Robust point and variance estimation for ecological and evolutionary meta-analyses with selective reporting and dependent effect sizes

Yefeng Yang[a], Malgorzata Lagisz[a], Coralie Williams, Jinming Pan[b,*], Daniel W. A. Noble[c] & Shinichi Nakagawa[a,d,*]

[a] Evolution & Ecology Research Centre and School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, NSW 2052, Australia

[b] Department of Biosystems Engineering, Zhejiang University, Hangzhou 310058, China

[c] Division of Ecology and Evolution, Research School of Biology, The Australian National University, Canberra, ACT, Australia

[d] Theoretical Sciences Visiting Program, Okinawa Institute of Science and Technology Graduate University, Onna, 904-0495, Japan

[*] Corresponding author

s.nakagawa@unsw.edu.au (S. Nakagawa), panhouse@zju.edu.cn (J. Pan).

Open Science

Raw data and analytical script to reproduce examples presented in the manuscript is archived at https://github.com/Yefeng0920/BiasRobustMA_tutorial. A webpage showing the implementation of the proposed method in combination with a visualisation tool can be accessed via https://yefeng0920.github.io/BiasRobustMA_tutorial/.

Abstract

Meta-analysis produces a quantitative synthesis of evidence-based knowledge, shaping not only research trends but also policy and practices in ecology and evolution. However, two statistical issues, selective reporting and statistical dependence, can severally distort meta-analytic evidence. Here, we propose a two-step procedure to tackle these challenges concurrently and re-analyse 448 ecological and evolutionary meta-analyses. First, we employ bias-robust weighting schemes under the generalized least square estimator to obtain less biased population mean effect sizes by mitigating selective reporting. Second, we use cluster-robust variance estimation to account for statistical dependence and reduce bias in estimating standard errors, ensuring valid statistical inference. Re-analyses of 448 meta-analyses show that ignoring the two issues tends to overestimate effect sizes by an average of 110% and underestimate standard errors by 120%. Our approach is effective at mitigating these biases in meta-analytic evidence. To facilitate the implementation, we have developed a website showing the step-by-step tutorial available on our website. Complementing the current meta-analytic practice with the proposed method can facilitate a transition to a more pluralistic approach in quantitative evidence synthesis in ecology and evolution.

Main

Quantitative synthesis of multiple research findings has become increasingly important for guiding scientific research and informing evidence-based decision-making [1]. Meta-analytic modelling is the most commonly used quantitative evidence synthesis method [2] and has been widely applied in various disciplines, including the natural and social sciences (e.g., ecology and evolution [3,4], medicine [5,6], environment [7], education [8], psychology [9], management [10], and economics [11,12]). In ecology and evolution, there are numerous statistical models available, but two basic ones are the fixed-effect (FE) and random-effects (RE) models [13]. The FE model assumes that true effect sizes are homogeneous across studies [14]. In contrast, the RE model assumes that true effect sizes are heterogeneous across studies [15]. Despite their popularity, both FE and RE models have limitations in dealing with ecological and evolutionary meta-analytic datasets with complex structures, which may result in unreliable parameter estimation (e.g., the point estimate of model coefficient) and statistical inference (e.g., null-hypothesis test and confidence intervals, CIs) [3].

One common feature of ecological and evolutionary meta-analytic datasets is statistical dependence [9,16,17], which arises mainly due to the presence of multiple effect size estimates from the same study (Figure 1) [18]. This issue is also pervasive across disciplines, as evidenced by the presence of multiple estimates per study (e.g., a feature of 100% of meta-analyses in environmental sciences [7] and 89% in animal science [4]). The dependence structure can be broadly classified into clustered and correlated structures (with some cases having a mixture of both) [4,9]. Clustered structures arise when the true effect sizes are clustered within a broader variable (e.g., the same study and species), while correlated structures arise when the sampling errors are correlated (e.g., longitudinal studies). Failure to account for statistical dependence can lead to underestimated standard errors and high Type I error (false positive) rates [3,4,19]. Fortunately, advanced statistical frameworks such as linear mixed-effects models [20-22] and structural equation models [23-26] have allowed for the development of extensions that can model dependent effect sizes. For example, the multilevel model, incorporating clusters as random effects, can handle the various sources of dependency [17,27], whereas the multivariate model, incorporating correlated random effects and errors, can deal with correlated dependency [4,28,29]. In some disciplines, the multilevel meta-analysis (MLMA) model has become a standard (benchmark) method for dealing with dependent effect sizes due to its flexible random-effects structure [4,7,17,27]. Meanwhile, a new method called cluster-robust variance estimation (CRVE) is receiving more attention [30,31], which can handle statistical dependence without knowing the exact nature of both clustered and correlated dependency structures [32].

Another common feature is selective reporting, which can bias parameter estimation. A well-known example of selective reporting is publication bias, which occurs when there is a tendency to publish only statistically significant findings (Figure 1) [33,34]. Publication bias can upwardly bias the point estimate of the population mean effect [35]. Unfortunately, publication bias is pervasive across scientific disciplines [36-40]. Given the high heterogeneity within ecological and evolutionary meta-analyses [41], RE models (and their MLMA extensions) are sensitive to publication bias because the typical inverse-variance weighting scheme can give equal weight to studies. This process results in less powerful / precise studies contributing more strongly to mean effect estimates, exaggerating the bias driven by publication bias [42,43]. The FE model, despite its tendency to inflate Type I error rates, has the advantage of being less susceptible to publication bias compared to the

RE/MLMA models [44,45]. In addition to the FE model, three other models that are often overlooked, but show resistance to publication bias, are unrestricted weighted least squares (UWLS) [44,46,47], inverse variance heterogeneity (IVhet) model [48], and the Henmi-Copas model [44]. Furthermore, there are *post-hoc* methods available that specifically aim to correct publication bias. These methods have pros and cons in mitigating the overestimation due to selective reporting [49,50].

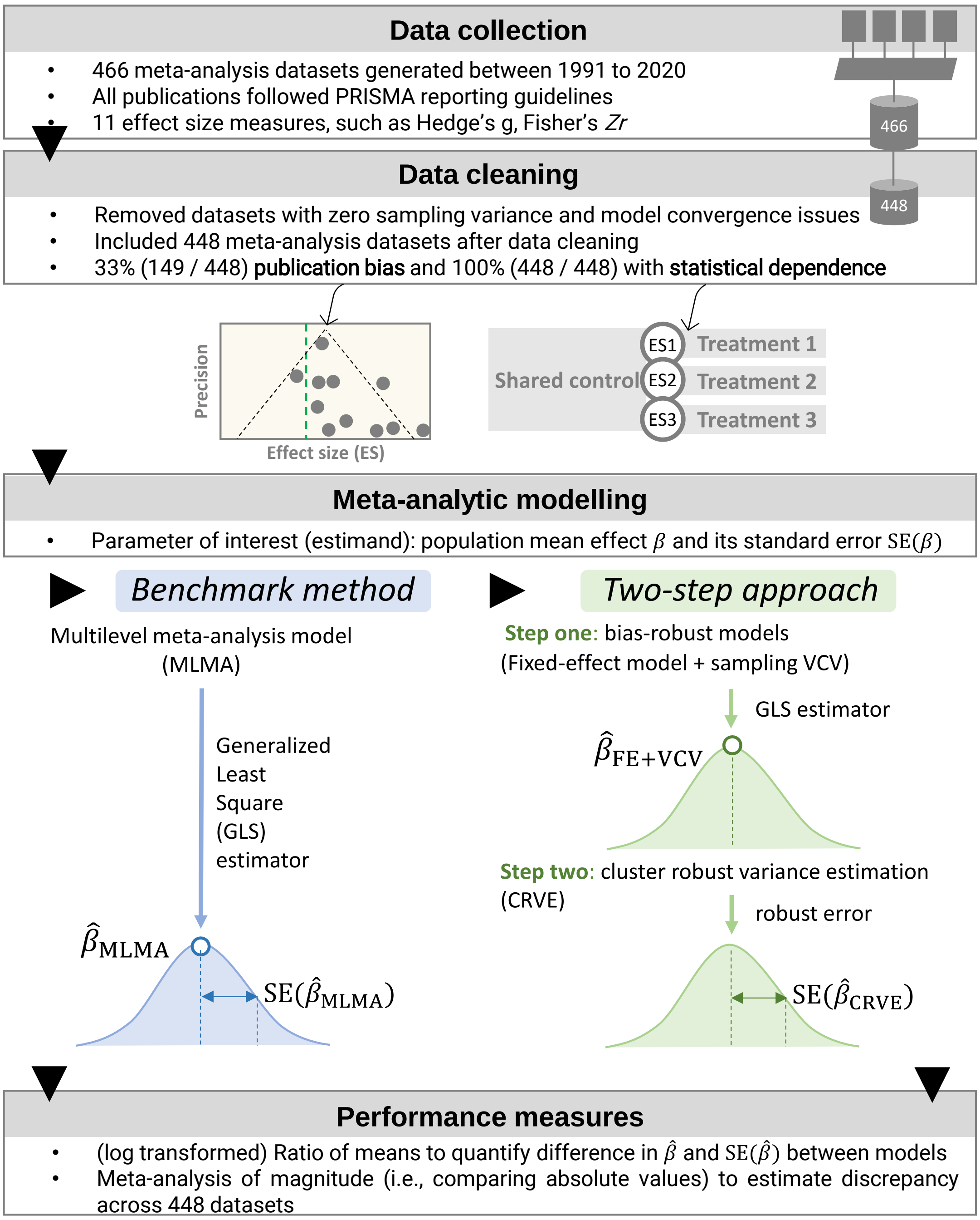

Figure 1. Workflow for data compilation, statistical modelling processes, and performance criteria.

When dealing with complex, yet realistic, ecological and evolutionary data structures, sophisticated methods, such as multilevel and multivariate models, can provide better statistical inference for dependent effect sizes. However, such models may overestimate the population mean effect size when publication bias is present because they weigh studies roughly equally in the presence of high heterogeneity. On the other hand, simple methods that are less vulnerable to publication bias, such as FE, UWLS, IVhet, and Henmi-Copas models [51-53], underestimate the standard error, inflating Type I error rates. This is because these simple models completely ignore the dependence structure (or, more precisely, have a mis-specified variance-covariance structure). Yet, few methods are currently available that can address statistical dependence and mitigate the impact of publication bias on meta-analyses simultaneously.

A potential solution to the problem is to separate out weighting schemes for addressing publication biases from estimating standard errors. To achieve this, we propose an effective and easy-to-implement two-step procedure (Figure 1). Below we explain the rationale of the proposed approach. Mathematical details can be found in the **Methods**, where we also demonstrate that the proposed approach is a generalized framework for existing models that can counteract selective reporting. We develop a helper function in the R package `orchaRd` [54] that can be used to visualise the impacts of publication bias on the population mean effect, providing a new graphical solution for visualising the impact publication bias has on parameter estimation and inference which can be used as a sensitivity analysis. Furthermore, our tutorial using the open-source software R, shows how to visualise and implement the proposed approach (https://yefeng0920.github.io/BiasRobustMA_tutorial/).

