMA			—			0.78	0.11	0.13	0.93	—
Medium	CMA			—			1.10	0.32	0.40	0.63	0.07
	RDMA			—			0.78	0.16	0.20	0.94	—
High	CMA			—			1.08	0.32	0.39	0.76	0.01
	RDMA			—			0.78	0.19	0.24	0.94	—

Note: True ES = true effect size (Hedges' g). ME est. = Mean estimated effect size across 10,000 iterations. MAE = Mean Absolute Error. RelMAE = Relative Mean Absolute Error (MAE/true ES). CI Cov. = Confidence Interval Coverage (proportion of CIs containing the true ES; target: 0.95). Mean p-hack % = Mean percentage of studies requiring p-hacking to achieve target publication bias across iterations. Heterogeneity levels are defined relative to effect size ($\tau^2 = 0.40\times$, $1.00\times$, $1.50\times$ ES), so absolute τ^2 values differ across effect size scenarios (see Section 2.2.1)

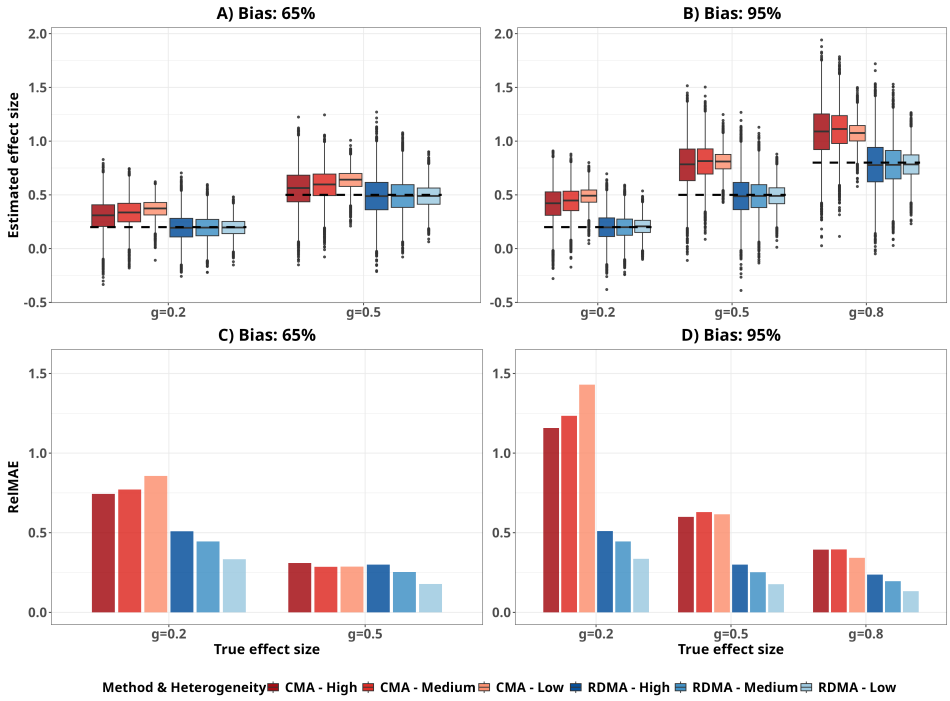

Figure 2: Effect size estimates and RelMAE comparing CMA and RDMA across scenarios. CMA is shown in red, RDMA in blue. The intensity of each colour represents the level of heterogeneity among the true effect sizes (darker colours represent higher heterogeneity). Panels A and B: Boxplots of estimated effect sizes under moderate (65%) and severe (95%) bias, summarising 10,000 iterations with dashed horizontal lines indicating true effect sizes. Panels C and D: RelMAE under moderate and severe bias. *Note: RDMA distributions are identical in panels A and B (and C and D) as this method samples from the unbiased population; RDMA is displayed in both bias scenarios to facilitate direct visual comparison with CMA*

For large effects with minimal p-hacking (0.1%), CMA's MAE followed the expected pattern, increasing with heterogeneity (0.27 → 0.32). In contrast, RDMA's MAE increased monotonically with heterogeneity, confirming that errors reflect natural sampling variance rather than bias. For medium and large effects with minimal p-hacking (0.9–9.7% and 0.0–0.1%), CMA's overestimation remained substantial, demonstrating that selective publication alone distorts estimates.

Under moderate bias (Figure 2, Panel A), CMA's estimates are closer to true values than under severe bias, yet systematic overestimation persisted, particularly for small effects with low heterogeneity, where p-hacking remained at 25.1%. For medium to large effects with high heterogeneity, where p-hacking was minimal (≤0.3%), CMA and RDMA distributions substantially overlap, indicating that CMA performs reasonably well in these specific conditions.

3.4 Confidence interval coverage

RDMA maintained near-nominal CI coverage (93–94%) across all simulation scenarios, closely approaching the 95% target expected of a well-calibrated method. CMA's CI coverage, in contrast, was severely degraded under severe bias, with the extent of coverage depending critically on the effect size magnitude and heterogeneity (Figure 3, Panel B).

Under severe bias (95%), CMA's coverage was worst under low heterogeneity: only 20% coverage for small and medium effects (i.e. 80% of CIs failed to capture the true effect), and 31% for large effects. CMA performed slightly better under medium heterogeneity (56% coverage for small to medium effects, 63% for large effects), while high heterogeneity brought coverage closer to the acceptable levels (70–76%), though still substantially below RDMA's consistent >92%. This pattern is largely mechanical: higher heterogeneity inflates the estimated between-study variance, widening the confidence intervals and increasing the probability of capturing the true effect despite systematic bias in point estimates.

Under moderate bias (65%), CMA approached acceptable performance for medium effects with medium-to-high heterogeneity (92–94% coverage), but still showed a substantial undercoverage for small effects with low heterogeneity (57%).

3.5 Relative mean absolute error

RDMA exhibited minimal systematic bias (the mean estimates within 2–5% of true values across all scenarios). RDMA's RelMAE ranged from 13–51% depending on effect size and heterogeneity, primarily reflecting a natural sampling variation from $k = 23$ studies rather than

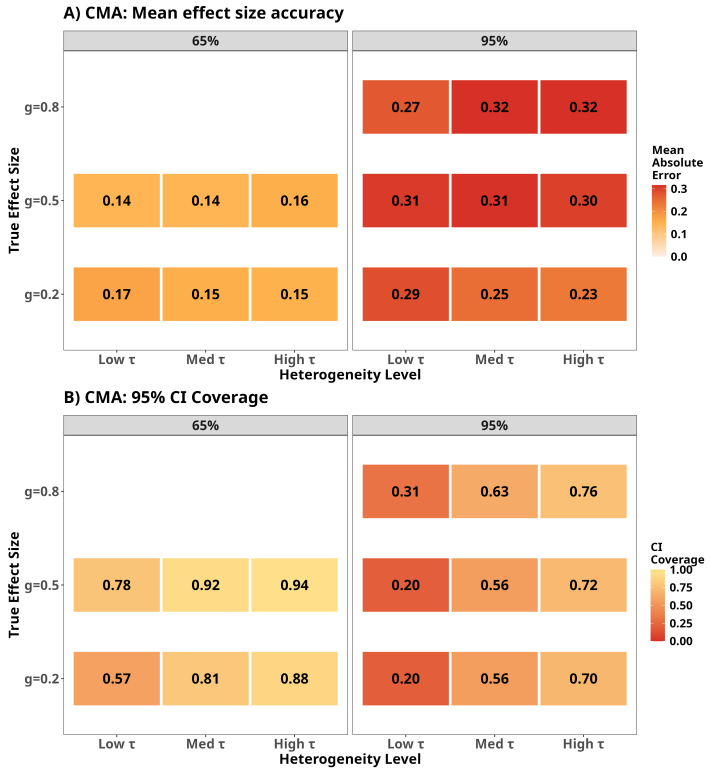

Figure 3: Performance heatmaps for CMA across simulation conditions. Panel A: Mean absolute error (MAE) in effect size estimates (lower values indicate larger accuracy). Panel B: 95% confidence interval coverage (target = 0.95). Rows represent true effect sizes ($g = 0.2, 0.5, 0.8$), columns represent heterogeneity levels (low, medium, high τ), and facets show bias scenarios (65% and 95% significant results amongst published results)

a systematic over- or underestimation. In some scenarios, RDMA showed slight underestimation (e.g., mean estimate of 0.2 for true $ES = 0.2$ under high heterogeneity), but RelMAE remained driven by the heterogeneity-induced variance. As heterogeneity increased within each magnitude of the effect size, RDMA's RelMAE increased proportionally (e.g., $ES = 0.2$: 34% → 45% → 51% across low, medium, and high heterogeneity), confirming that errors likely stem from sampling variance in heterogeneous populations rather than from bias.

