A robust and readily implementable method for the meta-analysis of response ratios with and without missing standard deviations

Shinichi Nakagawa*[1], Daniel W. A. Noble[2], Malgorzata Lagisz[1], Rebecca Spake[3], Wolfgang Viechtbauer[4] & Alistair M. Senior*[5]

Address:

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

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

[3] School of Biological Sciences, University of Reading, Whiteknights Campus, Reading RG6 6AS, UK

[4] Department of Psychiatry and Neuropsychology, School for Mental Health and Neuroscience, Faculty of Health, Medicine, and Life Sciences, Maastricht University, 6200 MD Maastricht, The Netherlands

[5] Charles Perkins Centre and School of Life and Environmental Sciences, University of Sydney, Camperdown, NSW 2006, Australia.

Running head: Improved meta-analyses of the response ratio

* Correspondence: Shinichi Nakagawa (s.nakagawa@unsw.edu.au); Alistair M. Senior (alistair.senior@Sydney.edu.au)

Abstract

The log response ratio, lnRR, is the most frequently used effect size statistic in ecology. However, missing standard deviations (SDs) are often present in meta-analytic datasets, preventing us from obtaining the sampling variance for lnRR. We propose three new methods to deal with missing SDs. All three methods use the square of the weighted average coefficient of variation CV to obtain sampling variances for lnRR when SDs are missing. Using simulation, we find that using the average CV to estimate the sampling variances for all observations, regardless of missingness, performs best. Surprisingly, even where SDs are missing, this simple method performs better than the conventional analysis with no missing SDs. This is because the conventional method incorporates biased estimates of sampling variances as opposed to less biased sampling variances with the average CV. All future meta-analyses of lnRR could take advantage of our new approach along with the other methods.

KEYWORDS

Missing data, multiple imputation, meta-regression, robust variance estimation, research synthesis

INTRODUCTION

Meta-analyses are frequently used to quantitatively synthesize the outcomes of ecological studies and explain inconsistencies among study findings (Gurevitch *et al.* 2018). However, incomplete reporting of necessary data in the primary literature threatens the validity of meta-analytic evidence. Specifically, many empirical papers fail to report standard deviations (SDs) or their derivatives, such as standard errors (SEs) and confidence intervals (CIs). SDs are required to calculate effect sizes and/or their precision for comparing means between two groups. The two best-known effect sizes for mean comparison are the standardized mean difference, SMD (well-known estimators include Cohen's *d* and Hedges' *g*) and the natural logarithm of the response ratio, lnRR (Hedges *et al.* 1999).

The use of these effect sizes, which requires SDs, is widespread in ecology (Nakagawa & Santos 2012; Koricheva & Gurevitch 2014). Yet, a recent review of 505 ecological meta-analytic studies showed nearly 70% of the datasets included studies with missing SDs (Kambach *et al.* 2020). The same review also showed that many meta-analysts did not use studies with missing SDs and performed a complete-case meta-analysis. Finally, this review also demonstrated that excluding studies with missing data could both upwardly and downwardly bias meta-analytic results, and that multiple imputation (MI) of missing SDs (and sample sizes) was an effective means of providing unbiased meta-analytic results.

Multiple imputation (MI) was introduced to ecologists more than a decade ago (Nakagawa & Freckleton 2008). However, there has been limited uptake of this method in ecological meta-analysis (cf. Ellington *et al.* 2015; Kambach *et al.* 2020). There are, we believe, two major reasons for this slow uptake. First, for many ecologists, the implementation of MI might be considered tedious, perhaps because it involves three steps: 1) creating multiple datasets with imputed missing data, e.g., $N_{dataset}$ = 100 with all missing SD data imputed, 2) analyzing each dataset separately, and

3) aggregating parameter estimates (e.g., regression coefficients) using Rubin's rules (Rubin 1987) (for details, see Nakagawa 2015; van Buuren 2018). Second, there may be uncertainty around its implementation for the complex models applied to ecological datasets. For example, it is unclear if Rubin's rules are always appropriate for aggregating variance components and related quantities (e.g., I^2 and R^2) and information criteria (e.g., AIC, BIC; cf. Nakagawa & Freckleton 2011). Furthermore, MI cannot be easily implemented for multilevel (mixed-effects / hierarchical) models unless they have only two levels (i.e., one random factor) (van Buuren 2018). Therefore, MI for meta-analytic studies using multilevel models is still seldom applied, reinforcing age-old approaches of excluding studies (observations) with missing SDs.

Here, we propose alternatives to MI for handling studies with missing SDs in meta-analyses that use the lnRR as the effect size measure (Nakagawa & Santos 2012; Koricheva & Gurevitch 2014; Kambach *et al.* 2020). We introduce three new methods to deal with missing SDs in a multilevel meta-analytic model as well as traditional random-effects models (summarized in Fig. 1). First, we introduce a method developed by Doncaster and Spake (2018) that uses an adjusted sampling variance formula for lnRR values. We then improve and extend this approach for missing SDs; we provide two procedures within this method: only using this adjustment method for effect sizes with missing SD (Method 1A) and using this adjustment method for all effect sizes regardless of missingness (Method 1B). Second, we describe a method that extends traditional weighted regression (Method 2). Next, we combine these two methods to provide a hybrid method (Method 3). We conduct a simulation study to compare these methods to the baseline method (a standard meta-analytic procedure without missing SDs) and show that Method 1B performs best. Surprisingly, for many cases, Method 1B (and Method 2) with missing SDs outperforms the baseline method without missing SDs (i.e., full data analysis). We make some recommendations for future meta-analyses accordingly. Importantly, we implement and illustrate these new methods

using the widely used R package, *metafor* (Viechtbauer 2010 ; all relevant data and code are available at a GitHub repository; see below).

NEW STATISTICAL METHODS

Estimating sampling variances prior to meta-analysis: Method 1A and 1B

The effect size statistic, lnRR, was first proposed by Hedges and colleagues (1999) as follows:

$$\mathrm{lnRR}_1 = \ln\left(\frac{m_1}{m_2}\right), \quad (1)$$

$$v(\mathrm{lnRR}) = \frac{sd_1^2}{n_1 m_1^2} + \frac{sd_2^2}{n_2 m_2^2} = \frac{\mathrm{CV}_1^2}{n_1} + \frac{\mathrm{CV}_2^2}{n_2}, \quad (2)$$

where m_1 and m_2 are the means of group 1 and group 2, respectively (e.g., experimental and control groups), v represents the sampling variance, sd and n are the corresponding SDs and sample sizes, respectively, and CV (sd/m) is the coefficient of variation.

However, when sample size (replicates) per effect size (study) is small, CVs in Equation 2 are often inaccurate. If we assume CV (CV^2) values for group 1 and group 2 (cf. Equation 2) are fairly constant or homogeneous across effect sizes (studies), we can obtain potentially better estimates of CV^2 values by averaging (Doncaster & Spake 2018; see also Hedges & Olkin 1985; Hunter & Schmidt 1990; Berkey *et al.* 1995):

$$v^*(\mathrm{lnRR}) = \frac{\sum_{i=1}^{K}\left(\mathrm{CV}_{1i}^2\right)/K}{n_1} + \frac{\sum_{i=1}^{K}\left(\mathrm{CV}_{2i}^2\right)/K}{n_2}, \quad (3)$$

where CV_{1i}^2 and CV_{2i}^2 are from the ith effect size (study; i = 1, 2, …, K; we assume the number of effect sizes = the number of studies = K). Indeed, Doncaster and Spake (2018) have demonstrated that the use of Equation 3 over Equation 2 improve the accuracy and precision of the overall (meta-analytic) mean estimate, especially when n is small (meaning n = 3–10 observations, with $n_1 + n_2$ =

6–20). Notably, they also suggested this formula could be used when SDs are missing from some studies, although this application was not investigated in their simulation.

Here we propose two improvements to Equation 3. Using simulations, Lajeunesse (2015) showed that Equation 1 and 2 are both biased when sample sizes are small to moderate, and showed the following estimators – based on the second-order Taylor expansion – can reduce these biases (see also Senior *et al.* 2020):

$$\mathrm{lnRR}_2 = \ln\left(\frac{m_1}{m_2}\right) + \frac{1}{2}\left(\frac{\mathrm{CV}_1^2}{n_2} - \frac{\mathrm{CV}_2^2}{n_1}\right), \quad (4)$$

$$v(\mathrm{lnRR}) = \frac{\mathrm{CV}_1^2}{n_1} + \frac{\mathrm{CV}_2^2}{n_2} + \frac{\mathrm{CV}_1^4}{2n_1^2} + \frac{\mathrm{CV}_2^4}{2n_2^2}. \quad (5)$$