Results

The multilevel meta-analysis (MLMA) model, which is a standard method to handle statistical dependence [4,7,9,17,27], shows systematic errors in estimating the population mean effect size in the presence of selective reporting. Our proposed two-step approach addressed this issue by employing bias-robust models within the cluster-robust variance estimation (CRVE) framework (Figure 1). In the first step, we employed bias-robust models with bias-robust weighting schemes to obtain less biased population mean effect size estimates $\hat{\beta}$. Specifically, we incorporated a within-study variance-covariance matrix into the fixed-effect model (FE + VCV) and used the unrestricted weighted least square (UWLS) model (for details, see the **Methods**). The bias-robust weighting schemes counteracted selective reporting by considering the underlying mechanisms that contribute to it. For example, using the inverse sampling VCV weighting scheme can downweigh smaller studies with low precision, thereby penalizing studies that appear to be "selectively reported".

Moving to the second step, we treated the fitted bias-robust models as the "working" model within the CRVE framework (Figure 1). This step helped mitigate potential biases in the standard error estimates $\mathrm{SE}(\hat{\beta})$ that could arise from violating the assumption of data independence in FE + VCV and UWLS (i.e., model misspecification). By employing the CRVE, we obtained robust standard error estimates $\mathrm{SE}(\hat{\beta})$ that ensured the validity of subsequent statistical inference, including null-hypothesis tests and confidence interval (CI) construction. We compared the performance of the MLMA model with the proposed method. The comparison was carried out on 448 ecological and evolutionary meta-analyses (Figure 1 and Supplementary data 1; details see Methods). The differences in the estimated $\hat{\beta}$ and $\mathrm{SE}(\hat{\beta})$ between models were used as performance criteria (Figure 1). To quantify the overall discrepancies between models across 448 meta-analyses, we employed the meta-analysis of magnitude technique, which involves comparing absolute values based on folded distribution [39,55].

Bias of the population mean effect size estimates $\hat{\boldsymbol{\beta}}$

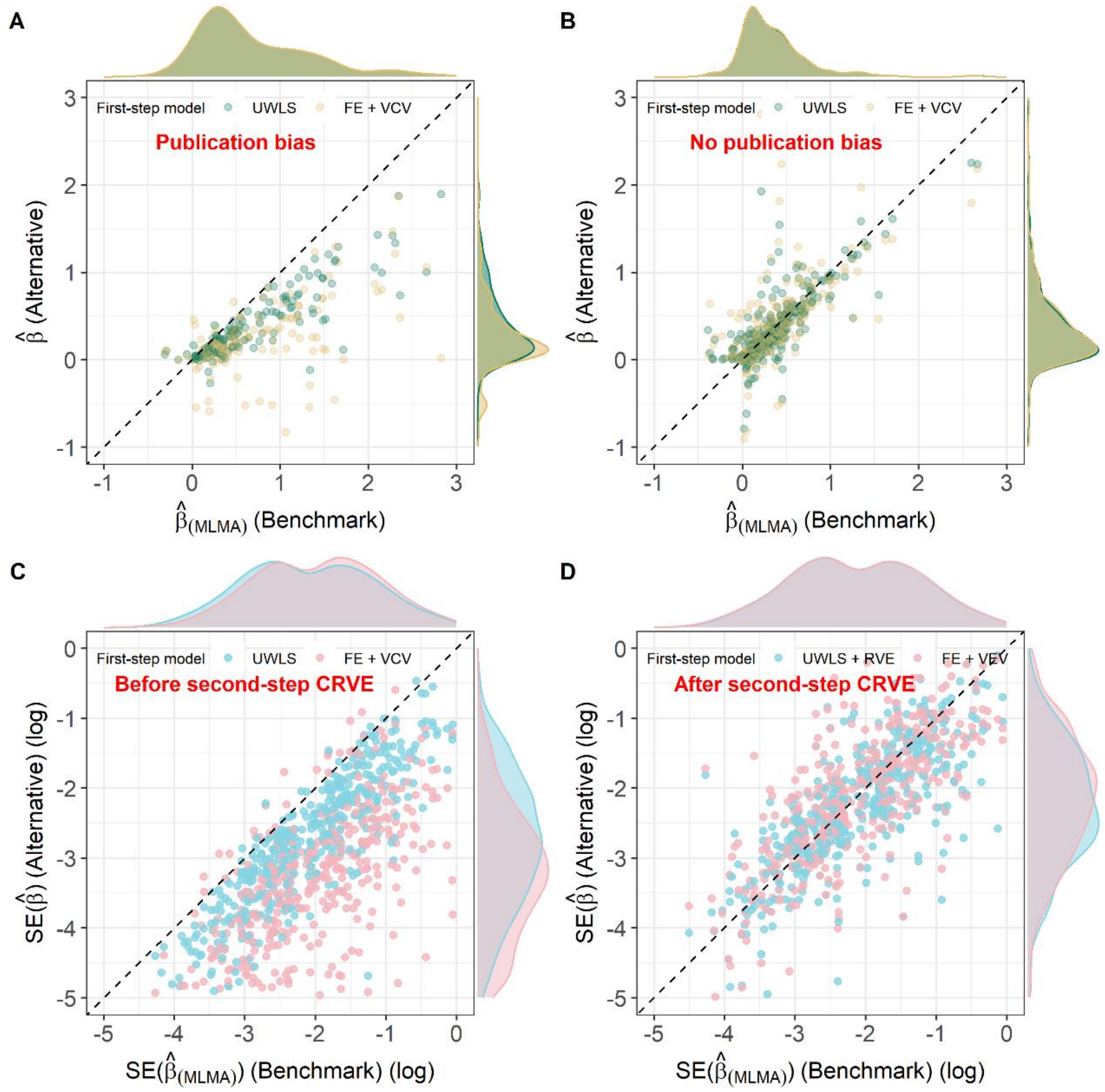

Figure 2. Comparison of the population mean effect size estimates $\hat{\beta}$ and their standard errors SE($\hat{\beta}$) obtained from different meta-analytic models. The dashed diagonal line is $y = x$, indicating where $\hat{\beta}$ and SE($\hat{\beta}$) would fall if there were no discrepancies between the models. The benchmark method, MLMA (multilevel meta-analysis model), is considered a standard method for addressing statistical dependence. The UWLS (unrestricted weighted least squares) and FE + VCV (fixed-effect model with

a within-study variance-covariance matrix) models are the first-step procedure of the proposed approach (i.e., bias-robust models), which employs bias-robust weighting schemes to mitigate the impact of publication bias (A and B). The publication bias was detected using the multilevel version of Egger's test with a significance level of $\alpha = 0.05$. Cluster robust variance estimation (CRVE) is the second step of our approach, which will address statistical dependence and compute the robust standard error $\mathrm{SE}(\hat{\beta})$ for subsequent statistical inference (C and D).

When the ecological and evolutionary meta-analyses were identified as having publication bias at the nominal significance level ($\alpha = 0.05$), the mean effect sizes $\hat{\beta}$, estimated by the benchmark method (MLMA model), consistently exceeded those obtained from the first step of our proposed method (UWLS and FE + VCV) (Figure 2).

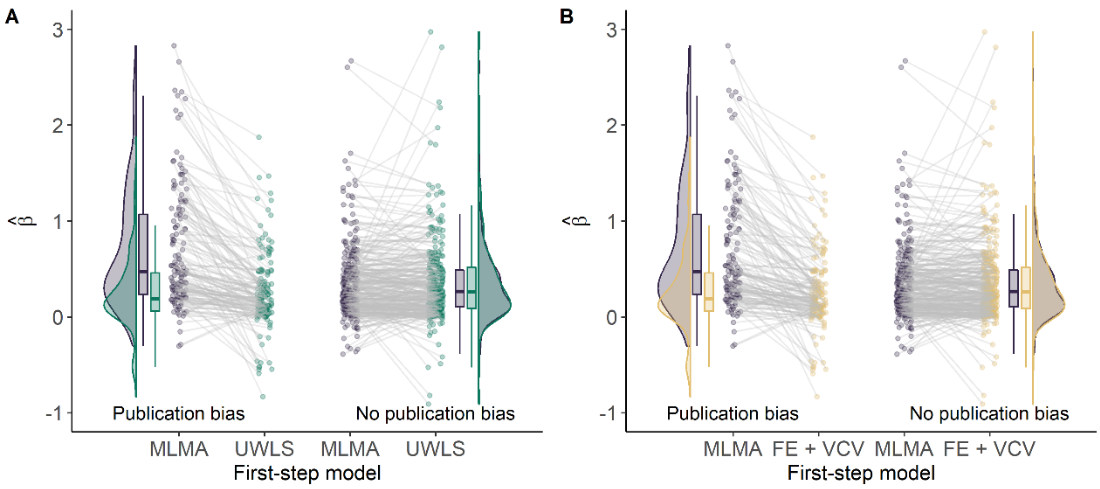

Figure 3. Paired comparison of the estimated population mean effect size estimates $\hat{\beta}$ obtained from the benchmark method (MLMA) and the first-step procedure of the proposed approach (UWLS and FE +VCV). For other details, refer to the legend in Figure 2.

Specifically, out of the 149 meta-analyses with publication bias, 141 exhibited an overestimated $\hat{\beta}$ when using the MLMA model (Figure 1), indicating 95% of meta-analyses exaggerated their mean effects if not employing any correction for the publication bias. On average, $\hat{\beta}$ derived from the MLMA model was 114.4% (0.763, 95% CI = [0.690,0.835]; Table S1) larger than that from the UWLS model and 155.3% (0.937, 95% CI = [0.848,1.026]) larger than that from the FE + VCV model (Figure 3). In contrast, for the meta-analyses without publication bias, there were no systematic differences in the estimated $\hat{\beta}$ among different models. Importantly, the magnitude of the discrepancy in the $\hat{\beta}$ between the models was statistically positively associated with the severity of publication bias (0.039, 95% CI = [0.001, 0.077]; Table S2; Figure 4). Similar patterns were observed when considering publication bias at a significance level of $\alpha = 0.1$ (see Figure S1 – S3).