In contrast, CMA showed both systematic overestimation and high RelMAE. For small effects with low heterogeneity under severe bias, CMA's mean estimate represented 143% overestimation with 143% RelMAE; nearly identical values indicating a systematic bias rather than a random variation. Standardising absolute error by effect size revealed different relative impacts: a similar MAE represents 143% error for $ES = 0.2$ but only 34% error for $ES = 0.8$, demonstrating why RelMAE is essential for comparing accuracy across effect sizes.

Under moderate bias (65%), CMA's RelMAE improved substantially (29–86% across scenarios), approaching RDMA's performance for medium effects with high heterogeneity where p-hacking was minimal (Table 1, Figure 2).

3.6 Heterogeneity estimation

Both methods estimated heterogeneity reasonably well, with RDMA performing slightly better than the CMA (Figure 4). RDMA's MAE for heterogeneity estimation (τ) was consistently lower than CMA's across scenarios, with the largest advantage under severe bias (95%): RDMA MAE 0.088 vs. CMA MAE 0.187, a difference of 0.099 units (ES = 0.2, high heterogeneity), and near-zero differences in the least biased scenarios. Complete heterogeneity estimation results for all the iterations are available in the accompanying data files.

Under moderate bias (65%), the performance gap nearly disappeared for medium and large effects with medium-to-high heterogeneity, although RDMA maintained its advantage for small effects and low heterogeneity scenarios (Table 1, Figure 4).

4 Discussion and Conclusions

We provide the first estimate of the benefits that raw (open) data can bring to meta-analysis by mitigating the effects of publication bias and p-hacking. Our results reveal that raw data meta-analysis yields widespread advantages over the classical meta-analysis (when the sample number of included primary studies and their sample size is the same between the two). These advantages depend on the extent of

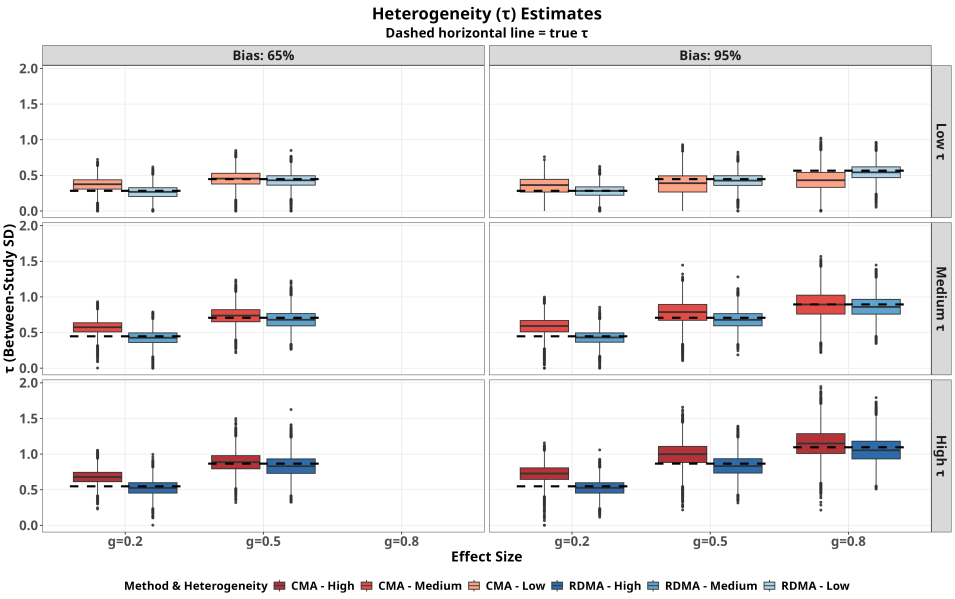

Figure 4: Heterogeneity estimates (τ, between-study standard deviation) from CMA (red) and RDMA (blue). Box-plots show distributions from 10,000 iterations. Dashed lines indicate true τ values under three scenarios of the true effect size values.

publication bias, with the magnitude of the true effect size and the heterogeneity level both playing important moderating roles. Further, our simulations also revealed under what conditions p-hacking is supported. We discuss our findings and their implications below.

4.1 Conditions favouring p-hacking

Our simulation revealed that the contribution of p-hacking to overall bias (which we define as the percentage of significant results in published literature) largely depends on three factors: the magnitude of the true effect size, the heterogeneity between the studies, and the severity of bias. The need for p-hacking is substantial when the true effects are small ($ES = 0.2$) and when bias is large (95% of published results are significant), peaking at 52.6% under low heterogeneity. When the heterogeneity of small effects is low, true effect sizes cluster tightly around a small mean value. Given that sample sizes typical for ecology (median $n \approx 40$) rarely achieve sufficient power to reach significance for such effects, p-hacking becomes necessary to reach publication thresholds of significant p-values. As heterogeneity of small effects increases, the distribution of true effect sizes widens, so some studies have larger true effects that achieve natural significance (statistical significance under the data-generating model), reducing the need for p-hacking. This explains why, paradoxically, CMA's mean absolute error decreases with increasing heterogeneity for $ES = 0.2$: higher heterogeneity reduces the proportion of studies requiring p-hacking, thereby reducing the severity of bias.

In contrast, under severe bias, minimal p-hacking was needed for medium effects (mean 4.5%) and essentially none for large effects (0.07%), where natural significance rates approach or exceed target levels due to adequate statistical power. Here, selective publication dominated over p-hacking. Under moderate bias (65%), p-hacking requirements were substantially reduced but not eliminated: small effects still required considerable p-hacking (25.1% under low heterogeneity), while medium effects required minimal manipulation ($\leq$0.3%). These patterns reflect our modelling choice of an approximate 50% publication rate, which is likely realistic, and consistent with empirical evidence across fields [18, 30, 47, 49]. This ratio makes our estimates conservative: lower publication rates would increase p-hacking requirements. Further, effects in ecology are typically small to medium (median $|g| = 0.55$; [45, 46]), and low statistical power is prevalent across many research fields [4, 7, 8]. Thus, our results demonstrate that the current incentives and publishing models, where significant results are favoured, are likely the key to pushing scientists to p-hack.

4.2 When to use raw data meta-analysis

Overall, the benefits of RDMA for accurately estimating the mean effect increase as true effect size and heterogeneity are smaller, yet remain substantial in most scenarios. The benefits are also larger under severe bias (95% of published effects are statistically significant) compared to moderate bias. These findings suggest that RDMA provides the greatest value for research questions with relatively homogeneous effects, such as controlled experiments with standardised protocols, well-defined interventions, or phenomena governed by consistent mechanisms. It provides somewhat less value when studying highly context-dependent phenomena with inherent large heterogeneity, although it still yields meaningful improvements over CMA in these contexts.

RDMA's advantages are most pronounced for small effects with low heterogeneity (76% error reduction) because RDMA circumvents both selective publication and severe p-hacking in these scenarios. For larger effects, RDMA primarily addresses selective publication alone (as p-hacking rates in this scenario are small), yielding smaller but still substantial advantages (38–62% error reduction). Higher heterogeneity reduces RDMA's relative advantage across all effect sizes because true variation among studies becomes large relative to the distortion introduced by bias. Critically, the type of error differs between the two methods. RDMA's errors stem from natural sampling variation: individual meta-analyses vary around the true effect, as expected when randomly sampling $k = 23$ studies from a larger population. This produces unbiased estimates with moderate precision. CMA's errors, in contrast, reflect systematic bias, with estimates consistently shifted away from true values in a predictable direction. This distinction is fundamental: random error decreases with more studies, while systematic bias persists regardless of sample size. RDMA therefore provides not just more accurate estimates, but estimates that improve predictably as data accumulate.