Therefore, Equation 3 can be further improved by including the extra-terms in Equation 5. Also, rather than using the average CV^2, we can use the square of the weighted average of CV for both the point estimate (effect size) and sampling variance as follows:

$$\mathrm{lnRR}_3 = \ln\left(\frac{m_1}{m_2}\right) + \frac{1}{2}\left(\frac{[\sum_{i=1}^{K}(n_{1i}\mathrm{CV}_{1i}) / \sum_{i=1}^{K} n_{1i}]^2}{n_1} - \frac{[\sum_{i=1}^{K}(n_{2i}\mathrm{CV}_{2i}) / \sum_{i=1}^{K} n_{2i}]^2}{n_2}\right) \quad (6)$$

$$\tilde{v}(\mathrm{lnRR}) = \frac{[\sum_{i=1}^{K}(n_{1i}\mathrm{CV}_{1i}) / \sum_{i=1}^{K} n_{1i}]^2}{n_1} + \frac{[\sum_{i=1}^{K}(n_{1i}\mathrm{CV}_{2i}) / \sum_{i=1}^{K} n_{2i}]^2}{n_2}$$

$$\frac{[\sum_{i=1}^{K}(n_{1i}\mathrm{CV}_{1i}) / \sum_{i=1}^{K} n_{1i}]^4}{2n_1^2} + \frac{[\sum_{i=1}^{K}(n_{2i}\mathrm{CV}_{2i}) / \sum_{i=1}^{K} n_{2i}]^4}{2n_2^2}, \quad (7)$$

One can use Equation 7 (along with Equation 6) where SDs are missing, because the weighted cross-study CV can substitute missing SDs, allowing the inclusion of these studies in meta-analyses (Method 1A; a mixture of Equations 4-7). Alternatively, one may use Equation 7 throughout regardless of the missingness of SDs (Method 1B with Equation 6 & 7; see Fig. 1). Note we discuss the use of the square of the weighted average of CV (Equations 6 & 7) rather than weighted average of CV^2 below (see Section "The accuracy and limitation of lnRR").

Using weighted-regression-like method: Method 2

In the absence of SDs, it has been suggested that information on sample sizes, which are commonly available, can be used to approximate the sampling variances in meta-analyses of lnRR or SMD, using the inverse of the following (e.g., Lajeunesse 2013; Kambach *et al.* 2020):

$$\tilde{n} = \frac{n_1 n_2}{n_1 + n_2}. \quad (8)$$

However, treating Equation 8 as an estimate of the 'exact' sampling variance is erroneous, because it ignores other terms in Equation 2 & 5 (i.e., mean and SD) (see the review by Kambach *et al.* 2020). A more realistic assumption is to treat $1/\tilde{n}$ as proportional to the sampling error; indeed, Equation 2 reduces to the inverse of Equation 8 when we set both CVs to 1. Weighted regression models, commonly used to correct for heteroscedasticity, make this assumption of proportionality, unlike the classical random-effects meta-analytical model, which assumes that the exact sampling variances are known. Many ecologists are likely to be familiar with weighted regression models that specify sample sizes as the weights (Fletcher & Dixon 2012).

The simplest random-effects meta-analytic model using lnRR can be written as follows:

$$\mathrm{lnRR}_i = \beta_0 + s_i + m_i, \quad (9)$$

$$s_i \sim \mathcal{N}(0, \sigma_s^2), \; m_i \sim \mathcal{N}(0, v_i),$$

where β_0 is the overall/average effect (or meta-analytic mean); s_i is the between-study effect for the *i*th effect size, normally distributed with a mean of zero and a variance of σ_s^2 (sometimes, referred to as τ^2), m_i is the sampling error for the *i*th effect size, distributed with the *i*th sampling variance (note that $i = 1, 2, \ldots, K$, the number of effect sizes = the number of studies). As mentioned earlier, we here assume that the sample variance of lnRR is known; that is, we use either Equation 2 or 5 as equal to the true sampling variances (v_i) in Equation 9. The ratio between σ_s^2 and the total variance is often used to quantify heterogeneity (I^2):

159 $$I^2 = \frac{\sigma_s^2}{\sigma_s^2 + \bar{v}}, \qquad (10)$$

160 where $\bar{v}$ is known as a 'typical' (or 'average') sampling variance (originally referred to as 'typical
161 within-study variance'; *sensu* Higgins & Thompson 2002), which can be estimated in several ways
162 (Xiong *et al.* 2010).

163

164 Unlike the meta-analytic model, in a weighted regression, the following is assumed:

165 $$v_i = \phi\left(\frac{1}{\tilde{n}_i}\right), \quad (11)$$

166 where ϕ, which is estimated by the model, functions as a multiplicative parameter fulfilling the
167 assumption of proportionality (i.e., $1/\tilde{n}_i$ is proportional to the sampling variance v_i). The key
168 distinction here is that Methods 1A and 1B both make the classical assumption from meta-analysis;
169 that is Equation 7 is an accurate estimate of the sampling variance for a study. However, Doncaster
170 and Spake's simulation suggests that the sampling variance (using Equation 3) is likely to be
171 inaccurate when sample sizes are small (e.g., $n_1 + n_2 = 6$ –20). Therefore, it may be advisable to
172 assume that v_i^* (Equation 3) is proportional to the true sampling variance. In a similar vein, we can
173 apply the same assumption to Equation 7; that is:

174 $$v_i = \phi\tilde{v}_i. \quad (12)$$

175 Here, our Method 2 makes the assumption of proportionality for the sampling variance using
176 Equation 12 in a weighted-regression-like model (Fig. 1).

177 **Combining Methods 1 and 2: Method 3**

178 In the second method, Equation 7 is used regardless of whether SDs are missing or not. We can,
179 however, combine Method 1 and Method 2 together into a new, hybrid Method 3 (Fig. 1). When
180 SDs are available, we can use Equation 5 to obtain the sampling variance of lnRR (along with
181 Equation 4 for the point estimate). When SDs are missing, we can use Equation 7 (and Equation 6)

combined with Method 2 in a weighted regression (i.e., $v_i = \phi\tilde{v}_i$). We can write this model (Method 3), using a multilevel meta-analysis (modelling multiple effect sizes per study) as follows:

$$\ln\mathrm{RR}_{ij} = \beta_0 + s_i + u_{ij} + m_{ij}, \quad (13)$$

$$s_i \sim \mathcal{N}(0, \sigma_s^2), \; u_{ij} \sim \mathcal{N}(0, \sigma_u^2), \; m_{ij} \sim \mathcal{N}(0, \mathbf{V})$$

where s_i is the between-study effect for the ith study (i = 1, 2, …, K), normally distributed with a mean of 0 and variance of σ_s^2 (often referred to as τ^2); u_{ij} is the between-effect-size effect (or within-study effect) for the jth effect size in the ith study, distributed with a mean of zero and variance of σ_u^2 (j = 1, 2, …, L_i, where L_i denotes the number of effect sizes within the ith study), **V** is a diagonal matrix with v_{ij} when no SDs are missing and $\phi\tilde{v}_{ij}$ when SDs are missing (i.e., Method 3). For example, when we have five effect sizes in three studies, **V** would be:

$$\mathbf{V} = \begin{bmatrix} v_{11} & 0 & 0 & 0 & 0 \\ 0 & v_{12} & 0 & 0 & 0 \\ 0 & 0 & \phi\tilde{v}_{21} & 0 & 0 \\ 0 & 0 & 0 & \phi\tilde{v}_{22} & 0 \\ 0 & 0 & 0 & 0 & v_{31} \end{bmatrix},$$

where 1st, 2nd and 5th cases (effect sizes) have SDs while the 3rd and 4th are without SDs, and as above, ϕ is estimated in the model. Because this model accounts for non-independence, it is appropriate in ecological meta-analyses that include correlations among-effect sizes such as when there is more than one effect size per study or species (Nakagawa & Santos 2012; Noble *et al.* 2017; Nakagawa *et al.* 2022; but for a more complex model with **V** including covariances, or sampling variances having dependencies, Appendix S1; https://alistairmcnairsenior.github.io/Miss_SD_Sim/).

SIMULATION

Simulation design

We conducted a simulation study to compare the performance of Methods 1A, 1B, 2, and 3 on meta-analytic datasets with varying proportions of missing SDs. We also computed a (baseline) meta-analytic model with full data, for reference (see Table 1 for which equation was used; see also Fig. 1). To represent a typical eco-evolutionary dataset, we simulated a hierarchical structure where each study contained ≥1, correlated effect size (i.e., we simulated an intra-class correlation for each study; $ICC_s = \sigma_s^2 / (\sigma_s^2 + \sigma_u^2)$ using the terms in Equation 13). For each simulated dataset we analyzed the full dataset, before deleting SDs for 5%, 15%, 25% 35%, 45%, or 55% of the studies. Missingness was imposed at the study-level, rather than the effect size-level. We then analyzed each dataset with the four proposed methods for handling missing SDs (Method 1A, 1B, 2, and 3). Datasets were analyzed using models that included a study-level and an effect-size random effect, specified using the 'rma.mv' function in *metafor*. For each model, bias was calculated as the difference between the estimated and parametrized value, for the following parameters: i) the meta-estimate of the overall mean effect size, ii) coverage of 95% confidence intervals (CIs), iii) total heterogeneity ($\tau^2 = \sigma_s^2 + \sigma_u^2$ in Equation 12 and $\tau^2 = \sigma_s^2$ in Equation 8; log ratio of estimated and parametrized value) and bias in the estimated ICC_s (difference between estimated and parametrized value). CIs were calculated as the estimated effect ± *t*-value × SE, where for t-values the degrees of freedom were the number of effect sizes minus 1, when $ICC_s = 0$, and the number of studies minus 1 when $ICC_s > 0$.