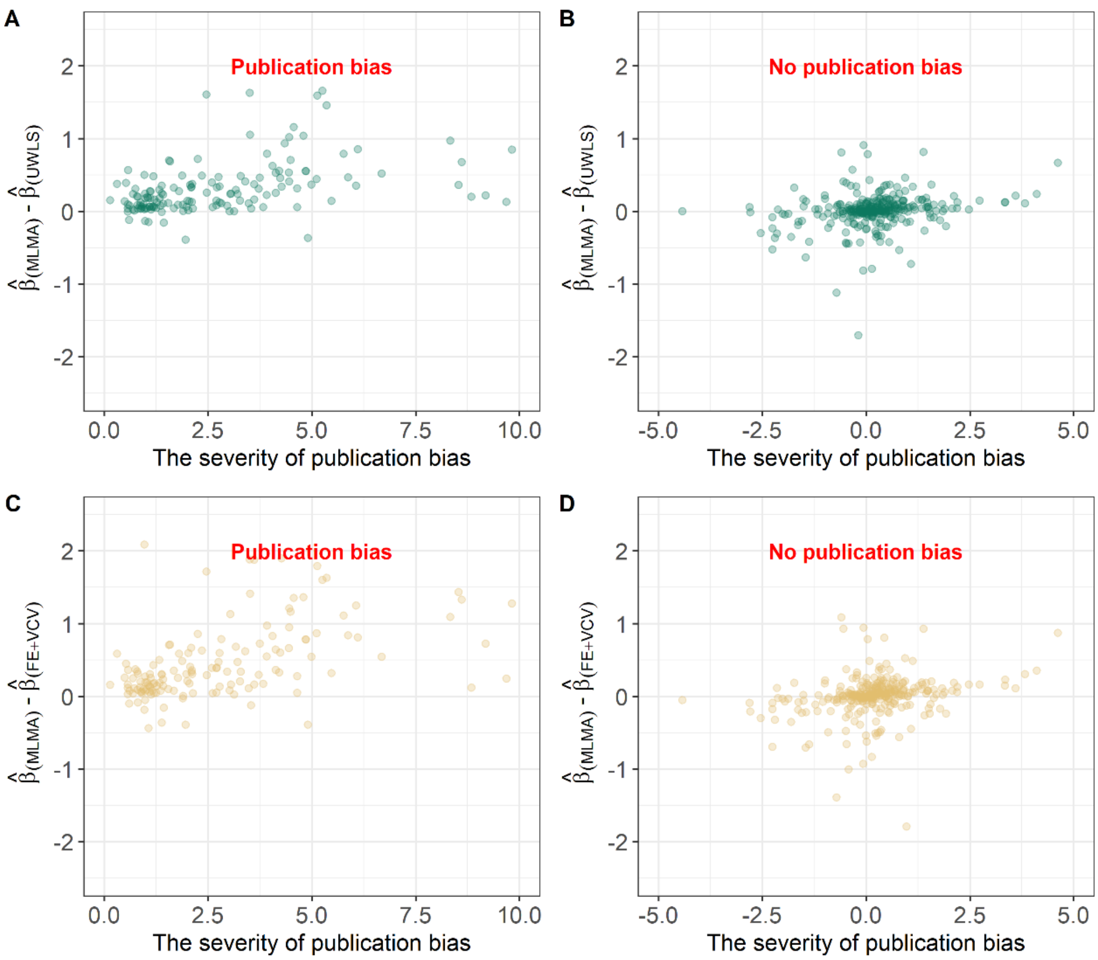

Figure 4. The relationship between the difference in the population mean effect size estimates $\hat{\beta}$ and the severity of publication bias.

Bias of the standard error estimates SE($\hat{\beta}$)

Before applying the second step of the proposed method, UWLS and FE + VCV consistently underestimated the SE($\hat{\beta}$) analysing meta-analyses with statistical dependence (Figure 2C). Among the 448 published meta-analyses analysed, 428 showed underestimation of SE($\hat{\beta}$) (Figure 2C), indicating that statistical inference of 96% of the published meta-analyses would be distorted if they were based on underestimated standard errors.

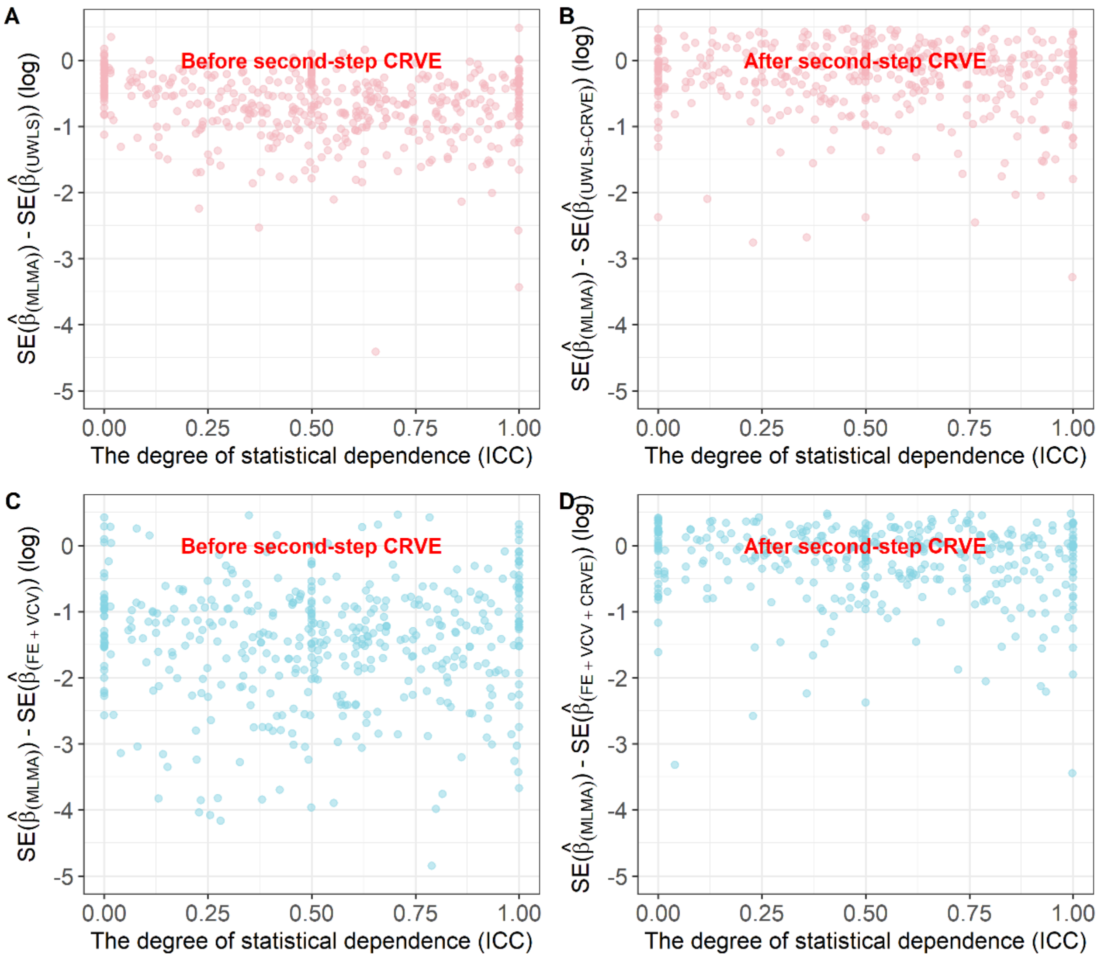

Figure 5. The relationship between the difference in the estimated standard error of the mean effect sizes SE($\hat{\beta}$) and the indicator of degree of statistical dependence. Intraclass correlation coefficient (ICC) represents the degree of dependence among effect sizes in a meta-analysis. For other details, refer to the legends in Figures 2 and 4.

On average, the $SE(\hat{\beta})$ derived from the UWLS and FE + VCV models were 123.3% (0.804, 95% CI = [0.736, 0.871]) and 382.3% (1.573, 95% CI = [1.469, 1.678]) smaller, respectively, than that obtained from the MLMA model when considering datasets with statistical dependence (Figure 2C and Table S3). After applying the second step of our proposed method, CRVE, no systematic difference was found in the estimated $SE(\hat{\beta})$ among different models (Figure 2D). There was no statistically significant association observed between the magnitude of the discrepancy in the estimated $SE(\hat{\beta})$ between the UWLS, FE +VCV and the MLMA models and the degree of statistical dependence (0.152, 95% CI = [-0.067, 0.371] and -0.108, 95% CI = [-0.445, 0.230], respectively; Figure 5).

Implementation and visualization: a real example

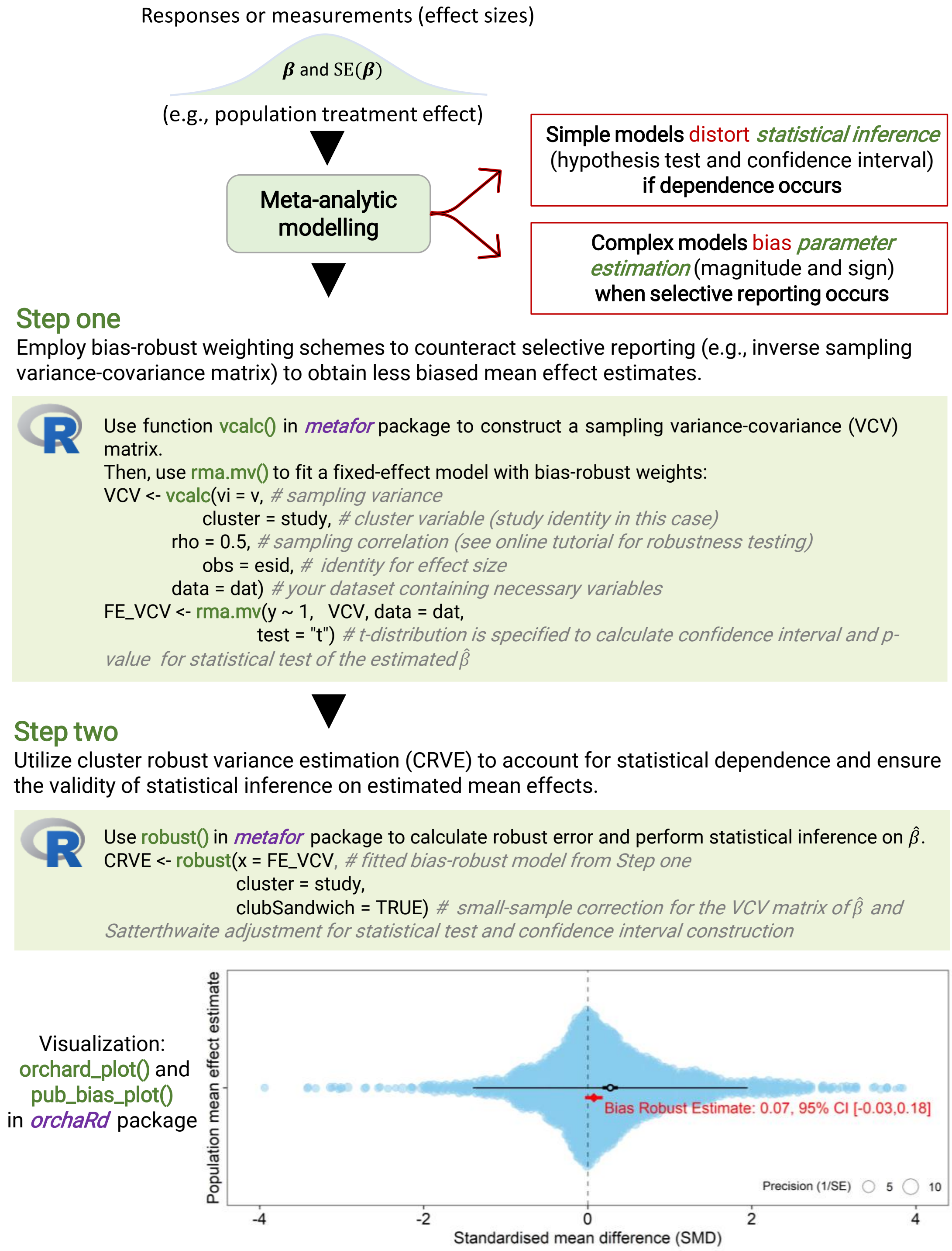

Figure 6. The implementation of the proposed two-step meta-analytic modelling approach, along with a novel visualization, using open-source software. The `metafor` [56] package is utilized for building two-step models, while the `orchaRd` [54] package is employed for graphical representation. For a detailed step-by-step tutorial demonstrating two alternative first-step models, please refer to the webpage (https://yefeng0920.github.io/BiasRobustMA_tutorial/). The bottom graph provides a visual solution to

show publication-bias-robust parameter estimation and inference derived from the proposed two-step approach, serving as a sensitivity analysis. It also includes essential elements for transparent reporting, such as 95% confidence intervals (CIs), and 95% prediction intervals (PIs), and precision (inverse standard error).