Our simulation suggests that RDMA's primary advantage lies in accessing the unbiased population of conducted studies, bypassing the publication selection filter that distorts CMA estimates. In our framework, both RDMA and CMA apply the same analytical method (random-effects meta-analysis of study-level effect sizes); the difference is solely in the sampling frame and the integrity of the effect estimates entering the meta-analysis. CMA uses summary statistics as reported in publications, which may be distorted by selective publication and p-hacking, whereas RDMA accesses the original unmanipulated data. Our results therefore quantify the access channel through which raw data benefits meta-analysis. This advantage may be shared by other approaches that bypass the publication filter and use original datasets, such as IPD meta-analysis in clinical research, though the mecha-

nisms differ. For simple treatment–control comparisons with continuous outcomes—the design that our simulation captures—published evidence suggests that one-stage participant-level modelling and two-stage study-level approaches yield practically equivalent summary treatment effects [59–61], though our simulation does not test this directly. The additional analytical advantages of IPD meta-analysis—such as consistent covariate adjustment, exploration of participant-level treatment-effect modifiers, and appropriate handling of time-to-event outcomes—become important for more complex research questions that our framework does not address.

4.3 Limitations and future directions

We make several simplifying assumptions. First, we conservatively modelled p-hacking only when needed to reach target bias rates, making our RDMA benefit estimates conservative; in reality, questionable practices are more diverse. Second, we assumed RDMA randomly samples all conducted studies with complete data. This design implies that RDMA has access to all conducted studies, including unpublished studies, and that it can use original unmanipulated data, randomly sampling the entire research population. In practice, selective data sharing and quality issues [41] could reduce RDMA's advantages, so results represent an upper bound. Similar constraints are well recognised in IPD meta-analysis, where data availability and access decisions can shape the final evidence base. However, RDMA can provide larger sample sizes than CMA when published literature is limited [40]. On the other hand, in an ideal scenario where open data are the norm, RDMA might have access to a larger pool of studies than CMA, likely increasing the benefits beyond those reported here.

Third, our simulation examined only a single effect size per study from simple treatment-control comparisons and therefore did not capture potential within-study heterogeneity in effect sizes for studies that report on more than one effect. Related to this, our representation of the publication selection process enforced a target prevalence of statistically significant results among published studies, with one effect size per study. In practice, selection mechanisms are more complex across disciplines: entire studies may remain unpublished (file drawer effect), researchers may selectively report significant results while omitting non-significant ones from the same study, and published studies typically contain multiple effect sizes rather than one. Clinical research may exhibit additional structured selection mechanisms, including time-lag bias and sponsor-dependent dissemination patterns. Fourth, we tested only random-effects models with REML estimation; other meta-analytic approaches might show different patterns. Finally, we did not examine meta-regression or subgroup analyses, which are

common in practice.

Future work should thus examine scenarios with multiple effects per study to understand how within-study selection interacts with publication bias. Future extensions could also implement explicit selection probability models, in which the probability of publication is modelled as a function of statistical significance (e.g., high probability for $p < 0.05$ and lower probability otherwise) and study characteristics, including sponsor status. Such models are widely used in clinical selection-model literature and would allow direct comparison with aggregate-data correction approaches developed for medical meta-analysis. Incorporating sponsor-dependent or outcome-level selection would be particularly relevant for extending our framework to broader bias scenarios encountered in regulatory or guideline-setting contexts.

Finally, our choice of 50% publication rate was based on empirical evidence from ecology and other fields [18, 30, 47, 49], and represents a conservative scenario. Our severe bias scenario (95% significant results) may be less realistic for ecology, where meta-analyses typically include studies that report multiple effect sizes, and only 64% of such studies contain at least one significant result among the effects included in the meta-analysis [45, 46]. However, this scenario may be more realistic in fields where primary studies commonly report a single effect size, as the proportion of significant results among published studies tends to be higher in such contexts. Publication rates might also vary across fields and research questions: if true publication rates in a specific domain are lower (e.g., 30%, requiring $N \approx 77$ conducted studies for $k = 23$ published), p-hacking requirements would be substantially higher, and RDMA's advantages would exceed those reported here. Conversely, if publication rates are higher (e.g., 70%, requiring only $N \approx 33$ conducted studies), RDMA's benefits would be smaller. Future work could explore sensitivity to this parameter by systematically varying the N:k ratio to establish boundary conditions for RDMA's benefits across different publication rate scenarios. This is easily done by adjusting the corresponding part of our simulation code. Such analyses would help researchers assess whether RDMA is worthwhile in their specific field based on field-specific publication rates.

Our estimates cannot at the moment be evaluated against real data as raw data meta-analysis remains very rare. However, if more meta-analyses begin using raw data, following existing guidelines [62], it might be possible to evaluate the assumptions of our model and its estimates. Hybrid approaches that combine published effect sizes with available raw data merit investigation, as they may provide a pragmatic middle ground. Empirical studies quantifying the time and resource costs of data acquisition would help researchers make informed decisions about when to pursue RDMA. Such extensions would strengthen understanding of when RDMA justifies the additional time investment

it requires.

4.4 Tool for meta-analysts

Our simulation framework is designed to be field-agnostic. Researchers can assess RDMA's potential benefits in their domain (or a particular research area) by replacing our ecological parameters with values relevant to their scenario. Required inputs include: (1) expected (typical) effect size magnitudes (Hedges' g or other standardised effect sizes), (2) between-study heterogeneity estimates (τ^2), (3) typical sample sizes of primary studies (e.g., treatment and control group sizes), (4) the typical number of published studies per meta-analysis ($k = 23$ in our simulations, representing the median in ecology), (5) the estimated publication rate in the field (approximately 50% in ecology, meaning $N = 46$ conducted studies are needed to yield $k = 23$ published studies), and (6) bias severity (proportion of significant results in the published literature). Both k and the publication rate are necessary because together they determine N, the population of conducted studies from which RDMA samples, and consequently how much p-hacking is required to achieve target bias rates: a lower publication rate (requiring a larger N of conducted studies to yield k published studies) increases p-hacking requirements and amplifies RDMA's advantages.

The code, which can be accessed at [43], automatically generates performance metrics (mean absolute error, confidence interval coverage, heterogeneity estimation accuracy) across user-specified scenarios. For researchers who want a quick preliminary estimate of expected CMA bias without running the full simulation, we also provide a closed-form approximation based on truncated normal theory (Supplementary Material S3) with an accompanying R script (`S3.R`) that returns an approximate bias estimate for a given combination of effect size, heterogeneity, sample sizes, and publication bias severity. Although our simulation used Hedges' g as the effect size measure, the framework can be adapted to other common meta-analytic metrics such as the log response ratio (lnRR) or Fisher's z-transformed correlation coefficient. The overall architecture—including the publication bias mechanism, the RDMA versus CMA sampling design, and all performance metrics—is effect size agnostic. Adaptation requires replacing the data-generating function with one appropriate for the targeted effect size, adjusting the p-hacking mechanism to match the corresponding test statistic, and recalibrating input parameters to reflect empirical distributions for the chosen metric. Detailed guidance is provided in the code repository [43].

4.5 Conclusions

RDMA is overall superior to CMA; however, its advantages depend on the interaction of bias severity, the true effect size magnitude, and the heterogeneity level. Under severe bias (95% significant results published) with low to medium heterogeneity, RDMA substantially reduces mean absolute error compared to CMA: by 56–76% for small effects ($g = 0.2$), 50–71% for medium effects ($g = 0.5$), and 38–62% for large effects ($g = 0.8$). RDMA consistently maintained 93–94% confidence interval coverage across all scenarios, while CMA's coverage ranged from only 20–31% under low heterogeneity to 70–76% under high heterogeneity. Under moderate bias (65%) with high heterogeneity and medium effects, the performance gap narrows substantially, with CMA performing adequately.

Given that raw data meta-analysis provides substantially more accurate effect size estimates across the scenarios most plausible for a range of research fields, researchers should more often seek to incorporate raw data into meta-analyses. This can be done by either using raw data from published studies or by identifying raw data independently [40, 62]. However, raw data are often difficult to find and use, as their quality is often very low [41, 42]. The research community, including funders and publishers, should thus seek to increase data quality and thereby increase the reliability of estimated effects.

Author contributions: Conceptualisation: D.Ž., A.C.; Methodology: D.Ž., S.G., A.T., A.C.; Software: D.Ž.; Validation: S.G., A.T.; Formal analysis: D.Ž.; Visualisation: D.Ž.; Funding acquisition: A.C.; Writing—original draft: D.Ž., A.C.; Writing—review and editing: S.G., A.T., A.C., L.P.

Competing interests: The authors declare that no competing interests exist.

Funding: This work was supported by the Croatian Science Foundation under the project number HRZZ-IP-2022-10-2872.