Each simulated dataset contained *K* studies (*K* = 12, 30, and 100 were tested). Because studies often vary in the number of effect sizes they contain, the number of effect sizes per study, *L*, was assigned as a random variable. We simulated *L* using a double Poisson distribution, which is a discrete probability distribution that can be under/over dispersed relative to a Poisson distribution via a multiplicative dispersion parameter. Using the 'rDPO' function in the *gamlss.dist* package, *L* was

parameterized by drawing values from a random double Poisson distribution with a mean of 2 and a multiplicative dispersion parameter of 2.88, before adding 1 (to prevent 0 values). This resulted in L having a minimum of 1, a mean of 3, and SD of 2.4 (i.e., dispersion of 1.92). We termed this set of parameters Set I. We also simulated a second set where L is fixed to 1 (i.e., each study had only one effect size; $L = 1$, dispersion = 0), which we called Set II. Set II is equivalent to a meta-analysis with just one effect size per study (i.e., no dependency), and which would be assessed using a standard random-effects meta-analysis (i.e., Equation 9) and its variants (i.e., Methods 2 & 3).

To simulate effect sizes that were correlated in a hierarchical manner, we assumed an overall lnRR (θ) of 0.3 ($e^{0.3} = 1.35$, or a 35% increase in the mean) with either negligible ($\tau^2 = 9\times10^{-6}$ or $\tau / \theta = 0.01$) or high total heterogeneity ($\tau^2 = 0.09$ or $\tau / \theta = 1$). This heterogeneity was partitioned between among- and within-study level effects assuming a given intra-class correlation (ICC_S; values of 0 and 0.5 were tested) such that the jth effect size ($j = 1 \ldots L_i$) in the ith ($i = 1 \ldots K$) study, θ_{ij} (cf. Equation 13) was drawn from a hierarchical pair of random normal distributions ('rnorm' function in *base* R) as:

$$\theta_i \sim N\left(\theta, \sqrt{\tau^2 \times \mathrm{ICC}_S}\right),$$

$$\theta_{ij} \sim N\left(\theta_i, \sqrt{\tau^2 \times (1 - \mathrm{ICC}_S)}\right).$$

To simulate variation in the precision of the studies in the dataset we treated the sample size of the underlying studies as a random variable, N. We assumed N varied at the level of the study such that each group/effect size within the same study had the same sample size. In our experience it is common for experimental designs to vary among, more than within, studies. We drew the simulated sample size for study k by drawing a random value from double Poisson distribution before adding a value of 3. The double Poisson distribution was parametrized with a mean of either 2 or 27 coupled with dispersion parameters of either 3.65 or 1.66. After adding the constant of 3, this resulted in two different distributions of N both with a minimum of 3, and (over) dispersion of 1.5, but with a mean (μ_N) of either 5 or 30. The smaller mean value of 5 is more typical in

terrestrial/ecosystem ecology (or some pre-clinical biomedical studies), while the larger mean value is more like evolutionary/behavioural ecology studies (or clinical trials) (Senior *et al.* 2016).

The underlying data in control and treatment groups in each effect size were drawn from random normal distributions 'shifted' to ensure both groups had a positive mean as is required for analysis using lnRR. From these individual simulated values, we calculated the mean and SD in each group for calculation of lnRR and downstream meta-analysis. The observations for the control group in effect size *j* in study *i* were drawn from the random normal distribution, $N(100, \sigma_i)$, and the paired treatment group from the random normal distribution, $N\left(100 \times e^{\theta_{ij}}, \sigma_i\right)$, where σ_i is the SD in the underlying individual observations in study *i*.

Because we are assessing the performance of methods to deal with missing SD values, we chose to treat the within-group (among-observation) SD as a random variable, *S*. The SD for study *i* was drawn from a random Gamma distribution (the 'rgamma' function in *base* R) with shape $\frac{\mu_S^2}{\sigma_S^2}$ and scale $\frac{\sigma_S^2}{\mu_S}$, where μ_S is the mean of *S* (i.e., mean SD of studies; here 15), and σ_S is the SD in *S*. This latter parameter thus specifies how heterogeneous the within-study (among-observation) variances are; we tested values of 10^{-10} (~0), 3.75, and 7.5 (i.e., entirely homogeneous variances, or the CV for the SD among studies is 0.25 or 0.5). A summary of the key parameters and their values is given in Table S1. Each combination of parameter values was simulated 10,000 times for both Set I and Set II. For Set I presented in the main text, we used the multilevel meta-analytic model (Equation 13) and its variants (Method 2 & 3). For Set II, we used the random-effects meta-analytic model (Equation 9) and its variants where the results are presented in the supplementary materials. For all three methods, we needed to calculate the average CV as in Equations 6 and 7. In Set I, this calculation was done by averaging CV within studies and then taking the weighted-average CV across studies (using mean *n* per study as the weight), disregarding rows containing missing SDs.

For Set II, we calculated the weighted CV among studies (using *n* per study as the weight) as we only had one CV value per study (see also Fig. S1-S3).

Simulation results

Fig. 2A shows the distribution of the median bias in the estimated overall effect under each simulated condition with complete data and using the four different methods for handling missing SDs. Even with full data, both upward and downward bias was possible in the estimated effect size, and this was reflected in the analyses using Method 1A and 3 to handle missing SDs. Notably, even at its most extreme, this bias only amounted to a little over 2% of the true effect size and was usually ~0.5%. Nonetheless, Method 1B and 2, both of which use the weighted average CV to estimate the sampling variance for all effect sizes regardless of missingness, yielded the lowest bias on average (Fig. 2A). The degree of bias across conditions in the full data analysis correlated very strongly with bias using Method 1A and 3, while bias in Method 1B and 2 correlated strongly (Fig 2B). This suggests that the methods fall into two classes that perform similarly across situations: Method 1A with Method 3, and Method 1B with Method 2. Contrasting Methods 1A and 1B directly, in almost all cases the absolute level of bias in Method 1A was higher than that for Method 1B (Fig. 2C). Further, where Method 1B had a higher bias than method 1A, this difference was small (Fig. 2C). Although Method 1B and Method 2 outperformed the other approaches on average, they were prone to producing excessively large bias on rare occasions; Fig. 2D shows the range in bias among the individual replicates under each simulated condition as a function of the different methods. Large ranges in bias occurred when the SDs among different studies were very heterogeneous, and the individual studies themselves had a low sample size (Fig. 2E).

All methods for handling missing data, and the full data analyses, could produce 95% CIs that were too narrow, or too wide under different scenarios (Fig. 3A). The full data, and Methods 1A and 3 tended to typically produce CIs that were slightly too narrow, whereas method 1B and 2 were prone to producing wider CIs (Fig 3A). Again, contrasting Method 1A and 1B, Method 1B's tendency to

303 produce a CI that is too wide occurs when the total heterogeneity among studies is low (Fig. 3B and
304 3C). However, where total heterogeneity is high, Method 1B performs as well as Method 1A (Fig.
305 3B and 3C).
306
307 Fig. 4A shows the median bias in the estimated heterogeneity under each condition and method.
308 Under most conditions, methods 1A, 1B, and 3 estimated heterogeneities with little bias, but all
309 including analysis of the full data were prone to substantially overestimating the total heterogeneity
310 (Fig. 3A). Method 2 tended to slightly underestimate heterogeneity (Fig. 3A). Overestimation of
311 heterogeneity occurred where the actual level of heterogeneity was low (Fig. 3B). On average most
312 methods did a good job of partitioning heterogeneity between the within- and among-study levels,
313 although Method 2 displayed a slight bias on average (Fig. 3C). Under some circumstances all
314 methods could be biased in partitioning heterogeneity (Fig. 3C). Method 1B was prone to bias
315 partitioning when the total heterogeneity was low; overestimating the ICC when the simulated study
316 effect was absent and underestimating when it was present (Fig 3D). In summary, while Method 1B
317 performed with the least bias under the broad range of simulated conditions tested, all the methods
318 fared surprisingly well, compared with the full data analysis (see Discussion for more).

319 **IMPLIMENTATION**

320 **The accuracy and limitation of lnRR**

321 The accuracy of the sampling variance for lnRR depends on whether lnRR is normally distributed.
322 Hedges et al. (1999) suggested a simple test to check the normality assumption based on Geary
323 (1930). This test was improved by Lajeunesse (2015) and expressed as:

324 $$\frac{1}{\mathrm{CV}}\left(\frac{4n^{\frac{3}{2}}}{1+4n}\right) \geq 3. \quad (14)$$

If many effect sizes fail to fulfil this relationship, then, meta-analytic results are unlikely to be robust. However, such tests are rarely used. This test is sometimes referred to as Geary's test whose original form is $\sqrt{n}/\mathrm{CV} \geq 3$. Lajeunesse (2015) has suggested a sensitivity analysis excluding effect sizes which fail to fulfill Equation 14 as a check on the robustness of the meta-analytic results.

It is important to realize that count data and abundance data, which are extremely common in ecology (Spake *et al.* 2021), may often fail this test (Equation 14). This is because such data is usually overdispersed meaning CV > 1. For example, it is not uncommon for count data to have, say, CV = 5, especially when the mean is close to zero (cf. Lajeunesse 2015). In such a case, one would need sample sizes larger than 226 for each treatment and control group to pass the test, which would be difficult for most ecological studies.