We created a tutorial to allow practitioners to apply the proposed approach. We selected a published ecological and evolutionary meta-analysis that claimed the existence of publication bias. This example meta-analysis examined the effect of herbivore interaction on fitness based on 179 species, 167 studies, and 1640 effect sizes [57]. In line with the publication bias test conducted in the original publication, our re-analysis also confirmed evidence for publication bias (0.876, 95% CI = [0.623, 1.128], see Figure S5 in tutorial). To address the statistical dependence among effect sizes (with 10 effect sizes/study), the original publication employed Bayesian multilevel meta-analytic modelling with phylogenetic relatedness, study, and observation identities as random effects.

To implement the first step of the proposed method, we first used the `vcalc()` function in the `metafor` package [56] to construct a sampling VCV matrix (Figure 6), assuming a constant sampling correlation of 0.5 [18]. Then, we used the `rma.mv()` function to fit a fixed-effect model using the constructed VCV matrix (FE + VCV model), obtaining the publication-bias-robust population mean effect of herbivore interaction. In the second step, we used the `coef_test()` in `metafor` (or `coef_test()` in the `clubSandwich` [58]) to address the misspecified dependence structure in the fitted FE + VCV model. This allowed us to compute robust standard errors and perform statistical inference on the interaction between herbivores. The result of the proposed approach indicates a minimal interaction between herbivores ($\hat{\beta}$ = 0.075, SE($\hat{\beta}$) = 0.054, 95% CI = [-0.034, 0.184], t_{52} = 1.375, *p*-value = 0.175;

Figure 6). In contrast, the original publication reported a statistically significant interaction between herbivores ($\hat{\beta}$ = 0.275, SE($\hat{\beta}$) = 0.047, 95% CI = [0.181, 0.368], t_{166} = 5.783, p-value < 0.001). We developed the `pub_bias_plot()` function, which can be used in conjunction with the `orchard_plot()` function from the `orchaRd` package [54]. Their combined use provides a new graphical plot for visualising the impact of publication bias on parameter estimation and inference (Figure 6), allowing for a visual assessment of the robustness of the meta-analytic findings and facilitating transparent reporting.

Discussion

To correctly estimate population mean effect sizes and make inferences when statistically dependent data [7,9] and selective reporting are present in ecology and evolution [36,51], we propose a readily implementable two-step approach. When selective reporting is present, the bias-robust models, used in the first step, consistently yielded less biased population mean effects compared to the standard multilevel meta-analysis model (MLMA; Figure 2). Population mean effect sizes were overestimated by an average of 110% when using the MLMA when publication bias was present (Figure 3). The severity of publication bias was also positively associated with the discrepancy in population mean effect estimates between the two models (Figure 4). Using a fixed-effect (FE) model with an adjusted VCV sampling matrix is effective at mitigating the impact of publication bias, particularly in cases where publication bias is more severe. Including cluster-robust variance estimation (CRVE) as the second step successfully addressed the issue of statistical dependence and achieved comparable estimates of standard errors across the models (Figure 2). On average, the CRVE corrected the estimates of standard errors by 120%. If these underestimated standard errors were to be used for statistical inference, it would lead to artefactually narrower CIs, increasing Type I error (false positive) rates in ecological and evolutionary meta-analytic evidence. Below, we discuss the innovations, extensions, limitations, and future perspectives of the proposed method.

Extending the tools for ascertaining the impacts of publication bias in ecological and evolutionary meta-analysis

Despite dedicated methods being available for adjusting selective reporting, including the fail-safe-*N* method, trim-and-fill, selection models and *p*-curve methods, these are not easily applied to complex hierarchical data. Failing to address statistical dependence (see Figure 7 in Nakagawa et al [50]) can result in unreliable publication-bias-corrected mean effect size estimates. Recently proposed modified Egger's approaches [50] show promise in addressing publication bias with complex hierarchical data by regressing effect size estimates against their sampling variances (also known as PEESE [51]), while controlling for statistical dependence. The intercept in the PEESE model can be interpreted as the publication-biased-adjusted mean effect, assuming an infinitely large study. The publication-biased-adjusted mean effect represents a marginalized mean in the context of a regression model, after accounting for impacts from informative predictors under infinite precision (or sample size). Unfortunately, informative predictors are rarely fully identified in published studies [39]. Additionally, extrapolation involved in marginalizing predictor effects may yield poor estimates of publication-biased-adjusted effect size and could also affect the magnitude of standard errors (Figure 6), influencing statistical inference. In contrast, the proposed two-step method does not rely on extrapolation and relaxes the assumption of infinite precision and information from predictors. Our proposed approach should be used as an effective sensitivity analysis to understand the effects of publication bias on the inferences drawn from a meta-analysis.

A general framework for incorporating weighting schemes

Our approach offers an effective way of addressing publication biases by separating out weighting schemes for point estimates from the estimation of robust standard errors. While the default inverse variance weighting scheme provides the best linear unbiased estimator of model coefficients only for representative data, it causes bias to the point estimate of the population mean effect (Equations 12 – 14). In contrast, the inverse sampling VCV weighting schemes used in the first step have prioritized reducing the bias of mean effects when the data is not representative (as in the case of selective reporting. Technically, the FE + VCV approach borrows from principles established under multivariate models but does not require one to parametrize random-effects structures [4], as shown in Equation 14 (for mathematical details, see the Methods). The CRVE separates the choice of weighting scheme from the estimation of standard errors and uses the residual distribution (see Equation 15) to approximate the true error distribution. Statistical inferences relying on robust standard errors remain valid under any weighting schemes [19,30,31,58,59]. This is particularly relevant when using bias-robust weighting schemes, as in our proposed approach, where the assumed error distribution deviates from the true distribution. Remedies can be made to address the issue of the small sample size [31]. These include using adjustment matrices (e.g., CR1, CR2 and CR3; Equation 15), robust-wild bootstrapping techniques and adapting the degrees of freedom for statistical tests of model coefficients [31,59,60].

More broadly, our method of tailoring and reconceptualising weighting can be extended to deal with three major sources of bias pertaining to meta-analytic weights. The first source of bias stems from questionable research practices. This includes selective reporting, such as publication bias, which has been addressed in the present study. Of relevance, poor quality studies or studies with a risk of bias can be potentially handled by using meta-analytic quality models within the framework CRVE [61], where study quality information is incorporated into the weighting schemes.

The second type of bias arises from using small sample sizes in primary studies. The proposed two-step approach can help ameliorate the impacts of small sample sizes on the estimation of sampling variances for various effect size measures, which, in turn, increases the accuracy of the weights converted from sampling variances. Current formulas used to compute sampling variance for common effect size measures (e.g., SMD, log response ratio, Fisher's *Zr*) are derived under the assumption of large sample sizes [14]. To mitigate this issue, one straightforward solution is to average the effect size estimates across all included studies and use this average to calculate the sampling variance (the so-called smooth estimator) [62]. As the number of studies increases, the averaged effect size converges to the true underlying effect. This improvement in handling small sample sizes also offers two additional benefits. It allows for the inclusion of primary studies with missing standard deviations [62] and enables data imputation for those missing standard deviations [63].

The third type of bias originates from the statistical properties of certain effect size measures, where the point estimate intrinsically correlates with its sampling variance. This is particularly relevant for effect size measures such as standardized mean difference (SMD), partial correlation coefficient (PCC), and log odds ratio [50,64]. To address this inherent correlation and mitigate bias, effective-sample-size-based or unit weighting schemes have been proposed [64]. However, these weighting schemes are not easily extendable to the framework of multilevel models, which are essential for accounting for statistical dependence and avoiding inflated Type I error rates.

Limitations and future opportunities

Two potential limitations should be noted about our proposed two-step approach for dealing with publication biases. The first potential limitation concerns using small-study effects as indicators of selective reporting or publication bias. While small-study effects, where studies with small sample sizes and large sampling errors tend to report large effect sizes [33,34], have been recognized as a typical signal of publication bias, we note that it is essentially a statistical representation of the asymmetry of a funnel plot [50]. Of note, Egger's test has limited statistical power to identify small-study effects under some circumstances [50]. Therefore, we conducted sensitivity analysis by relaxing the significance level to 0.1. The results from the significance level of 0.1 aligned with that from the nominal level of 0.05 (Figure S1 – S3), further supporting the effectiveness of the proposed approach in addressing selective reporting.

Second, we did not employ a simulation approach for assessing the empirical performance of statistical models. The reason for this decision was the limited quantitative knowledge available regarding the dependence structure and patterns of publication bias in real-world scenarios [49]. Designing simulations that accurately reflect these characteristics can be challenging. Our study leveraged the richness of published meta-analyses, which are more likely to capture the diverse range of dependence structures and publication bias patterns in real-world settings (see Figure 1 and the Methods). An extensive simulation would still be valuable in the future because our study specifically focused on the intercept-only meta-analysis models. While meta-regression models with categorical predictors can be transformed into subgroup intercept-only models, further investigation through simulations is needed to assess the generalizability of the proposed method to meta-regression models. Such a simulation study was out of the scope of the current work, as our focus was to establish a sensitivity analysis, which is robust against selective reporting, for multilevel meta-analytic models (i.e., estimating an overall effect).

Conclusion remarks

The development of meta-analytic models has strived to use sophisticated model structures to capture the underlying ecological and evolutionary data-generation process. In contrast, the proposed two-step approach takes a different perspective by prioritizing the adjustment for selective reporting in the first step and ensuring the validity of subsequent statistical inference by CRVE in the second step. This shift in focus emphasizes the development of appropriate weighting strategies to reduce bias in meta-analytic evidence when the data is not representative, which is often the case in ecology and evolution. We emphasize that the proposed approach is not intended to replace standard meta-analytic models (i.e., MLMA) in ecology and evolution. The performance of different models is contingent upon the true data-generation mechanisms that are rarely known. Therefore, we expect that the proposed approach serves as a sensitivity analysis to standard methods when interested in population mean effect.

In alignment with the move towards “multiverse” analytical workflows [65], we advocate for the routine use of our two-step method and its associated graphical tool as a sensitivity analysis (Figure 6). Complex hierarchical dependency structures and publication biases are typical of ecological and evolutionary meta-analyses. As such, it is becoming increasingly critical to explore and present multiple plausible analyses instead of relying solely on a single model (i.e., multiverse meta-analytic modelling) [66]. A more detailed assessment of the robustness of meta-analytic models would improve transparency and could be used to strengthen the meta-analytic evidence that is necessary to build the quantitative evidence that underpins ecological and evolutionary research and decision-making.

2 Methods

2.1 Dataset compilation

The dataset used in our study consisted of 448 ecological and evolutionary meta-analyses that were gathered by Costello and Fox [67] who followed systematic search principles to identify papers indexed in Web of Science Core Collection. All meta-analyses included claimed adherence to the PRISMA reporting guidelines (Preferred Reporting Items for Systematic Reviews and Meta-Analyses [68]). We further performed data cleaning to suit our analysis. Specifically, we eliminated cases with zero sampling variance, classified effect size measures into four categories (SMD family [i.e., Cohen's d and Hedge's g], log response ratio [lnRR], Fisher's *Zr*, and uncommon measures [i.e., mean difference, regression slope, risk ratio, and odds ratio]) [69]. We dropped meta-analysis datasets that had convergence issues in model fitting, despite adjusting different numerical optimizers and optimization specifications (i.e., number of iterations, step size and threshold). After the cleaning process, 448 meta-analysis datasets were included. Studies within meta-analyses reported eight effect size estimates on average, which indicates effect sizes were often clustered (ICC = 0.52) and possess substantial statistical dependency. About 33% (149/448) of the meta-analysis datasets showed evidence of publication bias based on the recently proposed multilevel version of Egger's test with a significance level of $\alpha = 0.05$ (Supplementary data 2), while 37% (166/448) of the meta-analysis datasets showed evidence of publication bias at $\alpha = 0.1$ [50].