Data availability: All simulation code, parameter estimation scripts, input data, and complete results are available in the OSF repository: https://doi.org/10.17605/OSF.IO/VEBYM [43].

Acknowledgements: We are grateful to Dr Alfredo Sánchez-Tójar for providing feedback on aspects of the methodology. AI tools (Claude, Anthropic) were used to assist with manuscript formatting and code documentation. All content was reviewed and verified by the authors.

Ethics: This study is based entirely on simulated data. No ethical approval was required.

References

[1] S. Nakagawa et al. "Quantitative evidence synthesis: A practical guide on meta-analysis, meta-regression, and publication bias tests for environmental sciences". In: *Environmental Evidence* 12.1 (2023), p. 8. DOI: 10.1186/s13750-023-00301-6.

[2] M. Borenstein et al. *Introduction to Meta-Analysis*. 2nd. Wiley, 2021. ISBN: 978-1-119-55835-4.

[3] J.P.T. Higgins et al. *Cochrane Handbook for Systematic Reviews of Interventions*. 2nd. Wiley-Blackwell, 2019. ISBN: 978-1-119-53662-8.

[4] M.D. Jennions and A.P. Møller. "A survey of the statistical power of research in behavioral ecology and animal behavior". In: *Behavioral Ecology* 14.3 (2003), pp. 438–445. DOI: 10.1093/beheco/14.3.438.

[5] Y. Yang et al. "Low statistical power and overestimated anthropogenic impacts, exacerbated by publication bias, dominate field studies in global change biology". In: *Global Change Biology* 28.3 (2022), pp. 969–989. DOI: 10.1111/gcb.15972.

[6] J. Cohen. "The statistical power of abnormal-social psychological research: A review". In: *The Journal of Abnormal and Social Psychology* 65.3 (1962), pp. 145–153. DOI: 10.1037/h0045186.

[7] K.S. Button et al. "Power failure: why small sample size undermines the reliability of neuroscience". In: *Nature Reviews Neuroscience* 14.5 (2013), pp. 365–376. DOI: 10.1038/nrn3475.

[8] J.P.A. Ioannidis, T.D. Stanley, and H. Doucouliagos. "The power of bias in economics research". In: *The Economic Journal* 127.605 (2017), F236–F265. DOI: 10.1111/ecoj.12461.

[9] M.L. Head et al. "The extent and consequences of p-hacking in science". In: *PLOS Biology* 13.3 (2015), e1002106. DOI: 10.1371/journal.pbio.1002106.

[10] D. Fanelli. ""Positive" results increase down the hierarchy of the sciences". In: *PLOS ONE* 5.4 (2010), e10068. DOI: 10.1371/journal.pone.0010068.

[11] D. Fanelli. "Negative results are disappearing from most disciplines and countries". In: *Scientometrics* 90.3 (2012), pp. 891–904. DOI: 10.1007/s11192-011-0494-7.

[12] A. Brodeur, N. Cook, and A. Heyes. "Methods matter: p-hacking and publication bias in causal analysis in economics". In: *American Economic Review* 110.11 (2020), pp. 3634–3660. DOI: 10.1257/aer.20190687.

[13] S.B. Bruns and J.P.A. Ioannidis. "p-Curve and p-hacking in observational research". In: *PLOS ONE* 11.2 (2016), e0149144. DOI: 10.1371/journal.pone.0149144.

[14] T.D. Sterling. "Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa". In: *Journal of the American Statistical Association* 54.285 (1959), pp. 30–34. DOI: 10.1080/01621459.1959.10501497.

[15] Y. Yang et al. "Publication bias impacts on effect size, statistical power, and magnitude (Type M) and sign (Type S) errors in ecology and evolutionary biology". In: *BMC Biology* 21.1 (2023), p. 71. DOI: 10.1186/s12915-022-01485-y.

[16] S. Nakagawa et al. "Methods for testing publication bias in ecological and evolutionary meta-analyses". In: *Methods in Ecology and Evolution* 13.1 (2022), pp. 4–21. DOI: 10.1111/2041-210X.13724.

[17] E.H. Turner et al. "Selective publication of antidepressant trials and its influence on apparent efficacy". In: *The New England Journal of Medicine* 358.3 (2008), pp. 252–260. DOI: 10.1056/NEJMsa065779.

[18] A. Franco, N. Malhotra, and G. Simonovits. "Publication bias in the social sciences: unlocking the file drawer". In: *Science* 345.6203 (2014), pp. 1502–1505. DOI: 10.1126/science.1255484.

[19] J.P. Simmons, L.D. Nelson, and U. Simonsohn. "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant". In: *Psychological Science* 22.11 (2011), pp. 1359–1366. DOI: 10.1177/0956797611417632.

[20] I.A. Cristea and J.P.A. Ioannidis. "P values in display items are ubiquitous and almost invariably significant: A survey of top science journals". In: *PLOS ONE* 13.5 (2018), e0197440. DOI: 10.1371/journal.pone.0197440.

[21] D. Chavalarias et al. "Evolution of reporting p values in the biomedical literature, 1990-2015". In: *JAMA* 315.11 (2016), pp. 1141–1148. DOI: 10.1001/jama.2016.1952.

[22] K. Kimmel, M.L. Avolio, and P.J. Ferraro. "Empirical evidence of widespread exaggeration bias and selective reporting in ecology". In: *Nature Ecology & Evolution* 7.9 (2023), pp. 1525–1536. DOI: 10.1038/s41559-023-02144-3.

[23] K. Kimmel and P.J. Ferraro. *Replicability in ecology*. Dataset, OSF. Accessed 15 December 2025. 2023. URL: https://osf.io/9yd2b.

[24] L.K. John, G. Loewenstein, and D. Prelec. "Measuring the prevalence of questionable research practices with incentives for truth telling". In: *Psychological Science* 23.5 (2012), pp. 524–532. DOI: 10.1177/0956797611430953.

[25] F. Agnoli et al. "Questionable research practices among Italian research psychologists". In: *PLOS ONE* 12.3 (2017), e0172792. DOI: 10.1371/journal.pone.0172792.

[26] K. Fiedler and N. Schwarz. "Questionable research practices revisited". In: *Social Psychological and Personality Science* 7.1 (2016), pp. 45–52. DOI: 10.1177/1948550615612150.

[27] H. Fraser et al. "Questionable research practices in ecology and evolution". In: *PLOS ONE* 13.7 (2018), e0200303. DOI: 10.1371/journal.pone.0200303.

[28] A.A. Baptista and F. Pereira. "Questionable research practices in engineering research". In: *International Journal of Information Systems and Project Management* 14.1 (2026), pp. 1–24. DOI: 10.12821/ijispm140101.

[29] G. Gopalakrishna et al. "Prevalence of questionable research practices, research misconduct and their potential explanatory factors: a survey among academic researchers in The Netherlands". In: *PLOS ONE* 17.2 (2022), e0263023. DOI: 10.1371/journal.pone.0263023.

[30] M. Purgar, T. Klanjscek, and A. Culina. "Quantifying research waste in ecology". In: *Nature Ecology & Evolution* 6.9 (2022), pp. 1390–1397. DOI: 10.1038/s41559-022-01820-0.

[31] M. Friese and J. Frankenbach. "p-Hacking and publication bias interact to distort meta-analytic effect size estimates". In: *Psychological Methods* 25.4 (2020), pp. 456–471. DOI: 10.1037/met0000246.

[32] A.M. Senior et al. "Heterogeneity in ecological and evolutionary meta-analyses: its magnitude and implications". In: *Ecology* 97.12 (2016), pp. 3293–3299. DOI: 10.1002/ecy.1591.

[33] S. van Erp et al. "Estimates of between-study heterogeneity for 705 meta-analyses reported in Psychological Bulletin from 1990–2013". In: *Journal of Open Psychology Data* 5 (2017), p. 4. DOI: 10.5334/jopd.33.

[34] J.P. Higgins et al. "Measuring inconsistency in meta-analyses". In: *BMJ* 327.7414 (2003), pp. 557–560. DOI: 10.1136/bmj.327.7414.557.

[35] Digital Science et al. *The State of Open Data 2023*. 2023. DOI: 10.6084/m9.figshare.24428194.v1. URL: https://digitalscience.figshare.com/articles/report/The_State_of_Open_Data_2023/24428194.