Importantly, because the Taylor expansion used in Equations 4-7 also assumes normality, these formulations of lnRR are even more sensitive than the original formulations (Equations 1 &2). Therefore, it may be advisable to use Equation 1 for the point estimate and the following estimator of the sampling variance (rather than Equation 7) when many effect sizes fail Geary's test:

$$\tilde{v}(\mathrm{lnRR}) = \frac{[\sum_{i=1}^{K}(n_{1i}\mathrm{CV}_{1i}) \,/\, \sum_{i=1}^{K} n_{1i}]^2}{n_1} + \frac{[\sum_{i=1}^{K}(n_{1i}\mathrm{CV}_{2i}) \,/\, \sum_{i=1}^{K} n_{2i}]^2}{n_2} \quad (15)$$

This formula still relies on the first-order Taylor expansion but is less sensitive than Equation 7. Of relevance, the square of the average of CV and the average of CV^2 tend to be very similar when the value of CV is small. However, when CV becomes larger (or fail to satisfy the 'improved' Geary's test), they diverge. When data are overdispersed, in particular, the average of CV^2 could be much higher than the square of the average of CV. This is the primary reason we chose the square of the average of CV over the average of CV^2; the former is less heavily influenced of individual effect

sizes based on overdispersed data. Other limitations (and advantages) of lnRR are discussed elsewhere (e.g., Spake *et al.* 2021; Yang *et al.* 2022).

Worked examples

Bird and colleagues (2019) conducted a meta-analysis exploring the impacts of competition on herbivorous insect fitness when occupying the same host plant with another species or in isolation. In brief, they collected data on a series of fitness measurements (e.g., abundance, body size, development time, fecundity; see Table 2 in Bird *et al.* 2019) and quantified the separate overall impact of competition on the various fitness measures, using phylogenetic multilevel meta-analyses (Cinar *et al.* 2022; Appendix S1).

For demonstration purposes, we focus on the largest dataset that makes use of abundance data. We restricted the analysis ratio scale (required for lnRR) and those effect sizes that passed the 'improved' Geary's test (Equation 14; see above). We use a multilevel meta-analytic model to estimate the overall impact of competition on focal insect fitness (i.e., intercept or overall meta-analytic mean) while controlling for phylogeny, research group, and research year (as per the analysis by Bird *et al.* 2019). We then introduced missing data at the study (article) level, so that a randomly selected ~20% of articles had effect sizes with missing SD in the control and experimental treatment; a scenario that is typical of many meta-analyses (cf. Kambach *et al.* 2020).

An analysis of these data applying the different methods compared to the whole data is provided in Table 2. We can see that the complete-case analysis (excluding all data with missing SDs) results in slightly larger confidence intervals and a reduction in the meta-analytic mean effect size, relative to the other methods. Methods 1A, 2, and 3 all suggest the overall meta-analytic is slightly smaller and result in greater precision around this estimated effect size. Using this example, we show how each approach is implemented in the supplement (Appendix S2) along with another example (McDonald *et al.* 2019; Appendix S3).

373 **DISCUSSION**

374 In this study, we have developed new methodological procedures to handle missing SDs in meta-
375 analyses of lnRR values. Our simulation suggested that the least biased estimates were obtained by
376 Method 1B, which uses the weighted average CV to calculate point estimates and sampling
377 variances for all effect sizes, regardless of missingness in SD. In terms of implementation, this is
378 the easiest method of all (see Supporting Information). We were surprised to find that Method 1B
379 (along with Method 2) outperformed the conventional meta-analysis of full data. This is especially
380 so given Method 1B along with other new proposed methods use 'single imputation' rather than
381 'multiple imputation', and analysis with single imputation should, in theory, fare worse than meta-
382 analysis with full data (using Equation 4 & 6, see Table 1; Nakagawa & Freckleton 2008;
383 Nakagawa 2015; van Buuren 2018; Kambach *et al.* 2020; see also Fletcher & Dixon 2012).
384
385 Given that Doncaster and Spake (2018) found the use of Equation 3 (the use of average CV)
386 performed better than analysis with Equation 2, we might have expected Method 2 to do well (see
387 also Lin & Aloe 2021). It is important to note that our simulation work clearly differed in at least
388 two respects. First, Doncaster and Spake never tested how their method fared with missing data.
389 Second, their simulation was restricted to non-multilevel models; however, many ecological
390 datasets require a multilevel approach. Method 1B and Doncaster and Spake's procedure (i.e., using
391 Equation 3 rather than Equation 2) perform well because even where they are reported, CV values
392 from individual studies are often inaccurate due to the small within-study sample size. This, in turn,
393 results in imprecise estimates of the sampling variances. However, using a pooled CV improves the
394 estimates of the sampling variances, and all subsequent downstream analyses. Improvements in the
395 estimation of the overall mean were demonstrated by Doncaster and Spake (2018), while our
396 simulation has shown that this method also improves the accuracy of heterogeneity estimates (i.e.,
397 variance components). Incidentally, Doncaster and Spake (2018) are not the first one to use the
398 'averaging' method. For example, Hedges and Olkin (1985) also proposed to use the average of the

observed standardized mean differences in the computation of their sampling variances when meta-analyzing a large number of small studies. Also, Hunter and Schmidt (Hunter & Schmidt 1990) proposed to use the weighted average of correlations in the sampling variance for the correlation coefficient. Similarly, Berkey et al. (1995) suggested using averages of counts or proportions in the equations for computing the sampling variances of log relative risks and odds ratios, which led to less biased estimates.

There were two conditions where Method 1B could result in biased estimates. The first scenario is when CVs are very different between studies and the number of studies is relatively small (e.g., $K < 20$). In such cases, we would be able to use Method 1A (or alternatively Method 3, although the latter is more difficult to implement) to check whether meta-analytic results are drastically different. In addition, a meta-analysis of lnCVR (log CV ratio) or lnCV (log CV) could help to evaluate how large the between-study variance in CV is (Nakagawa *et al.* 2015; Senior *et al.* 2020). Large variation in between-study CVs would violate our assumption that the CV stays fairly constant for our new methodologies (cf. Nakagawa *et al.* 2015). Note, however, that our simulation shows this assumption is less important when we have a relatively large number of studies (e.g., $K > 20$). The second scenario is when there is very low total heterogeneity ($\tau^2 = \sigma_s^2 + \sigma_u^2$, which usually translates to low I^2; see Higgins *et al.* 2003; Nakagawa & Santos 2012; also see Borenstein *et al.* 2017). As mentioned earlier, heterogeneity is typically high in eco-evolutionary meta-analyses. Indeed, Senior et al (2016) showed that on average, ecological and evolutionary meta-analyses have high heterogeneity with I^2 of around 90%.

Importantly, our simulation results are based on the assumptions that SDs are missing completely at random, and that SDs (or CVs) do not change systematically across studies. Regardless of whether these assumptions are true or not, we recommend that the default method should be Method 1A. This is because: 1) almost all meta-analytic datasets suitable for lnRR would have missing SD, 2)

we should avoid not using these data without SDs, and 3) Method 1A makes fewer assumptions than Method 1B. At the same time, one should always accompany Method 1A with Method 1B, which had the best performance in our simulation. Also, one may decide to use Method 1B in the main analysis when sample sizes (replications) per effect size are consistently low – a condition where Method 1B works best. Yet, in many cases, both Method 1A and 1B would provide quantitatively similar results. However, it is when results diverge quantitatively or qualitatively that we need to be careful about the robustness of one's results (cf. our worked example). These two approaches can be considered as each other's sensitivity analysis (Noble *et al.* 2017). Taken together, we recommend conducting a meta-analysis using Method 1A and 1B in tandem, and it is straightforward to do both (see Supplementary Information).

Finally, we emphasize that these methods represent an alternative, and not a replacement, to multiple imputation, MI. Indeed, if there are missing values in moderators, the only way to deal with such missing data is to use MI. However, our proposed methods (i.e., Method 1A & 1B) are easier to implement and readily extendable to complex models, as we showed above, especially when we do not have any missing data in moderators or are less concerned about missing data in moderators. We hope that meta-analysts in ecology and evolution will adopt these two new approaches (Method 1A & 1B) to improve their meta-analytic estimation under many circumstances. Importantly, we should all be aware of the limitations of the lnRR for meta-analysis, incorporating routine assessment of the underlying assumptions using the improved Geary's test.

ACKNOWLEDGEMENTS

SN, ML & DAWN were supported by an ARC (Australian Research Council) Discovery grant (DP200100367).

AUTHORS' CONTRIBUTIONS

SN and WV came up with the initial idea and statistical methods, which were discussed and expanded by the other co-authors. AMS led the simulation study, and DWAN put together Supplementary Information with the others' inputs. SN, AMS & DWAN wrote the first draft and all the authors edited and commented on earlier versions of the manuscript.