2.2 Generalized least square (GLS) estimation for meta-analytic models

To make our article mathematically rigorous, we provide a brief revisit to the key statistical framework and estimators in the context of meta-analytic modelling. This also serves as a refresher for readers who are already familiar with these theories, allowing them to skip ahead to the subsequent sections.

2.2.1 Meta-analytic model via linear mixed-effects model framework

Consider a meta-analytic dataset with statistical dependence where J primary studies are included and n_j effect size estimates y_{ij} and sampling variances s_{ij}^2 can be derived from the j-th primary study (where $i = 1, \ldots, n_j$ and $j = 1, \ldots, J$). Let $\mathbf{x}_{ij}$ be a row vector of p predictors (also known as, covariates or moderators) that induce systematic variations among the effect size parameters (true effects), thus being treated as fixed effects in the frequentist framework. Likewise, let $\mathbf{z}_{ij}$ be a row vector of q predictors that lead to random variations among the effect size parameters and are therefore considered random effects. Using the (generalized) linear mixed-model framework [21,70], the FE, RE models and their more complex variants can be unified as a general form with:

$$\mathrm{y}_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\mathrm{b}_{ij} + \mathrm{e}_{ij}, \tag{1}$$

where $\boldsymbol{\beta}$ denote the vector of model coefficients for fixed-effects predictors $\mathbf{x}_{ij}$, representing the change in the (predicted) y_{ij} resulting from each one-unit change in $\mathbf{x}_{ij}$; b_{ij} denote the model coefficients for random-effects predictors $\mathbf{z}_{\mathrm{ij}}$, indicating (residual) variation in the effect size parameters; e_{ij} denote the sampling error corresponding to y_{ij}, with $\mathrm{E}(\mathrm{e}_{ij}) = 0$, $\mathrm{Var}(\mathrm{e}_{ij}) = \mathrm{s}_{ij}^2$, and $\mathrm{Cov}(\mathrm{e}_{ij}, \mathrm{e}_{hj}) = \rho_{ihj}\mathrm{s}_{ij}\mathrm{s}_{hj}$ (where ρ_j is the sampling correlation or within-study correlation between two paired effect size estimates y_{ij} and y_{hj}).

For the sake of brevity, we stack effect size estimates y_{ij} for each j cluster (in this case, study) and express Equation 1 in the matrix notation as Equation 2, which is known as a specification of seemingly unrelated regressions (SURs):

$$\mathbf{y}_\mathrm{j} = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\boldsymbol{b}_j + \mathbf{e}_j, \tag{2}$$

If we further stack SUR for all J studies, we can obtain a more compact notation:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{b} + \mathbf{e}, \tag{3}$$

where $\mathrm{E}(\boldsymbol{b}) = \mathbf{0}$, $\mathrm{Var}(\boldsymbol{b}) = \mathbf{G} = \mathbf{I}_J \otimes \mathbf{U} = \mathrm{diag}(\mathbf{U}, \mathbf{U}, \dots, \mathbf{U})$ (with $\otimes$ representing Kronecker product that creates a block-diagonal matrix), with $\mathbf{G}$ being a $Jq \times Jq$ block-diagonal matrix with $\mathbf{U}$ as diagonal elements, $\mathbf{U}$ being a $q \times q$ random-effects variance-covariance matrix that can be estimated via common estimators such as restricted maximum likelihood (REML); $\mathrm{Var}(\mathbf{e}) = \mathbf{S}$, with $\mathbf{S}$ being an $Jn_j \times Jn_j$ within-study (sampling) variance-covariance matrix [71]. Under the frequentist framework, Equation 3 can be expressed as a marginal form of $\mathbf{y} \sim MVN(\mathbf{X\beta}, \mathbf{ZGZ}' + \mathbf{S})$. Therefore, the $Jn_j \times Jn_j$ matrix $\mathbf{\Sigma} = \mathrm{Var}(\mathrm{Zb} + \mathrm{e}) = \mathbf{ZGZ}' + \mathbf{S}$ defines the marginal variance-covariance of the effect sizes and explicitly reflects the true dependence structure. Because the re-analysis dataset (448 meta-analysis) did not include predictor variables. Therefore, in the present study, our main focus was on the intercept-only meta-analysis model, where no predictors ($\mathbf{X} = \mathbf{1}$) are included to explain the variation in effect size estimates.

2.2.2 The efficient weighting scheme: minimum-variance unbiased estimator (MVUE)

Let $\mathbf{W}$ be an $Jn_j \times Jn_j$ weighting matrix. We can obtain the GLS estimator of fixed-effects coefficient $\boldsymbol{\beta}$ (i.e., population mean effect) by minimizing the mean squared error of model coefficient MES($\boldsymbol{\beta}$):

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\boldsymbol{X})^{-1}\mathbf{X}'\mathbf{W}\boldsymbol{y} = \left(\sum_{j=1}^{J}\mathbf{X}_j'\mathbf{W}_j\mathbf{X}_j\right)^{-1}\sum_{j=1}^{J}\mathbf{X}_j'\mathbf{W}_j\mathbf{y}_j\,, \tag{4}$$

The associated sampling variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$ can be estimated with:

$$
\begin{aligned}
\mathrm{Var}(\widehat{\boldsymbol{\beta}}) &= (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\boldsymbol{\Sigma}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} \\
&= \left(\sum_{j=1}^{J}\mathbf{X}_j'\mathbf{W}_j\mathbf{X}_j\right)^{-1}\sum_{j=1}^{J}\mathbf{X}_j'\mathbf{W}_j\,\boldsymbol{\Sigma}_j\mathbf{W}_j\mathbf{X}_j\left(\sum_{j=1}^{J}\mathbf{X}_j'\mathbf{W}_j\mathbf{X}_j\right)^{-1}, \qquad (5)
\end{aligned}
$$

Since $\mathrm{E}(\mathbf{y}_j|\mathbf{X}_j) = \mathbf{X}_j\boldsymbol{\beta}$, equation 4 provides an unbiased estimator of $\boldsymbol{\beta}$, when there is no publication bias, regardless of the specifications of the $Jn_j \times jn_j$ weighting matrix $\mathbf{W}$:

$$
\mathrm{E}(\widehat{\boldsymbol{\beta}}) = \left(\sum_{j=1}^{J}\mathbf{X}_j'\mathbf{W}_j\mathbf{X}_j\right)^{-1}\sum_{j=1}^{J}\mathbf{X}_j'\mathbf{W}_j\mathrm{E}(\mathbf{y}_j) = \left(\sum_{j=1}^{J}\mathbf{X}_j'\mathbf{W}_j\mathbf{X}_j\right)^{-1}\sum_{j=1}^{J}\mathbf{X}_j'\mathbf{W}_j\mathbf{X}_j\boldsymbol{\beta} = \boldsymbol{\beta}, \quad (6)
$$

To obtain a minimum-variance unbiased estimator of $\boldsymbol{\beta}$, we need to pick a weighting matrix that produces the minimum $\mathrm{Var}(\widehat{\boldsymbol{\beta}})$ among all unbiased estimators (in this case, best linear unbiased prediction BLUP). Based on the generalized Gauss-Markov theorem, setting $\mathbf{W}_j = \boldsymbol{\Sigma}^{-1}$leads to the unique solution to minimise $\mathrm{Var}(\widehat{\boldsymbol{\beta}})$:

$$
\begin{aligned}
\mathrm{Var}(\widehat{\boldsymbol{\beta}}) &= \left(\sum_{j=1}^{J}\mathbf{X}_j'\mathbf{W}_j\mathbf{X}_j\right)^{-1}\sum_{j=1}^{J}\mathbf{X}_j'\mathbf{W}_j\,\boldsymbol{\Sigma}_j\boldsymbol{\Sigma}_j^{-1}\mathbf{X}_j\left(\sum_{j=1}^{J}\mathbf{X}_j'\mathbf{W}_j\mathbf{X}_j\right)^{-1} \\
&= \left(\sum_{j=1}^{J}\mathbf{X}_j'\mathbf{W}_j\mathbf{X}_j\right)^{-1} = \left(\sum_{j=1}^{J}\mathbf{X}_j'\boldsymbol{\Sigma}_j^{-1}\mathbf{X}_j\right)^{-1}, \qquad (7)
\end{aligned}
$$

$\mathbf{W}_j = \boldsymbol{\Sigma}_j^{-1}$is the so-called inverse variance-covariance weights, which is the default weighting scheme used in the typical meta-analytic models. Using $\mathbf{W}_j = \boldsymbol{\Sigma}_j^{-1}$ as weights requires the knowledge of dependence structure $\boldsymbol{\Sigma} = \mathbf{Z}_j\mathbf{G}_j\mathbf{Z}_j' + \mathbf{S}_j$. Based on the above equations, efficient estimation of $\boldsymbol{\beta}$ and unbiased $\mathrm{Var}(\widehat{\boldsymbol{\beta}})$ (thus valid statistical inference) can be achieved when dependence structure is known, and there is no selective reporting in the dataset. Unfortunately, the prevalence of publication bias and statistical dependence can compromise the estimation of the two estimands, as outlined above. In the subsequent sections, we elaborate on the proposed two-step approach that can simultaneously ameliorate the impact of selective reporting and account for statistical dependence and.

2.3 Step one: GLS with a bias-robust weighting scheme

The first step of the proposed two-step approach is to employ a bias-robust weighting scheme that can counteract selective reporting. The small-study effect is a common form of selective reporting, where studies with small sample sizes and large sampling errors tend to report large effect sizes [33,34]. When this form selective reporting occurs, a criterion to alleviate its impact on the estimation $\boldsymbol{\beta}$ is to assign small studies with small weights. However, the default inverse variance-covariance weighting scheme $\mathbf{W}_j = \boldsymbol{\Sigma}_j^{-1}$ is incapable of accomplishing this criterion because it assigns near equal weight to each study if the heterogeneity is large [13]. In contrast, the inverse sampling variance-covariance $\mathbf{W}_j = \mathbf{S}_j^{-1}$ is a typical bias-robust weighting scheme that satisfies the criterion of assigning smaller weights to small studies [44]. It turns out that existing bias-robust meta-analytic models that are more tolerant to publication bias all adhere to this criterion, albeit with different assumptions about dependence structure $\boldsymbol{\Sigma} = \mathbf{Z}_j\mathbf{G}_j\mathbf{Z}_j' + \mathbf{S}_j$. Below we briefly illustrate four special cases of such models.