[36] B.L. Houtkoop et al. "Data sharing in psychology: A survey on barriers and preconditions". In: *Advances in Methods and Practices in Psychological Science* 1.1 (2018), pp. 70–85. DOI: 10.1177/2515245917751886.

[37] A. Pepe et al. "How do astronomers share data? Reliability and persistence of datasets linked in AAS publications and a qualitative study of data practices among US astronomers". In: *PLOS ONE* 9.8 (2014), e104798. DOI: 10.1371/journal.pone.0104798.

[38] G. Colavizza et al. "The citation advantage of linking publications to research data". In: *PLOS ONE* 15.4 (2020), e0230416. DOI: 10.1371/journal.pone.0230416.

[39] L. Tedersoo et al. "Data sharing practices and data availability upon request differ across scientific disciplines". In: *Scientific Data* 8 (2021), p. 192. DOI: 10.1038/s41597-021-00981-0.

[40] A. Culina et al. "Navigating the unfolding open data landscape in ecology and evolution". In: *Nature Ecology & Evolution* 2.3 (2018), pp. 420–426. DOI: 10.1038/s41559-017-0458-2.

[41] D.G. Roche et al. "Public data archiving in ecology and evolution: how well are we doing?" In: *PLOS Biology* 13.11 (2015), e1002295. DOI: 10.1371/journal.pbio.1002295.

[42] D.G. Roche et al. "Slow improvement to the archiving quality of open datasets shared by researchers in ecology and evolution". In: *Proceedings of the Royal Society B: Biological Sciences* 289.1975 (2022), p. 20212780. DOI: 10.1098/rspb.2021.2780.

[43] D. Žanko. *Reliable inference: benefits of open raw data may be universal in meta-analysis - data and code*. OSF. Accessed 27 March 2026. 2026. URL: https://doi.org/10.17605/OSF.IO/VEBYM.

[44] D. Žanko et al. *Performance of open data versus classical meta-analysis in ecology - a simulation study*. OSF Preprints. Accessed 15 December 2025. 2025. URL: https://doi.org/10.17605/OSF.IO/KAH42.

[45] L. Costello and J.W. Fox. "Decline effects are rare in ecology". In: *Ecology* 103.6 (2022), e3680. DOI: 10.1002/ecy.3680.

[46] J.W. Fox and L. Costello. *Decline effects are rare in ecology: a meta-meta-analysis*. Accessed 15 December 2025. Dryad, 2022. URL: https://doi.org/10.5061/dryad.zkh1893b7.

[47] I. Chalmers and P. Glasziou. "Avoidable waste in the production and reporting of research evidence". In: *Lancet* 374.9683 (2009), pp. 86–89. DOI: 10.1016/S0140-6736(09)60329-9.

[48] J.W. Fox. "The sample size of the typical ecological correlation coefficient is small and slowly declining". In: *Oikos* 2025.12 (2025), e11430. DOI: 10.1002/oik.11430.

[49] S.C. Evans et al. ""Are you gonna publish that?" Peer-reviewed publication outcomes of doctoral dissertations in psychology". In: *PLOS ONE* 13.2 (2018), e0192219. DOI: 10.1371/journal.pone.0192219.

[50] F. Gambarota and G. Altoè. "Understanding meta-analysis through data simulation with applications to power analysis". In: *Advances in Methods and Practices in Psychological Science* 7.1 (2024). DOI: 10.1177/25152459231209330.

[51] L.V. Hedges. "Distribution theory for Glass's estimator of effect size and related estimators". In: *Journal of Educational Statistics* 6.2 (1981), pp. 107–128. DOI: 10.3102/10769986006002107.

[52] T. Deressa et al. *More Than Half of Statistically Significant Research Findings in the Environmental Sciences are Actually Not*. Accessed 15 December 2025. 2023. URL: https://doi.org/10.32942/x24g6z.

[53] R Core Team. *R: A language and environment for statistical computing*. Accessed 15 December 2025. R Foundation for Statistical Computing. Vienna, Austria, 2024. URL: https://www.R-project.org/.

[54] W. Viechtbauer. "Conducting meta-analyses in R with the metafor package". In: *Journal of Statistical Software* 36.3 (2010), pp. 1–48. DOI: 10.18637/jss.v036.i03.

[55] H. Wickham et al. *dplyr: A grammar of data manipulation*. R package version 1.1.4. Accessed 15 December 2025. 2023. URL: https://CRAN.R-project.org/package=dplyr.

[56] H. Wickham, D. Vaughan, and M. Girlich. *tidyr: Tidy messy data*. R package version 1.3.1. Accessed 15 December 2025. 2024. URL: https://CRAN.R-project.org/package=tidyr.

[57] H. Wickham. *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag, 2016. ISBN: 978-3-319-24277-4.

[58] T.L. Pedersen. *patchwork: The composer of plots*. R package version 1.2.0. Accessed 15 December 2025. 2024. URL: https://CRAN.R-project.org/package=patchwork.

[59] D.L. Burke, J. Ensor, and R.D. Riley. "Meta-analysis using individual participant data: one-stage and two-stage approaches, and why they may differ". In: *Statistics in Medicine* 36.5 (2017), pp. 855–875. DOI: 10.1002/sim.7141.

[60] R.D. Riley et al. "Two-stage or not two-stage? That is the question for IPD meta-analysis projects". In: *Research Synthesis Methods* 14.6 (2023), pp. 903–910. DOI: 10.1002/jrsm.1661.

[61] T.P. Morris et al. "Meta-analysis of Gaussian individual patient data: two-stage or not two-stage?" In: *Statistics in Medicine* 37.9 (2018), pp. 1419–1438. DOI: 10.1002/sim.7589.

[62] A. Culina et al. "How to do meta-analysis of open datasets". In: *Nature Ecology & Evolution* 2.7 (2018), pp. 1053–1056. DOI: 10.1038/s41559-018-0579-2.

Supplementary Material S1: Deviations from Preregistration

Our simulation approach evolved during empirical data analysis to better reflect realistic conditions in ecological meta-analysis. Key modifications align simulation parameters with empirical evidence while maintaining core objectives.

Number of studies per meta-analysis

We initially planned $k = 70$ studies based on the median number of effect sizes per meta-analysis. However, examining Fox and Costello[46] more carefully revealed that the median meta-analysis comprises 64 effect sizes from only 23 unique studies. Since our simulation generates one effect size per study, $k = 23$ better reflects typical primary studies in ecological meta-analyses.

Sample size allocation

Our preregistration specified 50% of studies with sample sizes $n = 20-30$, 25% with $n = 30-70$, and 25% with $n = 70-100$. We revised this to 25% with $n = 20-30$, 50% with $n = 30-70$, and 25% with $n = 70-100$ to better represent the range of sample sizes in ecological research while maintaining prevalence of small samples.

Heterogeneity parametrisation

We initially specified heterogeneity as $\tau = 10-30\%$ of effect size (τ being standard deviation). Empirical analysis revealed modest positive correlation between absolute effect size and τ^2 (variance) ($r = 0.41$, $p = 0.005$), leading us to express heterogeneity as proportions of effect size magnitude: $\tau^2 = 0.40\times$ effect size (low), $1.00\times$ effect size (medium), and $1.50\times$ effect size (high). This produces values spanning from below empirical Q_1 (0.26) to above Q_3 (0.88) observed by Fox and Costello[46].

Additional bias scenario

While our preregistration focused on severe bias (95%), we added a moderate bias scenario (65%) based on Fox and Costello[46], which showed 64% of studies in ecological meta-analyses contain significant results. These refinements improve our simulation's ecological real-

ism without changing fundamental research questions or analytical approach.

Supplementary Material S2: Mathematical formalisation of the simulation algorithm

Here we formalize the publication-bias and p-hacking mechanism exactly as implemented in the simulation code. We explicitly state the conditions under which the algorithm is successful, and calculate the expected total reported effect under those conditions.

Exact conditioning structure induced by the simulation algorithm

Let k denote the target number of published studies entering the classical meta-analysis, and let K denote the size of the conducted-study population generated in each attempt ($k = 23$, $K = 46$). The code repeats generation attempts until a valid published literature is obtained, so all expectations relevant to the simulation are conditional on the event that the generation algorithm succeeds.