DATA AVAILABILITY

All relevant code and data can be found at the GitHub repository (https://github.com/AlistairMcNairSenior/Miss_SD_Sim)

REFERENCES

1. Berkey, C.S., Hoaglin, D.C., Mosteller, F. & Colditz, G.A. (1995). A random-effects regression model for meta-analysis. *Stat Med*, 14, 395-411.
2. Bird, G., Kaczvinsky, C., Wilson, A.E. & Hardy, N.B. (2019). When do herbivorous insects compete? A phylogenetic meta-analysis. *Ecol Lett*, 22, 875-883.
3. Borenstein, M., Higgins, J.P.T., Hedges, L.V. & Rothstein, H.R. (2017). Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. *Res Synth Methods*, 8, 5-18.
4. Cinar, O., Nakagawa, S. & Viechtbauer, W. (2022). Phylogenetic multilevel meta-analysis: A simulation study on the importance of modelling the phylogeny. *Methods Ecol Evol*, 13, 383-395.
5. Doncaster, C.P. & Spake, R. (2018). Correction for bias in meta-analysis of little-replicated studies. *Methods Ecol Evol*, 9, 634-644.
6. Ellington, E.H., Bastille-Rousseau, G., Austin, C., Landolt, K.N., Pond, B.A., Rees, E.E. *et al.* (2015). Using multiple imputation to estimate missing data in meta-regression. *Methods Ecol Evol*, 6, 153-163.
7. Fletcher, D. & Dixon, P.M. (2012). Modelling data from different sites, times or studies: weighted vs. unweighted regression. *Methods Ecol Evol*, 3, 168-176.
8. Gurevitch, J., Koricheva, J., Nakagawa, S. & Stewart, G. (2018). Meta-analysis and the science of research synthesis. *Nature*, 555, 175-182.
9. Hedges, L.V., Gurevitch, J. & Curtis, P.S. (1999). The Meta-Analysis of Response Ratios in Experimental Ecology. *Ecology*, 80, 1150-1156.
10.

Hedges, L.V. & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press, Orlando.

Higgins, J.P.T. & Thompson, S.G. (2002). Quantifying heterogeneity in a meta-analysis. *Stat Med*, 21, 1539-1558.

Higgins, J.P.T., Thompson, S.G., Deeks, J.J. & Altman, D.G. (2003). Measuring inconsistency in meta-analyses. *Brit Med J*, 327, 557-560.

Hunter, J.E. & Schmidt, F.L. (1990). *Methods of meta-analysis : correcting error and bias in research findings*. Sage, Newbury Park ; London.

Kambach, S., Bruelheide, H., Gerstner, K., Gurevitch, J., Beckmann, M. & Seppelt, R. (2020). Consequences of multiple imputation of missing standard deviations and sample sizes in meta-analysis. *Ecol Evol*, 10, 11699-11712.

Koricheva, J. & Gurevitch, J. (2014). Uses and misuses of meta-analysis in plant ecology. *J Ecol*, 102, 828-844.

Lajeunesse, M.J. (2013). Recovering missing or partial data from studies: a survey. In: *The handbook of meta-analysis in ecology and evolution* (eds. Koricheva, J, Gurevitch, J & Mengersen, K). Princeton University Press Princeton, pp. 195-206.

Lajeunesse, M.J. (2015). Bias and correction for the log response ratio in ecological meta-analysis. *Ecology*, 96, 2056-2063.

Lin, L.F. & Aloe, A.M. (2021). Evaluation of various estimators for standardized mean difference in meta-analysis. *Stat Med*, 40, 403-426.

McDonald, S.E., Lawrence, R., Kendal, L. & Rader, R. (2019). Ecological, biophysical and production effects of incorporating rest into grazing regimes: A global meta-analysis. *J Appl Ecol*, 56, 2723-2731.

Nakagawa, S. (2015). Missing data: mechanisms, methods and messages. In: *Ecological statistics: contemporary theory and application* (eds. Fox, GA, Negrete-Yankelevich, S & Sosa, VJ). Oxford University Press Oxford, pp. 81-105.

Nakagawa, S. & Freckleton, R.P. (2008). Missing inaction: The dangers of ignoring missing data. *Trends in Ecology & Evolution*, 23, 592-596.

Nakagawa, S. & Freckleton, R.P. (2011). Model averaging, missing data and multiple imputation: a case study for behavioural ecology. *Behav Ecol Sociobiol*, 65, 103-116.

Nakagawa, S., Poulin, R., Mengersen, K., Reinhold, K., Engqvist, L., Lagisz, M. *et al.* (2015). Meta-analysis of variation: ecological and evolutionary applications and beyond. *Methods Ecol Evol*, 6, 143-152.

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

Nakagawa, S., Senior, A.M., Viechtbauer, W. & Noble, D.W.A. (2022). An assessment of statistical methods for nonindependent data in ecological meta-analyses: Comment. *Ecology*, 103, e03490.

26.

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

27.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley, New York ; Chichester.

28.

Senior, A.M., Grueber, C.E., Kamiya, T., Lagisz, M., O'Dwyer, K., Santos, E.S.A. *et al.* (2016). Heterogeneity in ecological and evolutionary meta-analyses: its magnitude and implications. *Ecology*, 97, 3293-3299.

29.

Senior, A.M., Viechtbauer, W. & Nakagawa, S. (2020). Revisiting and expanding the meta-analysis of variation: The log coefficient of variation ratio. *Res Synth Methods*, 11, 553-567.

30.

Spake, R., Mori, A.S., Beckmann, M., Martin, P.A., Christie, A.P., Duguid, M.C. *et al.* (2021). Implications of scale dependence for cross-study syntheses of biodiversity differences. *Ecol Lett*, 24, 374-390.

31.

van Buuren, S. (2018). *Flexible Imputation of Missing Data, Second Edition*. 2nd edn. CRC Press, Boca Raton.

32.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36, 1-48.

33.

Wickham, H. (2009). *ggplot2 : elegant graphics for data analysis*. Springer, New York ; London.

34.

Xiong, C.J., Miller, J.P. & Morris, J.C. (2010). Measuring Study-Specific Heterogeneity in Meta-Analysis: Application to an Antecedent Biomarker Study of Alzheimer's Disease. *Stat Biopharm Res*, 2, 300-309.

35.

Yang, Y., Hillebrand, H., Lagisz, M., Cleasby, I. & Nakagawa, S. (2022). Low statistical power and overestimated anthropogenic impacts, exacerbated by publication bias, dominate field studies in global change biology. *Glob Chang Biol*, 28, 969-989.

578 **Figure Legends**

579

580 **Figure 1**

581 Visual schematics of a hypothetical dataset with missing standard deviations (SDs) and five
582 different approaches used in this study, including 3 new methods. The symbols: $\ln RR_2$ (Equation 4),
583 $\ln RR_3$ (Equation 6), v (Equation 5), $\tilde{v}$ (Equation 7), and $\phi\tilde{v}$ (Equation 12). Note that, under some
584 circumstances, we could replace Equations 4 & 6 with Equation 1 while Equation 7 can be replaced
585 by Equation 15 (see the text for more details).

586

587 **Figure 2**

588 Results on overall meta-analytic mean from multi-level meta-analytic models: A) Violin plot
589 showing the distribution of median bias in the estimated effect under each simulated condition as a
590 function of the method used to handle missing data (distribution assuming full data shown for
591 reference). B) Pairwise correlations between the degree of bias under each simulated condition for
592 each method. C) Distribution of the difference between Method 1A and 1B in the absolute degree of
593 bias under each condition (positive values indicate greater median bias under Method 1A). D)
594 Violin plot showing the distribution of range bias ($\log_{10}$ transformed) in the estimated effect under
595 each simulated condition as a function of the method used to handle missing data. E. Violin plot
596 showing the distribution of range bias ($\log_{10}$ transformed) in the estimated effect under each
597 simulated condition as a function of the degree of heterogeneity in SDs among studies and typical
598 size of studies in the meta-analysis. Our plots were drawn using the R package ggplot2 (Wickham
599 2009).

600

601 **Figure 3**

602 Results on coverage from multi-level meta-analytic models: A) Violin plot showing the distribution
603 of coverage of 95% CIs under each simulated condition as a function of the method used to handle

missing data (distribution assuming full data shown for reference). B) Violin plot showing the distribution of coverage under each simulated condition as a function of the simulated level of total heterogeneity and the ICC for study using method 1.1 to handle missing SDs. C) Violin plot showing the distribution of coverage under each simulated condition as a function of the simulated level of total heterogeneity and the ICC for study using method 1.2 to handle missing SDs.

Figure 4

Results on coverage from multi-level meta-analytic models: A) Violin plot showing the distribution of median bias in the estimated heterogeneity under each simulated condition as a function of the method used to handle missing data (distribution assuming full data shown for reference). Bias in heterogeneity is calculated as the log ratio of the estimated and parametrized value. B. Box plot showing the median bias in estimated heterogeneity under each simulated condition as a function of the method used to handle missing data (colours as in panel A), and the simulated level of heterogeneity. C) Violin plot showing the distribution of the median bias in the estimated ICC for study under each simulated condition as a function of the method used to handle missing data. Bias in the ICC was calculated as the difference between the estimated and parameterized value. D) Violin plot showing the distribution of the median bias in the estimated ICC for study under each simulated condition as a function of the simulated level of total heterogeneity and the ICC for study using method 1.2 to handle missing SDs.