2.3.1 Fixed-effect (FE) model

It is well known that FE model assumes that the variances of the effect sizes are equal to the sampling variance, where diagonal elements of matrix $\mathbf{S}_j$ are sampling variances s_{ij}^2 and off-diagonal elements are sampling covariances $\rho_{ihj}s_{ij}s_{hj}$ with $\rho_{ihj} = 0$, indicating no correlation between the sampling errors [13]. FE model assumes there is no heterogeneity and thus $\mathbf{G} = \mathrm{diag}(\mathbf{0}, \mathbf{0}, \ldots, \mathbf{0})$ (between-study variance $\tau^2 = 0$). Consider an intercept-only model where $\mathbf{X}_j = \mathbf{1}$ (no predictors), $\mathbf{Z}_j = \mathbf{I}$ (random intercept), and $\mathbf{W}_j = \boldsymbol{\Sigma}_j^{-1} = \mathbf{S}_j^{-1}$. Hence, the estimator of $\boldsymbol{\beta}$ (Equation 4) and sampling variance $\mathrm{Var}(\widehat{\boldsymbol{\beta}})$ (Equation 5) reduce to

$$\widehat{\boldsymbol{\beta}}_{\mathrm{FE}} = \left(\sum_{j=1}^{J} \mathbf{1}_j' \mathbf{W}_j \mathbf{1}_j\right)^{-1} \sum_{j=1}^{J} \mathbf{1}_j' \mathbf{W}_j \mathbf{1}_j \mathbf{Y}_j\,, \tag{8}$$

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}_{\mathrm{FE}}) = \left(\sum_{j=1}^{J} \mathbf{1}_j' \mathbf{W}_j \mathbf{1}_j\right)^{-1}\,, \tag{9}$$

After simple matrix algebra, Equations 8 and 9 can be converted into typical summation expressions used in meta-analytic literature:

$$\hat{\beta}_{\mathrm{FE}} = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j} \mathrm{w}_{ij}\, \mathrm{y}_{ij}}{\sum_{j=1}^{J}\sum_{i=1}^{n_j} \mathrm{w}_{ij}} = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j} (1/\mathrm{s}_{ij})\, \mathrm{y}_{ij}}{\sum_{j=1}^{J}\sum_{i=1}^{n_j} (1/\mathrm{s}_{ij})}, \tag{10}$$

$$\mathrm{Var}\left(\hat{\beta}_{\mathrm{FE}}\right) = \frac{1}{\sum_{j=1}^{J}\sum_{i=1}^{n_j} \mathrm{w}_{ij}} = \frac{1}{\sum_{j=1}^{J}\sum_{i=1}^{n_j} (1/\mathrm{s}_{ij})}, \tag{11}$$

Note that if there is no statistical dependence ($\mathrm{n}_j = 1$), Equations 10 and 11 will collapased to normal estimators of FE model. After estimaing $\hat{\beta}_{\mathrm{FE}}$ and $\mathrm{Var}(\hat{\beta}_{\mathrm{FE}})$ from Equations 10 and 11, we need to perform statistical inference on the estiamted $\hat{\beta}_{\mathrm{FE}}$. One common method for this is the Wald-type test. Alternative methods are also available, such as likelihood ratio test and permutation test [72]. The Wald-type test involves comparing a test statistic t against critical values of a known distribution to test the null hypothesis H_0: $\beta = 0$. The test statistic t can be calcualted as $(\hat{\beta}_{\mathrm{FE}} - \beta)/\mathrm{SE}\left(\hat{\beta}_{\mathrm{FE}}\right)$, where $\mathrm{SE}\left(\hat{\beta}_{\mathrm{FE}}\right)$ is the square-root of $\mathrm{Var}(\hat{\beta}_{\mathrm{FE}})$. Under H_0, test statistic T follows (asymptotically) a standard normal distribution or t distribution with $df = Jn_j - 1$ degrees of freedom (adjustments to df are possible such as $df = J - 1$ [73-75]), which can be used to construct a confidence interval (CI) and calculate a p-value for the test. However, while Equation 10 can reduce the bias of $\hat{\beta}$ with respect to publication bias, Equation 11 apparently underestimates $\mathrm{Var}(\hat{\beta})$, inflated test statistic t, Type I error rate and p-value due to neglection of random-effects (i.e., misspecified dependence structure $\boldsymbol{\Sigma}_j^{-1} = \mathbf{Z}_j \mathbf{G}_j \mathbf{Z}_j' + \mathbf{S}_j = \mathbf{S}_j$).

2.3.2 Unrestricted weighted least squares (UWLS) model

The UWLS model is a weighted linear regression used to address heteroskedasticity (i.e., an unequal variance of observation) in ordinary least squares regression. The UWLS was initially conceptualized as a statistical model for meta-analysis in particle physics [47,76]. Stanley and his colleagues subsequently fleshed out its theories in meta-analysis and termed it as unrestricted weighted least squares [46,52,53]. Unlike FE model, UWLS model relaxes the assumption that sampling variance s_{ij} is precisely known without uncertainty, instead assuming that s_{ij} is only known up to a proportionality constant σ_e^2. This results in a weighting scheme of $\mathbf{W}_j = \boldsymbol{\Sigma}_j^{-1} = \sigma_e^2 \mathbf{S}_j^{-1}$ in the UWLS model. The estimator of $\hat{\beta}$ is identical to Equation 10 in FE model, while the estimator of $\mathrm{Var}(\hat{\beta})$ becomes:

$$\mathrm{Var}(\hat{\beta}_{\mathrm{UWLS}}) = \frac{1}{\sigma_e^2 \sum_{j=1}^{J} \sum_{i=1}^{n_j} (1/\mathrm{s}_{ij})}, \tag{12}$$

The term σ_e^2 technically refers to the residual variance or weighted mean squared error. This parameter can be estimated from data in contrast to the fixed value of 1 in FE and RE models, which is why the UWLS model is called the “unrestricted” WLS [46]. The term σ_e^2 is beneficial to statistical inference on $\hat{\beta}_{\mathrm{UWLS}}$ at two aspects. On the one hand, it represents the overdispersion of effect sizes and thus accounts for heterogeneity in a multiplicative way [46]. This is why UWLS is also known as the “multiplicative” method for meta-analysis. On the other hand, σ_e^2 can act as a scaling factor of $\mathrm{Var}(\hat{\beta}_{\mathrm{UWLS}})$ to account for the uncertainty in estimating $\boldsymbol{\Sigma}_j^{-1}$ and improve statistical inference [77]. The estimator of $\mathrm{Var}(\hat{\beta}_{\mathrm{UWLS}})$ in UWLS model (Equation 12) is mathematically equivalent to Hartung-Knapp-Sidik-Jonkman adjustment method [73-75].

2.3.3 Inverse variance heterogeneity (IVhet) model

The IVhet model is a meta-analysis model that assumes no heterogeneity (between-study variance $\tau = 0$) when estimating $\hat{\beta}$, but accounts for model overdispersion when estimating $\mathrm{Var}(\hat{\beta})$ using a quasi-likelihood-based variance structure [48]. It uses an intra-class correlation (ICC) as a scale parameter to inflate $\mathrm{Var}(\hat{\beta})$ derived from the FE model. This "overdispersion" strategy relaxes distributional assumptions with variance and purely relies on mean-variance relationship. Interestingly, there is an unappreciated model called Henmi-Copas model [44] that can counteract publication bias by using the FE model to estimate $\hat{\beta}$ and then centering the 95% CI derived from the RE model on this estimate. Technically, IVhet model is similar to Henmi-Copas model because both of them use $\hat{\beta}$ from FE model and incorporate estimated τ into $\mathrm{Var}(\hat{\beta})$ and the corresponding CI. Both models are essentially equivalent to the RE model with an inverse sampling variance weighting scheme, although they differ in their theoretical underpinnings and the estimators used to estimate τ^2. For the sake of illustration, consider an intercept-only model where $\mathbf{X}_j = \mathbf{1}$ (no predictors), $\mathbf{Z}_j = \mathbf{I}$ (random intercept), $\mathbf{W}_j = \mathbf{S}_j^{-1}$ and $\boldsymbol{\Sigma}_j = \tau^2\mathbf{I}_j + \mathbf{S}_j$. Hence, the estimator of β is identical to that in FE and UWLS models, while the estimator of sampling variance $\mathrm{Var}(\hat{\beta})$ (Equation 5) simplifies to:

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \left(\sum_{j=1}^{J} \mathbf{1}_j'\mathbf{S}_j^{-1}\mathbf{1}_j\right)^{-1} \sum_{j=1}^{J} \mathbf{1}_j'\mathbf{S}_j^{-1}\,(\tau^2\mathbf{I}_j + \mathbf{S}_j)\mathbf{S}_j^{-1}\mathbf{1}_j \left(\sum_{j=1}^{J} \mathbf{1}_j'\mathbf{S}_j^{-1}\mathbf{1}_j\right)^{-1}, \qquad (13)$$

Obtaining the summation expression by evaluating the above matrix expression:

$$\mathrm{Var}\left(\hat{\beta}_{\mathrm{IVhet}}\right) = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}(\frac{1}{\mathrm{s}_{ij}})^2(\tau^2 + s_{ij}^2)}{\left(\sum_{j=1}^{J}\sum_{i=1}^{n_j}(1/\mathrm{s}_{ij})\right)^2}, \tag{14}$$

Although, the estimator $\mathrm{Var}\left(\hat{\beta}_{\mathrm{IVhet}}\right)$ accounts for heterogeneity, it still underestimates the true $\mathrm{Var}\left(\hat{\beta}\right)$. Because it assumes that there is no correlation in the random-effects matrix $\mathbf{G}_j$ and sampling variance-covariance matrix $\mathbf{S}_j$ (zero off-diagonal elements).

2.3.4 FE model with a sampling variance-covariance matrix (FE + VCV)

Imagine that if the j-th study contributes more effect size estimates than the k-th study ($n_j > n_k$), then FE, UWLS and IVhet models tend to assign more weights to the j-th study. In other words, studies reporting more effect size estimates are assigned more weights, which would bias the estimates of β (i.e., pull the estimated β towards study with more effect size estimates) since effect size estimates from the same study do not represent independent information. Computationally, above models ignore the correlation between effect size estimates within the same study when estimating β. To address this, a straightforward approach is to incorporate this correlation when estimating β. This can be achieved by using a sampling variance-covariance matrix $\mathbf{S}_j$ with non-zero correlation ρ_{ihj} between sampling errors e_{ij} and e_{hj} (see Equation 1). For example, assume $\mathbf{S}_j$ having a compound symmetry structure given as $\mathbf{VCV}_j$ matrix:

$$\mathbf{VCV}_j = \begin{bmatrix} s_{1j}^2 & \cdots & \rho_{1n_jj}s_{1j}s_{n_jj} \\ \vdots & \ddots & \vdots \\ \rho_{1n_jj}s_{1j}s_{n_jj} & \cdots & s_{n_jj}^2 \end{bmatrix}, \tag{15}$$

In Equation 15, the off-diagonal element $\rho_{ihj}s_{ij}s_{hj}$ represents the covariance between effect size estimates, which can be used to adjust for multiple effect size estimates derived from the same study. Therefore, we propose using the FE model in combination with a sampling variance-covariance matrix with non-zero correlation (FE + VCV). In this way, the weighting scheme becomes $\mathbf{W}_j = \mathbf{\Sigma}_j^{-1} = \mathbf{VCV}_j^{-1}$ or

$\mathbf{W} = \operatorname{diag}\left(\mathbf{VCV}_1^{-1}, \mathbf{VCV}_2^{-1}, \ldots, \mathbf{VCV}_J^{-1}\right)$. According to matrix algebra, the off-diagonal elements of $\mathbf{VCV}_j^{-1}$ are negative values, which would de-emphasize the studies reporting more effect size estimates when estimating β. However, this approach still underestimates $\operatorname{Var}\left(\hat{\beta}\right)$ due to the same reason mentioned earlier. The FE model with a within-study VCV matrix is essentially a simplified version of the multivariate models that do not involve random effects (e.g., τ^2) and distinguish different outcomes [4].