For each conducted study $i = 1, \ldots, K$, a total sample size N_i is generated, then split as

$$n_{Ti} = \left\lceil \frac{N_i}{2} \right\rceil, \qquad n_{Ci} = \left\lfloor \frac{N_i}{2} \right\rfloor, \qquad \mathrm{df}_i = N_i - 2.$$

A true study-specific effect is then generated as

$$\Delta_i = \mu + U_i, \qquad U_i \sim \mathcal{N}(0, \tau^2).$$

Conditional on (N_i, Δ_i), the study-level raw data are generated as

$$Y_{Ci1}, \ldots, Y_{Cin_{Ci}} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1), \qquad Y_{Ti1}, \ldots, Y_{Tin_{Ti}} \overset{\text{iid}}{\sim} \mathcal{N}(\Delta_i, 1).$$

Let

$$\bar{Y}_{Ci} = \frac{1}{n_{Ci}} \sum_{j=1}^{n_{Ci}} Y_{Cij}, \qquad \bar{Y}_{Ti} = \frac{1}{n_{Ti}} \sum_{j=1}^{n_{Ti}} Y_{Tij},$$

and let the pooled standard deviation be

$$S_{pi} = \left[\frac{(n_{Ti} - 1) S_{Ti}^2 + (n_{Ci} - 1) S_{Ci}^2}{N_i - 2} \right]^{1/2}.$$

The simulated Cohen's d and the Hedges correction factor are

$$d_i = \frac{\bar{Y}_{Ti} - \bar{Y}_{Ci}}{S_{pi}}, \; J_i = 1 - \frac{3}{4(N_i - 2) - 1}.$$

The observed study estimate is therefore

$$Y_i = J_i d_i,$$

and the within-study sampling variance is

$$V_i = \left(\frac{N_i}{n_{Ti} n_{Ci}} + \frac{Y_i^2}{2N_i} \right) J_i^2.$$

The associated test statistic and two-sided p-value are

$$T_i = \frac{Y_i}{\sqrt{V_i}}, \qquad P_i = 2\Big(1 - F_{t,\mathrm{df}_i}(|T_i|)\Big),$$

where $F_{t,\mathrm{df}}$ is the cumulative distribution function of the Student t distribution with df degrees of freedom.

The simulation partitions the conducted studies into three sets:

$$A = \{i : P_i \le 0.05\}, \qquad B = \{i : 0.05 < P_i \le 0.30\}, \qquad C = \{i : P_i > 0.30\}.$$

Set A contains the originally significant studies, B contains the non-significant studies eligible for p-hacking, and C contains the remaining non-significant studies. Let

$$a = |A|, \qquad b = |B|, \qquad c = |C|, \qquad M = a + b + c.$$

Note that we retain only studies with finite Y_i, finite V_i, and $V_i > 0$; all formulas below are conditional on the retained conducted literature.

Let

$$q \in \{0.65, 0.95\}$$

denote the target proportion of published studies that must be significant. We have

$$s^\star = \mathrm{round}(qk), \qquad \ell^\star = k - s^\star.$$

so that $s^\star$ is the target number of significant published studies and $\ell^\star$ is the target number of non-significant published studies.

If $a \ge s^\star$, then no p-hacking is needed. If $a < s^\star$, he number of additional significant studies that must be created by p-hacking is $h = s^\star - a$.

In all cases define

$$s = \min(a, s^\star).$$

which is the number of originally significant studies that are actually published.

A generation attempt is accepted only if the algorithm can construct the published sample and the resulting sample size falls within a tolerance window of the target. Let $\delta = \max(3, \mathrm{round}(0.15\,k))$. The exact success event is

$$\mathcal{E} = \{M \ge K\} \cap \{b \ge h\} \cap \{M - a - h \ge \ell^\star\} \cap \{|k_{\mathrm{actual}} - k| \le \delta\},$$

where $k_{\text{actual}} = s+h+\ell^{\star}_{\text{actual}}$ is the realised number of published studies. The first condition is the code's requirement that at least $K = 2k$ valid studies remain after filtering. The second condition ensures that there are enough eligible studies for p-hacking. The third condition ensures that after removing the hacked studies from the non-significant pool there are still enough remaining non-significant studies to fill the $\ell^{\star}$ non-significant publication slots. The fourth condition is the code's acceptance tolerance, which in practice is rarely binding ($\delta = 3$ for $k = 23$).

All expectations corresponding to the actual simulation are therefore conditional on $\mathcal{E}$.

Exact construction of the published sample

Conditional on the realized conducted literature and on $\mathcal{E}$, we construct the published literature as follows.

If $s > 0$, the set of originally significant published studies, denoted S, is drawn uniformly without replacement from A so that $|S| = s$. Hence

$$\Pr(S = \mathcal{S} \mid A, \mathcal{E}) = \binom{a}{s}^{-1}$$

for every $\mathcal{S} \subseteq A$ such that $|\mathcal{S}| = s$.

If $h > 0$, eligible p-hackable studies in B are ordered by increasing p-value:

$$P^B_{(1)} \leq P^B_{(2)} \leq \cdots \leq P^B_{(b)}.$$

The hacked set is then chosen deterministically as the h studies closest to significance:

$$H = \{i_{(1)}, \ldots, i_{(h)}\}.$$

Let

$$r = \left\lceil \frac{h}{2} \right\rceil.$$

The first r hacked studies, namely $i_{(1)}, \ldots, i_{(r)}$, undergo "mild" p-hacking, while the remaining $h-r$ hacked studies, namely $i_{(r+1)}, \ldots, i_{(h)}$, undergo "aggressive" p-hacking.

For the mild subgroup, the reported effect size is unchanged:

$$Y^{\star}_{i_{(j)}} = Y_{i_{(j)}}, \qquad j = 1, \ldots, r,$$

and only the reported p-value is replaced by an independent draw

$$P^{\star}_{i_{(j)}} \sim \text{Unif}(0.01, 0.049).$$

For the aggressive subgroup, we generate

$$U_{i_{(j)}} \overset{\text{iid}}{\sim} \text{Unif}(0.01, 0.049), \qquad j = r+1, \ldots, h,$$

then define

$$T^{\star}_{i_{(j)}} = F^{-1}_{t,\mathrm{df}_{i_{(j)}}}\left(1 - \frac{U_{i_{(j)}}}{2}\right).$$

The observed effect size becomes

$$Y^{\star}_{i_{(j)}} = T^{\star}_{i_{(j)}}\sqrt{V_{i_{(j)}}}.$$

The reported p-value is set equal to

$$P^{\star}_{i_{(j)}} = U_{i_{(j)}}.$$

Note the within-study variance is not altered, so for all hacked studies $V^{\star}_{i_{(j)}} = V_{i_{(j)}}$.

The remaining non-significant studies are

$$R = (B \cup C) \setminus H.$$

The published non-significant set N is sampled uniformly without replacement from R so that $|N| = \ell^{\star}$. Hence

$$\Pr(N = \mathcal{N} \mid R, \mathcal{E}) = \binom{|R|}{\ell^{\star}}^{-1}$$

for every $\mathcal{N} \subseteq R$ such that $|\mathcal{N}| = \ell^{\star}$.

The final published set is

$$\mathcal{P} = S \cup H \cup N, \qquad |\mathcal{P}| = k.$$

Expected effect among originally significant published studies

Because S is a simple random sample without replacement from A, the sample average of the originally significant published studies is unbiased for the finite-population average over A. Therefore, conditional on the realized conducted literature,

$$\mathbb{E}\left(\frac{1}{s}\sum_{i\in S} Y_i \,\middle|\, A, \mathcal{E}\right) = \frac{1}{a}\sum_{i\in A} Y_i, \qquad a > 0.$$

Equivalently, the exact expected effect among originally significant published studies is

$$\mu^{\mathrm{pub}}_{S}(D) = \frac{1}{a}\sum_{i\in A} Y_i, \qquad a > 0,$$

where D denotes the realized conducted literature.