Figure S1

Results on overall meta-analytic mean from multi-level meta-analytic models: A) Violin plot showing the distribution of median bias in the estimated effect under each simulated condition as a function of the method used to handle missing data (distribution assuming full data shown for reference). B) Pairwise correlations between the degree of bias under each simulated condition for each method. C) Distribution of the difference between Method 1A and 1B in the absolute degree of

630 bias under each condition (positive values indicate greater median bias under Method 1A). D)
631 Violin plot showing the distribution of range bias (log_{10} transformed) in the estimated effect under
632 each simulated condition as a function of the method used to handle missing data. E. Violin plot
633 showing the distribution of range bias (log_{10} transformed) in the estimated effect under each
634 simulated condition as a function of the degree of heterogeneity in SDs among studies and typical
635 size of studies in the meta-analysis.
636
637 **Figure S2**
638 Results on coverage from multi-level meta-analytic models: A) Violin plot showing the distribution
639 of coverage of 95% CIs under each simulated condition as a function of the method used to handle
640 missing data (distribution assuming full data shown for reference). B) Violin plot showing the
641 distribution of coverage under each simulated condition as a function of the simulated level of total
642 heterogeneity and the ICC for study using method 1.1 to handle missing SDs. C) Violin plot
643 showing the distribution of coverage under each simulated condition as a function of the simulated
644 level of total heterogeneity and the ICC for study using method 1.2 to handle missing SDs.
645
646 **Figure S3**
647 Results on coverage from random-effects meta-analytic models: A) Violin plot showing the
648 distribution of median bias in the estimated heterogeneity under each simulated condition as a
649 function of the method used to handle missing data (distribution assuming full data shown for
650 reference). Bias in heterogeneity is calculated as the log ratio of the estimated and parametrized
651 value. B. Box plot showing the median bias in estimated heterogeneity under each simulated
652 condition as a function of the method used to handle missing data (colours as in panel A), and the
653 simulated level of heterogeneity.
654
655

Table 1 Different methods including the case with no missing data, used for simulations and which equations were used for which method

Methods	**Point estimate**[1]	**Sampling variance (full data)**	**Sampling variance (missing SD)**
Method 1A	Equations 4 and 6	Equation 5	Equations 7
Method 1B	Equations 4 and 6	Equation 7	Equations 7
Method 2	Equations 4 and 6	Equation 12	Equation 12
Method 3	Equations 4 and 6	Equations 5	Equation 12
No missing data	Equation 4	Equations 5	Not applicable

[1] Applying both Equations 4 & 6 (the latter for observations/rows with missing SD) or applying only Equation 6 for all observations when SDs are missing would make little differences for (effect size) point estimates, unless effect sizes fulfill Equation 14.

662 **Table 2** Results from the re-analyses of a subset of data from Bird et al. (2019) using the methods
663 we propose to deal with missing SD data estimating the overall effects of competition on focal
664 insect abundance (LCI = lower, or 2.5%, confidence limit; UCI = upper, or 97.5%, confidence
665 limit). Note that these results exclude effect size estimates that fail Geary's test.

Method	Est.	SE	95% LCI	95% UCI
Full Data	0.202	0.085	0.036	0.369
Complete Case	0.176	0.102	-0.024	0.377
Method 1A	0.186	0.091	0.008	0.364
Method 1B	0.146	0.096	-0.043	0.334
Method 2	0.192	0.083	0.03	0.354
Method 3	0.185	0.086	0.017	0.353

666

667

668

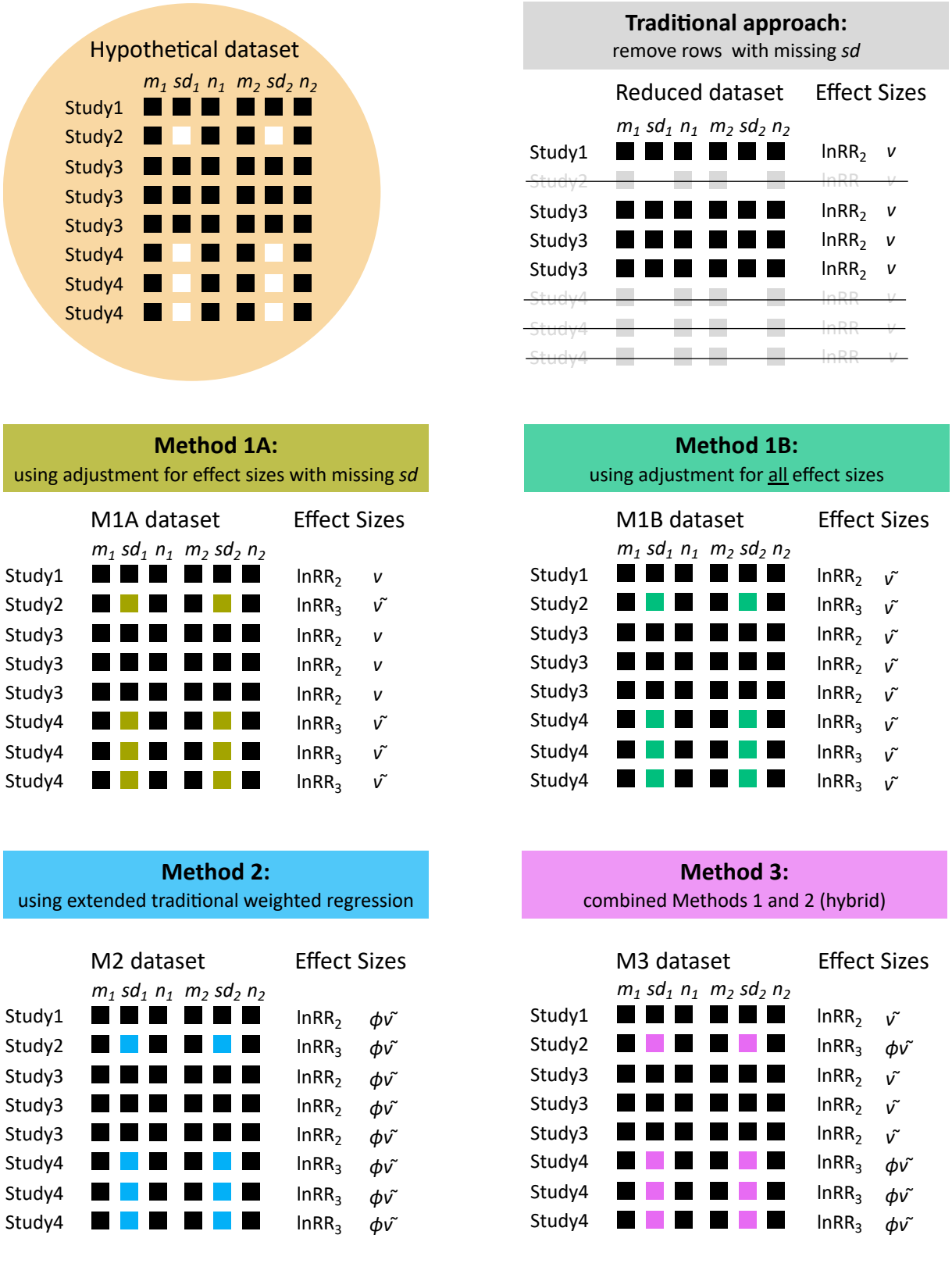
669 **Table S1.** Variables/parameters in simulations.

Variable (Notation)	Description and details	Value(s)
% Studies Missing SD	Percentage of studies that have missing SDs	5, 15, 25, 35, 45 or 55
Overall Effect Size (θ)	The overall mean lnRR effect size	0.3
Number of Studies (K)	Total number of studies within the meta-analytic dataset	12, 30, 100
Number of Effect Sizes in Study (L)	The number of effect sizes within a study. Values for each study were randomly distributed using a double Poisson distribution	Random with mean of 5, and dispersion 1.5
Total Heterogeneity (τ^2)	The total heterogeneity among effect sizes	0.09
Intra-Class Correlation for Study (ICC_s)	The proportion of total heterogeneity that is attributable to study-level effects	0 or 0.5
Sample Size in Study (N)	The sample size of groups within studies; individual sample sizes were randomly distributed using a double Poisson distribution	Random with mean of either 5 or 30, and dispersion 1.5
Standard Deviation in Study (S)	The within-study SDs. Individual within-study SDs were randomly distributed following a Gamma distribution	Random with mean 15 and SD of either 10^{-10}, 3.75 or 7.5