2.4 Step two: cluster-robust variance estimation (CRVE)

The second step of the proposed two-step approach involves using CRVE to guard against misspecification in the marginal variance-covariance matrix $\mathbf{\Sigma}_j = \mathbf{Z}_j \mathbf{G}_j \mathbf{Z}_j' + \mathbf{S}_j$ (as seen in Equations 11, 12 and 14). In **Section 2.2.2 Minimum-variance unbiased estimator (MVUE)**, it was discussed that achieving the MVUE for β requires accurately specifying the matrix configurations of $\mathbf{\Sigma}_j$ for each study in the meta-analysis and using the inverse of $\mathbf{\Sigma}_j$ as the weighting scheme to minimize $\operatorname{Var}(\hat{\beta})$. The bias-robust weighting scheme reduces the bias of β at the expense of sampling variance $\operatorname{Var}(\hat{\beta})$. Fortunately, CRVE can provide consistent estimates of the population parameters even when the matrix $\mathbf{\Sigma}_j$ is misspecified [30]. In the context of meta-analysis, CRVE was introduced by Sidik and Jonkman [32], and Hedges and colleagues [30] formalized its use to account for misspecification in the marginal variance-covariance matrix of effect sizes. The challenge in estimating the sampling variance of the model coefficients β arises from the unknown $\mathbf{\Sigma}_j$ in Equation 7. In the framework of CRVE, the outer products of the residual vector $\hat{\mathbf{e}}_j = \mathbf{y}_\mathrm{j} - \mathbf{X}_j \widehat{\boldsymbol{\beta}}$ for j-th study can be used to empirically approximate the marginal variance-covariance matrix:

$\hat{\mathbf{e}}_j\hat{\mathbf{e}}_j' \approx \boldsymbol{\Sigma}_j$. Thus, the estimator of the sampling variance $\mathrm{Var}(\hat{\beta})$ becomes the so-called robust sandwich estimator:

$$\mathrm{Var}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{CRVE}}\right) = \left(\sum_{j=1}^{J} \mathbf{X}_j'\mathbf{W}_j\mathbf{X}_j\right)^{-1} \sum_{j=1}^{J} \mathbf{X}_j'\mathbf{W}_j\,\hat{\mathbf{e}}_j\hat{\mathbf{e}}_j'\mathbf{W}_j\mathbf{X}_j \left(\sum_{j=1}^{J} \mathbf{X}_j'\mathbf{W}_j\mathbf{X}_j\right)^{-1}, \qquad (16)$$

When the number of studies J is sufficiently large, it is reasonable to assume $\mathrm{E}(\hat{\mathbf{e}}_j\hat{\mathbf{e}}_j') \approx \boldsymbol{\Sigma}_j$. Therefore, $\mathrm{Var}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{CRVE}}\right)$ converges to the true sampling variance as $J \to \infty$ (see proof [30]). As a result, statistical inferences on the model coefficients are valid based on asymptotic inference theory. In cases where J is small, correction methods can be used to reduce the bias of robust standard errors and maintain valid statistical inferences. These methods include using small-sample adjusting matrices to estimated residuals $\hat{\mathbf{e}}_j$ [31,59] or employing robust-wild bootstrapping techniques [60].

2.5 Standard (benchmark) method: multilevel meta-analytic (MLMA) model

To assess the performance of the proposed two-step approach, we compare it with a benchmark method commonly used for handling dependent effect sizes, known as the multilevel meta-analytic (MLMA) model. The basic MLMA model is a three-level meta-analytic model that includes random-effects at the between-study and within-study levels. It can be expressed as:

$$\mathrm{Y}_{ij} = \beta + \mathrm{u}_{(b)j} + \mathrm{u}_{(w)ij} + \mathrm{e}_{ij}, \qquad (17)$$

where β denotes the model intercept, representing the mean effect size; $\mathrm{u}_{(\mathrm{b})j}$ is a random-effects at between-study level with $\mathrm{Var}\left(\mathrm{u}_{(b)j}\right) = \tau_b^2$, which captures between-study heterogeneity; $\mathrm{u}_{(b)j}$ is a random-effects at within-study level with $\mathrm{Var}\left(\mathrm{u}_{(w)j}\right) = \tau_w^2$, which captures within-study heterogeneity; e_{ij} is the corresponding sampling error, with $\mathrm{Var}(\mathrm{e}_{ij}) = \mathrm{s}_{ij}^2$ and $\mathrm{Cov}(\mathrm{e}_{ij}, \mathrm{e}_{hj}) = 0$. The MVUE for β and sampling variance $\mathrm{Var}(\beta)$ can be reached using Equation 4 and 7, with weights equal to inverse variance-covariances. The weights for MLMA for each study are written as (as in the `metafor` package) [15]:

$$\mathbf{W}_{j_{\text{MLMA}}} = \begin{bmatrix} \hat{\tau}_b^2 + \hat{\tau}_w^2 + s_{1j}^2 & \cdots & \hat{\tau}_b^2 \\ \vdots & \ddots & \vdots \\ \hat{\tau}_b^2 & \cdots & \hat{\tau}_b^2 + \hat{\tau}_w^2 + s_{1j}^2 \end{bmatrix}^{-1}, \tag{18}$$

Therefore, $\mathbf{W}_{\text{MLMA}} = \operatorname{diag}\left(\mathbf{W}_{1\text{MLMA}}, \mathbf{W}_{2\text{MLMA}}, \ldots, \mathbf{W}_{J\text{MLMA}}\right)$. The MLMA model offers several advantages, which have made it a standard benchmark method for dealing with dependent effect sizes in many disciplines [4,7,17,27]. First, its implementation is straightforward and easy. Second, it does not require the reporting of sampling correlations, which are often unavailable in the literature. Third, it performs well in terms of statistical inferences and provides extra insights into heterogeneity compared to multivariate models and CRVE. However, one unappreciated limitation of MLMA is that it uses inverse variance-covariance as the weighting scheme, which can lead to an overestimation of β when publication bias is present.

2.6 Performance criterion

The first criterion focused on the biasedness of the mean effects ($\hat{\beta}$) when no predictors were included in the models. We applied the MLMA model (benchmark) and the first step of our proposed method (two alternatives: UWLS and FE + VCV models) to each of the 448 meta-analysis datasets. The sampling correlation ρ_{ihj} was set to 0.5 for FE + VCV model [4,18]. The sensitivity analysis should be conducted to examine the extent to which the mean effect is sensitive to the assumption of within-study (sampling) correlation ρ_{ihj} values used for constructing VCV matrix (see tutorial: https://yefeng0920.github.io/BiasRobustMA_tutorial/)

We compared the mean effects $\hat{\beta}$ obtained from the MLMA model with those from the two alternatives by calculating the log transformed ratio of means as the effect size measure. Our expectation was that the magnitudes of $\hat{\beta}$ derived from the benchmark model would consistently be larger than those from our proposed method when publication bias was present. The second criterion addressed the biasedness of the standard errors of the mean effects $\mathrm{SE}(\hat{\beta})$, represented by the square-root of $\mathrm{Var}(\hat{\beta})$. The robust $\mathrm{SE}(\hat{\beta})$ was calculated by the second step of our proposed method, which involved applying the CRVE to the fitted UWLS, and FE + VCV models. We calculated the paired log-transformed ratio of $\mathrm{SE}(\hat{\beta})$ obtained from the MLMA model to robust $\mathrm{SE}(\hat{\beta})$ [78]. Theoretically, we predicted the $\mathrm{SE}(\hat{\beta})$ derived from MLMA model to consistently be larger than those from the UWLS and FE + VCV models, but similar to robust $\mathrm{SE}(\hat{\beta})$. We further computed the folded mean and sampling variance of the effect size using the folded distribution [39,55]. Finally, we used the meta-analysis of magnitude technique to assess the overall discrepancies between models [55].

Data availability

The raw data is available at https://github.com/Yefeng0920/BiasRobustMA_tutorial.

Code availability

The analytical script to reproduce examples presented in the manuscript is archived at GitHub: https://github.com/Yefeng0920/BiasRobustMA_tutorial. A webpage showing the implementation of the proposed method in combination with a visualisation tool can be accessed via https://yefeng0920.github.io/BiasRobustMA_tutorial/.

References

1 Cooper, H., Hedges, L. V. & Valentine, J. C. *The handbook of research synthesis and meta-analysis*. (Russell Sage Foundation, 2019).

2 Gurevitch, J., Koricheva, J., Nakagawa, S. & Stewart, G. Meta-analysis and the science of research synthesis. *Nature* **555**, 175-182 (2018).

3 Nakagawa, S., Noble, D. W., Senior, A. M. & Lagisz, M. Meta-evaluation of meta-analysis: ten appraisal questions for biologists. *BMC biology* **15**, 1-14 (2017).

4 Yang, Y., Macleod, M., Pan, J., Lagisz, M. & Nakagawa, S. Advanced nethods and implementations for the meta-analyses of animal models: current practices and future recommendations. *Neuroscience & Biobehavioral Reviews*, 105016 (2022).

5 Seidler, A. L. *et al.* A guide to prospective meta-analysis. *bmj* **367** (2019).

6 Egger, M., Smith, G. D. & Phillips, A. N. Meta-analysis: principles and procedures. *Bmj* **315**, 1533-1537 (1997).

7 Nakagawa, S., Yang, Y., Macartney, E. L., Spake, R. & Lagisz, M. Quantitative evidence synthesis: a practical guide on meta-analysis, meta-regression, and publication bias tests for environmental sciences. *Environmental Evidence* **12**, 8, doi:10.1186/s13750-023-00301-6 (2023).

8 Pigott, T. D. & Polanin, J. R. Methodological guidance paper: High-quality meta-analysis in a systematic review. *Review of Educational Research* **90**, 24-46 (2020).

9 Cheung, M. W.-L. A guide to conducting a meta-analysis with non-independent effect sizes. *Neuropsychology review* **29**, 387-396 (2019).

10 Hansen, C., Steinmetz, H. & Block, J. 1-19 (Springer, 2021).

11 Stanley, T. D. *et al.* Meta‐analysis of economics research reporting guidelines. *Journal of economic surveys* **27**, 390-394 (2013).

12 Andrews, I. & Kasy, M. Identification of and correction for publication bias. *American Economic Review* **109**, 2766-2794 (2019).

13 Borenstein, M., Hedges, L. V., Higgins, J. P. & Rothstein, H. R. A basic introduction to fixed‐effect and random‐effects models for meta‐analysis. *Research synthesis methods* **1**, 97-111 (2010).

14 Borenstein, M., Hedges, L. V., Higgins, J. P. & Rothstein, H. R. *Introduction to meta-analysis*. (John Wiley & Sons, 2021).

15 Viechtbauer, W. Accounting for heterogeneity via random-effects models and moderator analyses in meta-analysis. *Zeitschrift für Psychologie* **215**, 104-121 (2007).

16 Van Den Noortgate, W. & Onghena, P. Multilevel meta-analysis: A comparison with traditional meta-analytical procedures. *Educational and psychological measurement* **63**, 765-790 (2003).