Passing from the realized literature to the simulation distribution, the exact expected effect among originally significant published studies is

$$\mu_S^{\text{pub}} = \mathbb{E}\left(\frac{1}{|A|} \sum_{i \in A} Y_i \,\middle|\, |A| > 0, \mathcal{E} \right).$$

By exchangeability of the K conducted studies, this can also be written as

$$\mu_S^{\text{pub}} = \mathbb{E}\left(Y_1 \,|\, P_1 \leq 0.05, \mathcal{E} \right).$$

Expected reported effect among hacked studies

Let

$$\bar{Y}_H^\star = \frac{1}{h} \sum_{i \in H} Y_i^\star$$

denote the average reported effect among hacked studies, defined only when $h > 0$. Conditional on the realized conducted literature D and on the success event $\mathcal{E}$, the hacked set H is deterministic, and randomness arises only from the uniforms used in the severe hacking step. Hence

$$\mathbb{E}(\bar{Y}_H^\star \mid D, \mathcal{E}, h > 0) = \frac{1}{h} \left[\sum_{j=1}^{r} Y_{i_{(j)}} + \sum_{j=r+1}^{h} \mathbb{E}\left(Y_{i_{(j)}}^\star \,\middle|\, D, \mathcal{E} \right) \right].$$

For $j = r+1, \ldots, h$,

$$Y_{i_{(j)}}^\star = \sqrt{V_{i_{(j)}}}\, F_{t,\mathrm{df}_{i_{(j)}}}^{-1}\left(1 - \frac{U_{i_{(j)}}}{2} \right), \qquad U_{i_{(j)}} \sim \mathrm{Unif}(0.01, 0.049),$$

so

$$\mathbb{E}\left(Y_{i_{(j)}}^\star \,\middle|\, D, \mathcal{E} \right) = \sqrt{V_{i_{(j)}}}\, m(\mathrm{df}_{i_{(j)}}),$$

where

$$m(d) = \frac{1}{0.039} \int_{0.01}^{0.049} F_{t,d}^{-1}\left(1 - \frac{u}{2} \right) du.$$

Therefore the exact conditional expectation among hacked studies is

$$\mu_H(D) := \mathbb{E}(\bar{Y}_H^\star \mid D, \mathcal{E}, h > 0) = \frac{1}{h} \left[\sum_{j=1}^{r} Y_{i_{(j)}} + \sum_{j=r+1}^{h} \sqrt{V_{i_{(j)}}}\, m(\mathrm{df}_{i_{(j)}}) \right].$$

Averaging over the simulation law yields the exact unconditional expectation

$$\mu_H = \mathbb{E}\left(\mu_H(D) \,|\, h > 0, \mathcal{E} \right).$$

Equivalently,

$$\mu_H = \mathbb{E}\left[\frac{1}{h} \left(\sum_{j=1}^{\lceil h/2 \rceil} Y_{i_{(j)}} + \sum_{j=\lceil h/2 \rceil + 1}^{h} \sqrt{V_{i_{(j)}}}\, m(\mathrm{df}_{i_{(j)}}) \right) \middle| h > 0, \mathcal{E} \right].$$

Expected reported effect among all significant published studies

The set of significant published studies is

$$G = S \cup H, \qquad |G| = s + h = s^\star.$$

Its average reported effect is

$$\bar{Y}_G^\star = \frac{1}{s^\star}\left(\sum_{i\in S} Y_i + \sum_{i\in H} Y_i^\star\right).$$

Conditional on D and $\mathcal{E}$,

$$\mathbb{E}(\bar{Y}_G^\star \mid D, \mathcal{E}) = \frac{1}{s^\star}\left[\frac{s}{a}\sum_{i\in A} Y_i + \sum_{j=1}^{r} Y_{i_{(j)}} + \sum_{j=r+1}^{h} \sqrt{V_{i_{(j)}}}\, m(\mathrm{df}_{i_{(j)}})\right].$$

Exact expectation and bias of the CMA estimator

In the simulation, an intercept-only random-effects meta-analysis model is fitted to the published sample by REML. Let

$$\mathbf{Y}^\star = (Y_1^\star, \ldots, Y_k^\star)^\top, \qquad \mathbf{V}^\star = (V_1^\star, \ldots, V_k^\star)^\top$$

denote the published effect sizes and within-study variances after the publication and p-hacking algorithm has been applied. The published-study ordering is immaterial. Define

$$\hat{\tau}^2_{\mathrm{REML}}(\mathbf{y}, \mathbf{v})$$

to be the REML estimate of the between-study variance in the intercept-only model returned by the `rma` fit, and define

$$\psi_{\mathrm{REML}}(\mathbf{y}, \mathbf{v}) = \frac{\sum_{j=1}^{k} \dfrac{y_j}{v_j + \hat{\tau}^2_{\mathrm{REML}}(\mathbf{y}, \mathbf{v})}}{\sum_{j=1}^{k} \dfrac{1}{v_j + \hat{\tau}^2_{\mathrm{REML}}(\mathbf{y}, \mathbf{v})}}.$$

Then the CMA point estimator is exactly

$$\hat{\mu}_{\mathrm{CMA}} = \psi_{\mathrm{REML}}(\mathbf{Y}^\star, \mathbf{V}^\star).$$

Conditional on the realized conducted literature D and on $\mathcal{E}$, the only remaining randomness comes from the uniform subset S of originally significant studies, the uniform subset N of non-significant published studies, and the uniforms used in the severe p-hacking step. Hence the exact conditional expectation of the CMA estimator is

$$\mathbb{E}(\hat{\mu}_{\mathrm{CMA}} \mid D, \mathcal{E}) = \frac{1}{\binom{a}{s}\binom{M-a-h}{\ell^\star}} \sum_{\substack{\mathcal{S}\subseteq A \\ |\mathcal{S}|=s}} \sum_{\substack{\mathcal{N}\subseteq (B\cup C)\setminus H \\ |\mathcal{N}|=\ell^\star}} \left[\frac{1}{0.039^{\,h-r}} \int_{[0.01,0.049]^{h-r}} \psi_{\mathrm{REML}}\big(\mathbf{y}^\star, \mathbf{v}^\star\big)\, d\mathbf{u}\right],$$

where

$$\psi_{\mathrm{REML}}\big(\mathbf{y}^\star, \mathbf{v}^\star\big) = \psi_{\mathrm{REML}}\big(\mathbf{y}^\star(D, \mathcal{S}, \mathcal{N}, \mathbf{u}), \mathbf{v}^\star(D, \mathcal{S}, \mathcal{N})\big),$$

$\mathbf{y}^\star(D, \mathcal{S}, \mathcal{N}, \mathbf{u})$ is the k-vector of published effect sizes formed by concatenating

$$\{Y_i : i \in \mathcal{S}\},\ \{Y_{i_{(j)}} : j = 1, \ldots, r\},\ \left\{\sqrt{V_{i_{(j)}}}\, F^{-1}_{t, \mathrm{df}_{i_{(j)}}}\left(1 - \frac{u_j}{2}\right) : j = r+1, \ldots, h\right\},\ \{Y_i : i \in \mathcal{N}\},$$

and $\mathbf{v}^\star(D, \mathcal{S}, \mathcal{N})$ is the matching vector of within-study variances, which is simply the corresponding collection of original V_i's.

The exact unconditional expectation of the CMA estimator under the simulation is therefore

$$\mathbb{E}(\hat{\mu}_{\mathrm{CMA}} \mid \mathcal{E}) = \mathbb{E}\left[\mathbb{E}(\hat{\mu}_{\mathrm{CMA}} \mid D, \mathcal{E}) \,|\, \mathcal{E}\right].$$

Accordingly, the exact bias of the CMA estimator induced by the simulation algorithm is

$$\mathrm{Bias}^{\mathrm{REML}}_{\mathrm{CMA}} = \mathbb{E}(\hat{\mu}_{\mathrm{CMA}} \mid \mathcal{E}) - \mu.$$

This expression is exact for finite sample representation as defined by the algorithm. It is not reducible to a simpler elementary closed form without some approximations and inexact simplifications.

Supplementary Material S3: Closed-form approximation for CMA bias

We derive a closed-form approximation for CMA bias under selective publication and p-hacking, using the truncated normal distribution. This avoids the combinatorial complexity of the exact expressions in S2 while closely matching the simulation results (Table 2).

Setup

We use the notation of S2 throughout. Each conducted study i has true effect $\Delta_i \sim \mathcal{N}(\mu, \tau^2)$, sample size N_i, observed effect $Y_i = J_i d_i$ (Hedges' g), and within-study variance $V_i \approx (N_i / n_{Ti} n_{Ci})\, J_i^2$. A study is significant when $|Y_i| > Y_{\text{crit},i}$, where $Y_{\text{crit},i} = t_{0.975,\, N_i - 2}\sqrt{V_i}$. As in S2, $s^\star = \text{round}(qk)$ studies in the published sample must be significant and $\ell^\star = k - s^\star$ are non-significant.