670

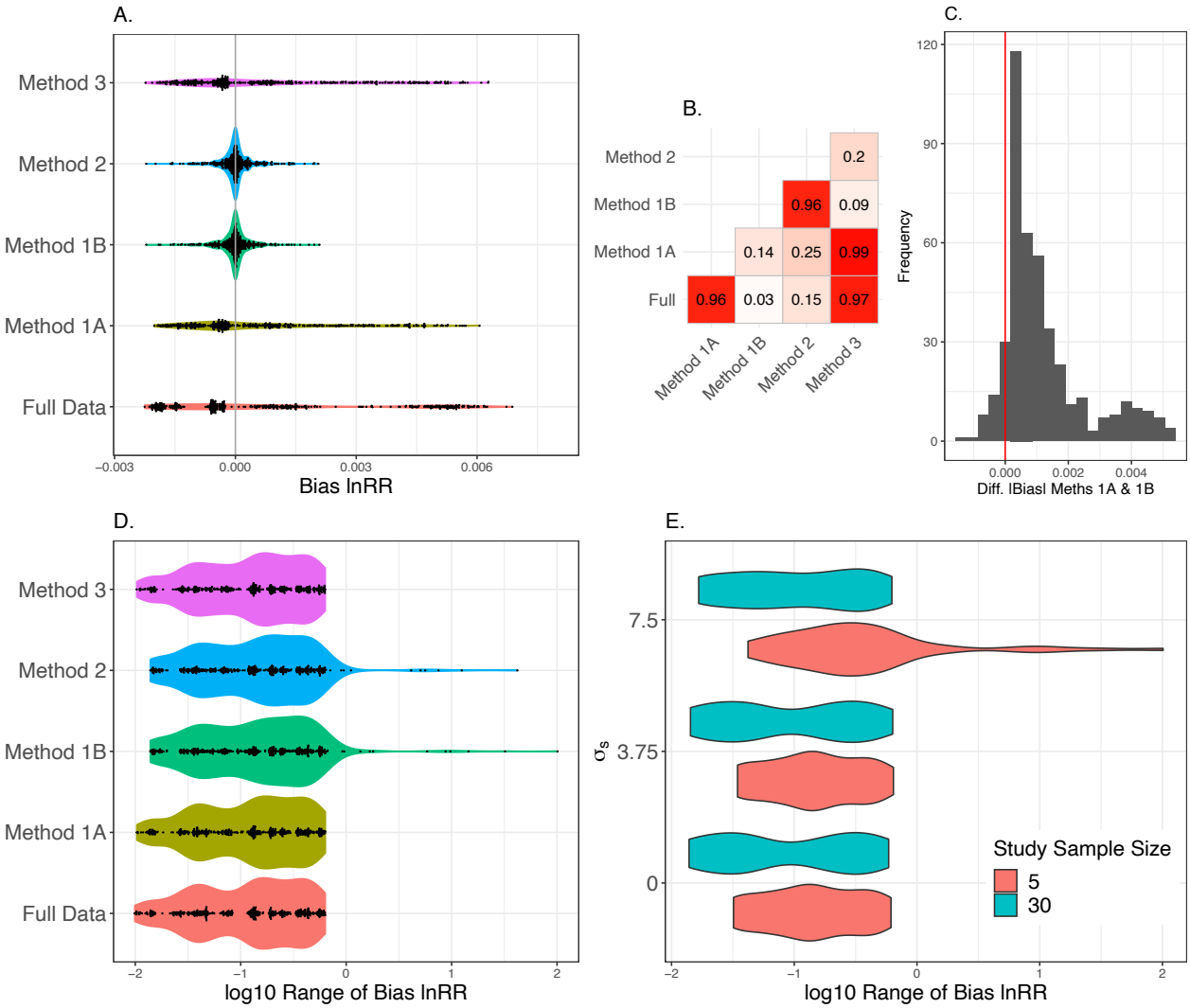
671

672 **Figure 1**



673
674

675 **Figure 2**



676

677

678

Figure 3

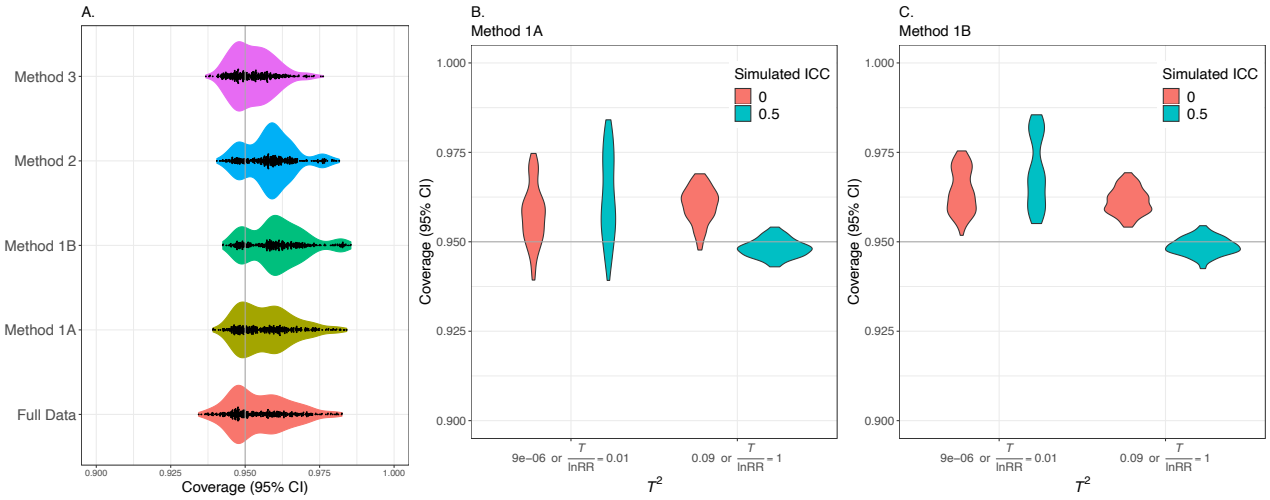
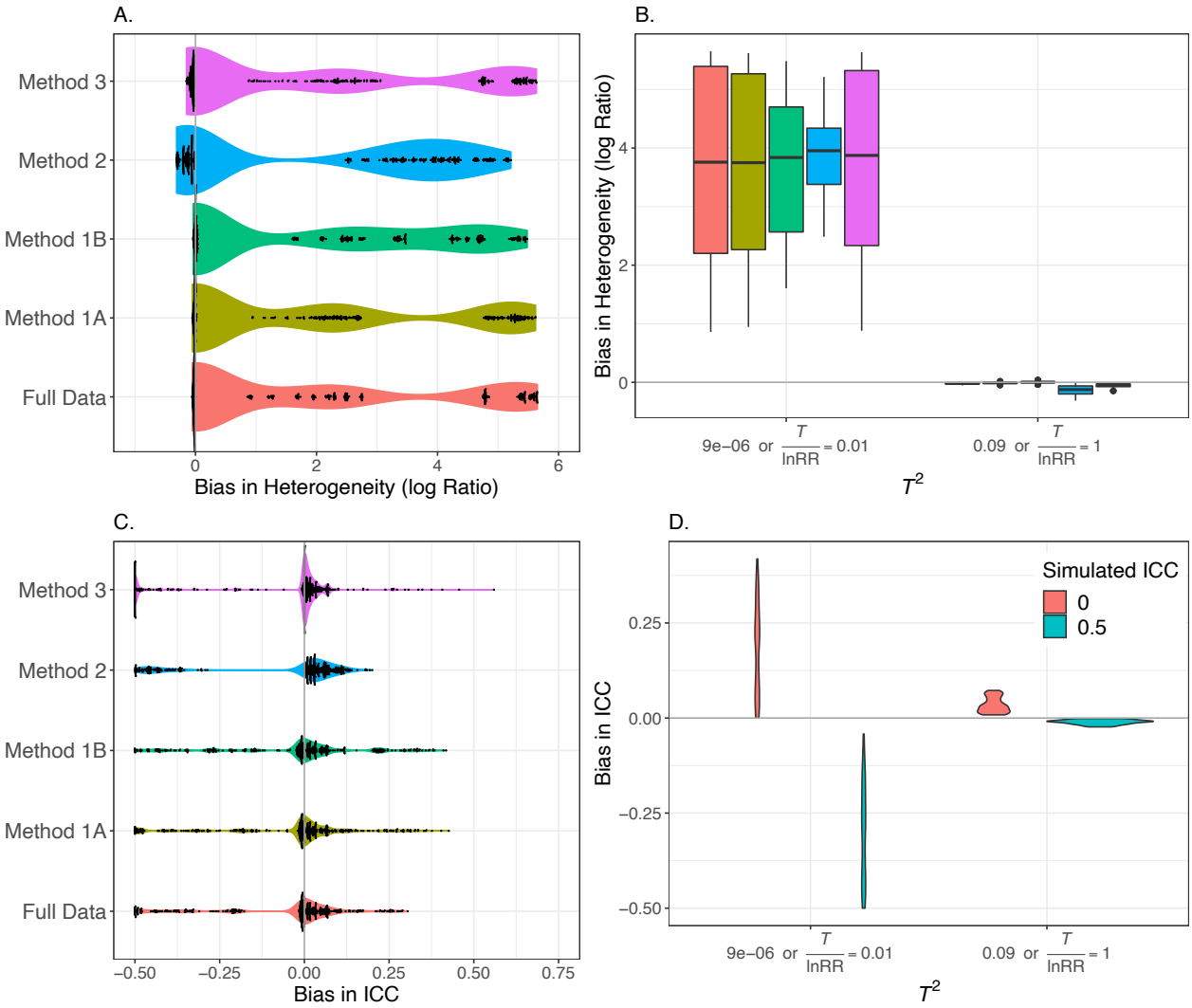
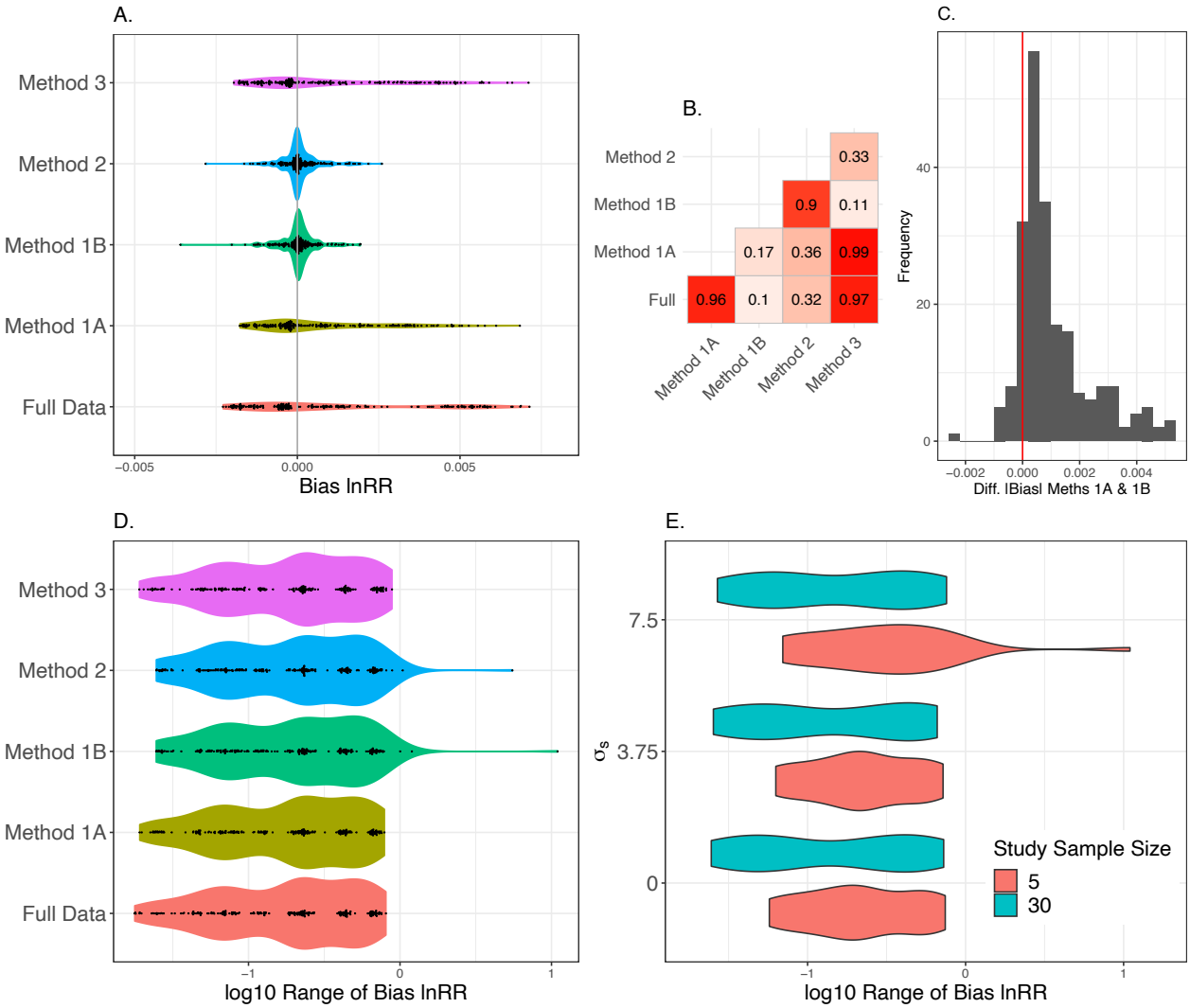