17 Van den Noortgate, W., López-López, J. A., Marín-Martínez, F. & Sánchez-Meca, J. Three-level meta-analysis of dependent effect sizes. *Behavior research methods* **45**, 576-594 (2013).

18 Noble, D. W., Lagisz, M., O'dea, R. E. & Nakagawa, S. Nonindependence and sensitivity analyses in ecological and evolutionary meta‐analyses. *Molecular Ecology* **26**, 2410-2425 (2017).

19 Tanner‐Smith, E. E. & Tipton, E. Robust variance estimation with dependent effect sizes: Practical considerations including a software tutorial in Stata and SPSS. *Research synthesis methods* **5**, 13-30 (2014).

20 Kalaian, H. A. & Raudenbush, S. W. A multivariate mixed linear model for meta-analysis. *Psychological methods* **1**, 227 (1996).

21 Stram, D. O. Meta-analysis of published data using a linear mixed-effects model. *Biometrics*, 536-544 (1996).

22 Sheu, C.-F. & Suzuki, S. Meta-analysis using linear mixed models. *Behavior Research Methods, Instruments, & Computers* **33**, 102-107 (2001).

23 Cheung, M. W. L. Some reflections on combining meta‐analysis and structural equation modeling. *Research Synthesis Methods* **10**, 15-22 (2019).

24 Tu, Y.-K. & Wu, Y.-C. Using structural equation modeling for network meta-analysis. *BMC medical research methodology* **17**, 1-9 (2017).

25 Cheung, M. W.-L. Multivariate meta-analysis as structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal* **20**, 429-454 (2013).

26 Cheung, M. W.-L. Modeling dependent effect sizes with three-level meta-analyses: a structural equation modeling approach. *Psychological methods* **19**, 211 (2014).

27 Nakagawa, S. & Santos, E. S. Methodological issues and advances in biological meta-analysis. *Evolutionary Ecology* **26**, 1253-1274 (2012).

28 Gasparrini, A., Armstrong, B. & Kenward, M. G. Multivariate meta‐analysis for non‐linear and other multi‐parameter associations. *Statistics in medicine* **31**, 3821-3839 (2012).

29 Jackson, D., Riley, R. & White, I. R. Multivariate meta‐analysis: potential and promise. *Statistics in medicine* **30**, 2481-2498 (2011).

30 Hedges, L. V., Tipton, E. & Johnson, M. C. Robust variance estimation in meta‐regression with dependent effect size estimates. *Research synthesis methods* **1**, 39-65 (2010).

31 Pustejovsky, J. E. & Tipton, E. Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics* **36**, 672-683 (2018).

32 Sidik, K. & Jonkman, J. N. A note on variance estimation in random effects meta-regression. *Journal of biopharmaceutical statistics* **15**, 823-838 (2005).

33 Rosenthal, R. The file drawer problem and tolerance for null results. *Psychological bulletin* **86**, 638-641 (1979).

34 Thornton, A. & Lee, P. Publication bias in meta-analysis: its causes and consequences. *Journal of clinical epidemiology* **53**, 207-216 (2000).

35 Ioannidis, J. P., Munafo, M. R., Fusar-Poli, P., Nosek, B. A. & David, S. P. Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends in cognitive sciences* **18**, 235-241 (2014).

36 Franco, A., Malhotra, N. & Simonovits, G. Publication bias in the social sciences: Unlocking the file drawer. *Science* **345**, 1502-1505 (2014).

37 Franco, A., Malhotra, N. & Simonovits, G. Underreporting in psychology experiments: Evidence from a study registry. *Social Psychological and Personality Science* **7**, 8-12 (2016).

38 Head, M. L., Holman, L., Lanfear, R., Kahn, A. T. & Jennions, M. D. The extent and consequences of p-hacking in science. *PLoS biology* **13**, e1002106 (2015).

39 Yang, Y. *et al.* 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**, 1-20 (2023).

40 Sterne, J. A., Gavaghan, D. & Egger, M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *Journal of clinical epidemiology* **53**, 1119-1129 (2000).

41 Senior, A. M. *et al.* Heterogeneity in ecological and evolutionary meta‐analyses: its magnitude and implications. *Ecology* **97**, 3293-3299 (2016).

42 Easterbrook, P. J., Gopalan, R., Berlin, J. & Matthews, D. R. Publication bias in clinical research. *The Lancet* **337**, 867-872 (1991).

43 Fanelli, D., Costas, R. & Ioannidis, J. P. Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences* **114**, 3714-3719 (2017).

44 Henmi, M. & Copas, J. B. Confidence intervals for random effects meta-analysis and robustness to publication bias. *Statistics in medicine* **29**, 2969-2983 (2010).
45 Bramley, P., López-López, J. A. & Higgins, J. P. Examining how meta-analytic methods perform in the presence of bias: a simulation study. *Research Synthesis Methods* **12**, 816-830 (2021).
46 Stanley, T. D. & Doucouliagos, H. Neither fixed nor random: weighted least squares meta-analysis. *Statistics in medicine* **34**, 2116-2127 (2015).
47 Baker, R. D. & Jackson, D. Meta-analysis inside and outside particle physics: two traditions that should converge? *Research Synthesis Methods* **4**, 109-124 (2013).
48 Doi, S. A., Barendregt, J. J., Khan, S., Thalib, L. & Williams, G. M. Advances in the meta-analysis of heterogeneous clinical trials I: the inverse variance heterogeneity model. *Contemporary clinical trials* **45**, 130-138 (2015).
49 Bramley, P., López-López, J. A. & Higgins, J. P. T. Examining how meta-analytic methods perform in the presence of bias: A simulation study. *Research Synthesis Methods* **12**, 816-830, doi:https://doi.org/10.1002/jrsm.1516 (2021).
50 Nakagawa, S. *et al.* Methods for testing publication bias in ecological and evolutionary meta-analyses. *Methods in Ecology and Evolution* **13**, 4-21 (2022).
51 Stanley, T. D. & Doucouliagos, H. Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods* **5**, 60-78 (2014).
52 Stanley, T. D. & Doucouliagos, H. Neither fixed nor random: weighted least squares meta-regression. *Research synthesis methods* **8**, 19-42 (2017).
53 Stanley, T., Doucouliagos, H. & Ioannidis, J. P. Beyond Random Effects: When Small-Study Findings Are More Heterogeneous. *Advances in Methods and Practices in Psychological Science* **5**, 25152459221120427 (2022).
54 Nakagawa, S. *et al.* The orchard plot: cultivating a forest plot for use in ecology, evolution, and beyond. *Research Synthesis Methods* **12**, 4-12 (2021).
55 Morrissey, M. B. Meta-analysis of magnitudes, differences and variation in evolutionary parameters. *Journal of Evolutionary Biology* **29**, 1882-1904 (2016).
56 Viechtbauer, W. Conducting meta-analyses in R with the metafor package. *Journal of statistical software* **36**, 1-48 (2010).
57 Bird, G., Kaczvinsky, C., Wilson, A. E. & Hardy, N. B. When do herbivorous insects compete? A phylogenetic meta-analysis. *Ecology Letters* **22**, 875-883 (2019).
58 Pustejovsky, J. E. & Tipton, E. Meta-analysis with robust variance estimation: Expanding the range of working models. *Prevention Science* **23**, 425-438 (2022).
59 Tipton, E. Small sample adjustments for robust variance estimation with meta-regression. *Psychological methods* **20**, 375-393 (2015).
60 Joshi, M., Pustejovsky, J. E. & Beretvas, S. N. Cluster wild bootstrapping to handle dependent effect sizes in meta-analysis with a small number of studies. *Research Synthesis Methods* **13**, 457-477 (2022).
61 Doi, S. A. & Thalib, L. A quality-effects model for meta-analysis. *Epidemiology*, 94-100 (2008).
62 Doncaster, C. P. & Spake, R. Correction for bias in meta-analysis of little-replicated studies. *Methods in Ecology and Evolution* **9**, 634-644 (2018).
63 Nakagawa, S. *et al.* A robust and readily implementable method for the meta-analysis of response ratios with and without missing standard deviations. *Ecology Letters* **26**, 232-244 (2023).
64 Bakbergenuly, I., Hoaglin, D. C. & Kulinskaya, E. Estimation in meta-analyses of response ratios. *BMC medical research methodology* **20**, 1-24 (2020).

65 Steegen, S., Tuerlinckx, F., Gelman, A. & Vanpaemel, W. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science* **11**, 702-712 (2016).

66 Wagenmakers, E.-J., Sarafoglou, A. & Aczel, B. One statistical analysis must not rule them all. *Nature* **605**, 423-425 (2022).

67 Costello, L. & Fox, J. W. Decline effects are rare in ecology. *Ecology*, e3680 (2022).

68 Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G. & Group, P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS med* **6**, e1000097 (2009).

69 Nakagawa, S. & Cuthill, I. C. Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological reviews* **82**, 591-605 (2007).

70 Platt, R. W., Leroux, B. G. & Breslow, N. Generalized linear mixed models for meta-analysis. *Statistics in medicine* **18**, 643-654 (1999).

71 Searle, S. R. Another Look at Henderson's Methods of Estimating Variance Components. *Biometrics* **24**, 749-787, doi:10.2307/2528870 (1968).

72 Viechtbauer, W., López-López, J. A., Sánchez-Meca, J. & Marín-Martínez, F. A comparison of procedures to test for moderators in mixed-effects meta-regression models. *Psychological Methods* **20**, 360-374 (2015).

73 Hartung, J. An alternative method for meta-analysis. *Biometrical Journal: Journal of Mathematical Methods in Biosciences* **41**, 901-916 (1999).

74 Knapp, G. & Hartung, J. Improved tests for a random effects meta-regression with a single covariate. *Statistics in medicine* **22**, 2693-2710 (2003).

75 Sidik, K. & Jonkman, J. N. A simple confidence interval for meta-analysis. *Statistics in medicine* **21**, 3153-3159 (2002).

76 Rosenfeld, A. H. The particle data group: Growth and operations-eighteen years of particle physics. *Annual Review of Nuclear Science* **25**, 555-598 (1975).

77 Van Aert, R. C. & Jackson, D. A new justification of the Hartung-Knapp method for random-effects meta-analysis based on weighted least squares regression. *Research synthesis methods* **10**, 515-527 (2019).

78 Nakagawa, S. *et al.* Meta-analysis of variation: ecological and evolutionary applications and beyond. *Methods in Ecology and Evolution* **6**, 143-152 (2015).

Acknowledgement

We highly value the contributions and insights shared by Wolfgang Viechtbauer and James Pustejovsky in the R Special Interest Group for Meta-Analysis and the publication bias workshop at the 2023 Evidence Synthesis and Meta-Analysis in R conference. YY was funded by the National Natural Science Foundation of China (NO. 32102597). SN, YY, and ML were funded by the Australian Research Council Discovery Grant (DP210100812). DN was supported by an ARC Discovery Grant (DP210101152). A part of this research was conducted while visiting the Okinawa Institute of Science and Technology (OIST) through the Theoretical Sciences Visiting Program (TSVP) to SN.

Contributions

YY and SN conceptualised the idea and drafted the manuscript. YY analysed the data with the help of SN. YY led the creation of the accompanying webpage working with CW and SN. ML and YY led all the visualisations. DN developed the helper function with the help of SN. All authors read, commented on, and edited the manuscript and approved the final submission.

Competing interests

The author reported no conflict of interest.

Supplementary Information

Table S1 – S3 and Figure S1 – S3