Assumptions

We make two simplifications. First, conditional on (Δ_i, N_i):

$$Y_i \mid \Delta_i, N_i \;\dot{\sim}\; \mathcal{N}(J_i \Delta_i,\; V_i).$$

Second, we approximate the REML-weighted estimator $\hat{\mu}_{\text{CMA}} = \psi_{\text{REML}}(\mathbf{Y}^\star, \mathbf{V}^\star)$ (S2) by a simple average of published effects. This is reasonable because the REML weights $1/(V_i + \hat{\tau}^2)$ become approximately uniform as $\hat{\tau}^2$ grows.

Truncated normal expectations

For a study with $\Delta_i = \delta$, $N_i = n$, let $\sigma_i = \sqrt{V_i}$, $\bar{Y}_i = J\delta$, and define

$$\alpha_u = \frac{Y_{\text{crit}} - \bar{Y}_i}{\sigma_i}, \qquad \alpha_\ell = \frac{-Y_{\text{crit}} - \bar{Y}_i}{\sigma_i},$$

where ϕ and Φ are the standard normal density and CDF. The expected effect among significant studies (right tail, which dominates for positive effects) is

$$\mathbb{E}[Y_i \mid Y_i > Y_{\text{crit}},\, \delta, n] \;=\; \bar{Y}_i + \sigma_i \cdot \frac{\phi(\alpha_u)}{1 - \Phi(\alpha_u)},$$

with probability of significance $\pi(\delta, n) = 1 - \Phi(\alpha_u) + \Phi(\alpha_\ell)$. For non-significant studies (doubly-truncated normal):

$$\mathbb{E}[Y_i \mid |Y_i| < Y_{\text{crit}},\, \delta, n] \;=\; \bar{Y}_i + \sigma_i \cdot \frac{\phi(\alpha_\ell) - \phi(\alpha_u)}{\Phi(\alpha_u) - \Phi(\alpha_\ell)}.$$

Integration over the study population

Since Δ_i and N_i vary across studies, we integrate over $\delta \sim \mathcal{N}(\mu, \tau^2)$ and average over sample-size strata $\{(n_j, w_j)\}$ (midpoints: $n = 25, 50, 85$ with weights $0.25, 0.50, 0.25$):

$$\mathbb{E}[Y \mid \text{sig}] = \frac{\displaystyle\int \sum_j w_j \, \mathbb{E}[Y \mid Y > Y_{\text{crit}},\, \delta, n_j] \; \pi(\delta, n_j) \; f(\delta) \, d\delta}{\displaystyle\int \sum_j w_j \; \pi(\delta, n_j) \; f(\delta) \, d\delta},$$

where $f(\delta)$ is the $\mathcal{N}(\mu, \tau^2)$ density. The non-significant expectation $\mathbb{E}[Y \mid \text{non-sig}]$ is analogous. The marginal significance probability is $\bar{\pi} = \int \sum_j w_j \, \pi(\delta, n_j) \, f(\delta) \, d\delta$.

P-hacking components

When $K\bar{\pi} < s^\star$, the shortfall $h = s^\star - K\bar{\pi}$ studies require p-hacking (S2). The first $r = \lceil h/2 \rceil$ undergo mild hacking (effect Y_i unchanged), the remaining $h - r$ undergo aggressive hacking. Their expected effects are:

Mild: these studies have $p \in (0.05, 0.30]$; their original effects sit between the $p = 0.30$ and $p = 0.05$ boundaries, so the expected original effect is approximated as

$$\mathbb{E}[Y \mid \text{mild}] \;\approx\; \tfrac{1}{2}(Y_{\text{low}} + Y_{\text{crit}}),$$

where $Y_{\text{low}} = t_{0.85,\,\bar{N}-2}\,\bar{\sigma}$ is the $p = 0.30$ boundary.

Aggressive: following S2, $Y_i^\star = T_i^\star \sqrt{V_i}$ with $T_i^\star = F^{-1}_{t,\mathrm{df}_i}(1 - U_i/2)$, $U_i \sim \text{Unif}(0.01, 0.049)$, giving

$$\mathbb{E}[Y^\star \mid \text{aggressive}] \;=\; \bar{\sigma} \cdot m(\bar{d}), \qquad m(d) = \frac{1}{0.039} \int_{0.01}^{0.049} F^{-1}_{t,d}\big(1 - \tfrac{u}{2}\big)\, du \quad \text{(S2)}.$$

Combined approximation

With $s = \min(K\bar{\pi},\, s^\star)$ originally significant published studies:

$$\mathbb{E}[\hat{\mu}_{\text{CMA}}] \;\approx\; \frac{1}{k}\Big[\, s\,\mathbb{E}[Y \mid \text{sig}] + r\,\mathbb{E}[Y \mid \text{mild}] + (h-r)\,\mathbb{E}[Y^\star \mid \text{aggressive}] + \ell^\star\,\mathbb{E}[Y \mid \text{non-sig}] \Big],$$

$$\text{Bias}^{\text{approx}}_{\text{CMA}} = \mathbb{E}[\hat{\mu}_{\text{CMA}}] - \mu.$$

When $h = 0$ (no p-hacking needed), this simplifies to $\mathbb{E}[\hat{\mu}_{\text{CMA}}] \approx (s^\star/k)\,\mathbb{E}[Y \mid \text{sig}] + (\ell^\star/k)\,\mathbb{E}[Y \mid \text{non-sig}]$.

These expressions allow researchers to obtain a quick estimate of expected CMA bias for their own field-specific parameters—effect size magnitude, heterogeneity, sample sizes, and publication bias severity—without running the full simulation. An accompanying R script (`S3.R`) is provided in the repository.

Interpretation

The bias is driven by the ratio of the significance threshold Y_{crit} to the true effect. When $Y_{\text{crit}} \gg \mu$ (small effects, small samples), only studies with large positive sampling errors pass the filter, and the uplift $\phi(\alpha_u)/(1-\Phi(\alpha_u))$ is large. When μ is large enough that most studies are naturally significant, the truncation point falls into the left tail and the uplift vanishes.

Increasing τ^2 widens the distribution of Δ_i, so more studies achieve significance through genuinely large effects rather than sampling luck, attenuating the selection bias. This matches the simulation finding that RDMA's advantage shrinks with heterogeneity.

CI coverage and error bounds

Under selective publication, the CMA estimate is centred on $\mu + \text{Bias}$ while the CI is constructed as if unbiased. With half-width $w = z_{0.975} \cdot \text{SE}$:

$$\text{Coverage} \;\approx\; \Phi\left(\frac{w - \text{Bias}}{\text{SE}}\right) - \Phi\left(\frac{-w - \text{Bias}}{\text{SE}}\right).$$

When $\text{Bias} \gg w$, coverage collapses, consistent with 20% coverage for small effects under severe bias. The MAE decomposes as $|\text{Bias}| + \mathbb{E}[|\varepsilon|]$ where $\varepsilon \sim O(\sqrt{(\bar{V} + \tau^2)/k})$. The bias is a structural property of the publication filter and does not diminish with k; only the random component shrinks. RDMA eliminates the bias term entirely.

Validation

Table 2 compares the approximation with simulation (Table 1, 10,000 iterations). Mean absolute difference: 0.018; maximum: 0.06. The largest gaps occur for small effects under severe bias with low heterogeneity, where p-hacking is heaviest and the midpoint approximation for borderline studies is coarsest.

Table 2: Approximation vs. simulation for $\mathbb{E}[\hat{\mu}_{\text{CMA}}]$

True ES	**Het.**	**Bias**	**Approx.**	**Sim.**	**Diff.**
0.2	Low	65%	0.41	0.37	+0.04
0.2	Low	95%	0.55	0.49	+0.06
0.2	Medium	65%	0.33	0.33	+0.00
0.2	Medium	95%	0.50	0.44	+0.06
0.2	High	65%	0.30	0.30	+0.00
0.2	High	95%	0.45	0.42	+0.03
0.5	Low	65%	0.63	0.64	−0.01
0.5	Low	95%	0.83	0.81	+0.02
0.5	Medium	65%	0.58	0.59	−0.01
0.5	Medium	95%	0.81	0.81	+0.00
0.5	High	65%	0.55	0.56	−0.01
0.5	High	95%	0.77	0.78	−0.01
0.8	Low	95%	1.07	1.07	+0.00
0.8	Medium	95%	1.10	1.10	+0.00
0.8	High	95%	1.07	1.08	−0.01