A.
Method 3
Method 2
Method 1B
Method 1A
Full Data
0.900
0.925
0.950
0.975
1.000
Coverage (95% CI)
B.
Method 1A
Simulated ICC
0
0.5
Coverage (95% CI)
9e−06 or $\frac{T}{\ln RR} = 0.01$
0.09 or $\frac{T}{\ln RR} = 1$
T^2
C.
Method 1B
Simulated ICC
0
0.5
Coverage (95% CI)
9e−06 or $\frac{T}{\ln RR} = 0.01$
0.09 or $\frac{T}{\ln RR} = 1$
T^2

Figure 4


A.
Method 3
Method 2
Method 1B
Method 1A
Full Data
0
2
4
6
Bias in Heterogeneity (log Ratio)
B.
Bias in Heterogeneity (log Ratio)
0
2
4
9e−06 or T/lnRR = 0.01
0.09 or T/lnRR = 1
T²
C.
Method 3
Method 2
Method 1B
Method 1A
Full Data
−0.50
−0.25
0.00
0.25
0.50
0.75
Bias in ICC
D.
Simulated ICC
0
0.5
Bias in ICC
0.25
0.00
−0.25
−0.50
9e−06 or T/lnRR = 0.01
0.09 or T/lnRR = 1
T²

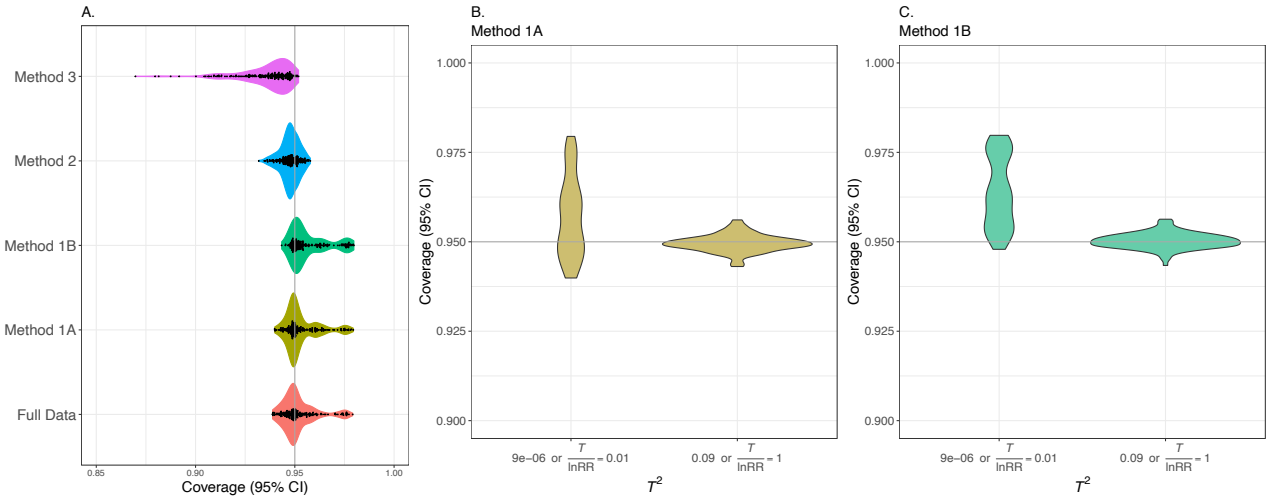
688 **Figure S1**



689

690

691 **Figure S2**

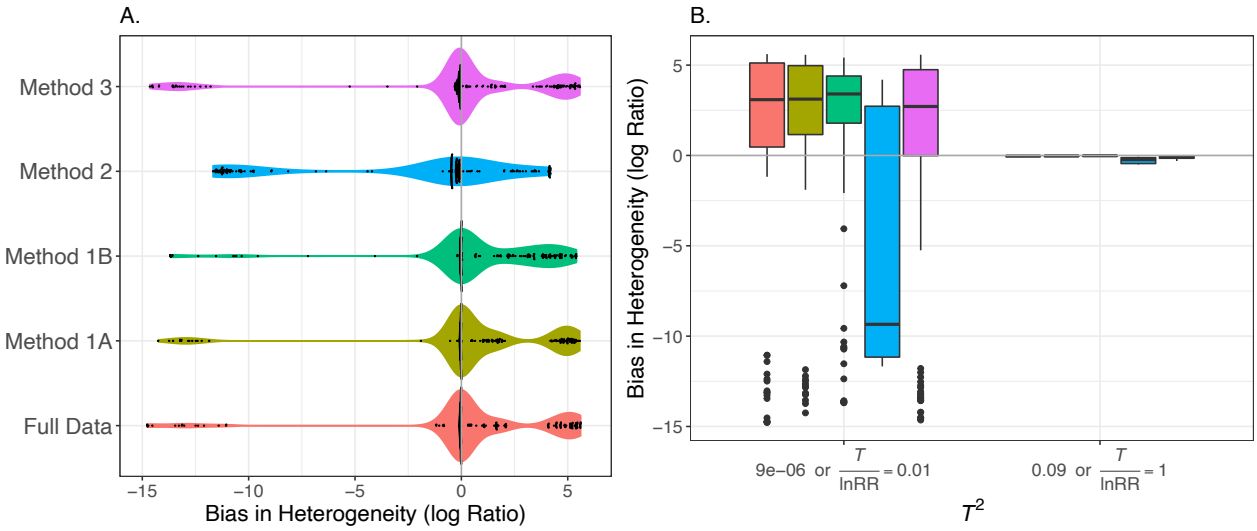


692

693

694

Figure S3


A.
Method 3
Method 2
Method 1B
Method 1A
Full Data
−15
−10
−5
0
5
Bias in Heterogeneity (log Ratio)
B.
Bias in Heterogeneity (log Ratio)
5
0
−5
−10
−15
9e−06 or $\frac{T}{\text{lnRR}} = 0.01$
0.09 or $\frac{T}{\text{lnRR}} = 1$
T^2