Received: 20 August 2017 | Revised: 24 April 2018 | Accepted: 2 July 2018

DOI: 10.1002/jrsm.1316

RESEARCH ARTICLE

WILEY Research Synthesis Methods

A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses

Dean Langan[1,7] | Julian P.T. Higgins[2] | Dan Jackson[3] | Jack Bowden[2] | Areti Angeliki Veroniki[4,5,6] | Evangelos Kontopantelis[7] | Wolfgang Viechtbauer[8] | Mark Simmonds[9]

[1]Great Ormond Street Institute of Child Health, UCL, London, UK

[2]School of Social and Community Medicine, University of Bristol, Bristol, UK

[3]Statistical Innovation Group, AstraZeneca, Cambridge, UK

[4]Li Ka Shing Knowledge Institute, St. Michael's Hospital, 209 Victoria Street, East Building, Toronto, Ontario, M5B 1T8

[5]Department of Primary Education, School of Education, University of Ioannina, Ioannina, Greece

[6]Institute of Reproductive and Developmental Biology, Department of Surgery & Cancer, Faculty of Medicine, Imperial College, London, W12 0NN, UK

[7]Centre for Health Informatics, Institute of Population Health, University of Manchester, Manchester, UK

[8]Department of Psychiatry and Neuropsychology, Maastricht University, Maastricht, The Netherlands

[9]Centre for Reviews and Dissemination, University of York, York, UK

Correspondence
Dean Langan, Great Ormond Street Institute of Child Health, UCL, London WC1E 6BT, UK.
Email: d.langan@ucl.ac.uk

Studies combined in a meta-analysis often have differences in their design and conduct that can lead to heterogeneous results. A random-effects model accounts for these differences in the underlying study effects, which includes a heterogeneity variance parameter. The DerSimonian-Laird method is often used to estimate the heterogeneity variance, but simulation studies have found the method can be biased and other methods are available. This paper compares the properties of nine different heterogeneity variance estimators using simulated meta-analysis data. Simulated scenarios include studies of equal size and of moderate and large differences in size. Results confirm that the DerSimonian-Laird estimator is negatively biased in scenarios with small studies and in scenarios with a rare binary outcome. Results also show the Paule-Mandel method has considerable positive bias in meta-analyses with large differences in study size. We recommend the method of restricted maximum likelihood (REML) to estimate the heterogeneity variance over other methods. However, considering that meta-analyses of health studies typically contain few studies, the heterogeneity variance estimate should not be used as a reliable gauge for the extent of heterogeneity in a meta-analysis. The estimated summary effect of the meta-analysis and its confidence interval derived from the Hartung-Knapp-Sidik-Jonkman method are more robust to changes in the heterogeneity variance estimate and show minimal deviation from the nominal coverage of 95% under most of our simulated scenarios.

KEYWORDS
DerSimonian-Laird, heterogeneity, random-effects, REML, simulation

1 | INTRODUCTION

Meta-analysis is the statistical technique of combining the results of multiple comparable studies. These studies often have differences in their design and conduct that lead to heterogeneity in their underlying effects. When heterogeneity is thought to be present, researchers should first attempt to find its causes, but these causes may be too numerous to isolate or may simply be unknown. Unexplained heterogeneity of study effects can be quantified in a random-effects model. This model typically assumes a normal distribution of the underlying effects across studies. A reliable estimate of the variance of this distribution can provide valuable insight into the degree

of heterogeneity between studies, even if such studies are not formally synthesised in a meta-analysis.

The moment-based method proposed by DerSimonian and Laird[1] is most commonly used to estimate the heterogeneity variance. However, this method has been shown in previous simulation studies to be negatively biased in meta-analyses containing small studies,[2] particularly in meta-analyses of binary outcomes.[3,4] There are many other available methods,[5] including those proposed by Paule and Mandel,[6] Hartung and Makambi,[7] Sidik and Jonkman,[4,8] and the restricted maximum likelihood (REML) method.[9] Estimates derived from these methods in the same meta-analysis can often be notably different, and in a small number of cases, these estimates can produce discordant conclusions on the summary effect and its confidence interval.[10] Therefore, the choice of heterogeneity variance method is an important consideration in a meta-analysis. Research based on simulated meta-analysis data can allow a researcher to make a more informed decision.

A recent systematic review collated simulation studies that compare the properties of heterogeneity variance estimators.[11] Its aim was to assess if there is consensus on which heterogeneity variance methods (if any) have better properties than DerSimonian-Laird (DL). The review identified 12 relevant simulation studies, but there was little consensus across the various authors' recommendations.[2-4,8,12-19] This may have been caused by a potential conflict of interest among the authors of all but four of these studies[3,12,13,17]; the authors of these eight studies recommended their own newly proposed methods over existing methods. Three of the simulation studies[3,12,13] compared only preexisting methods and made an explicit recommendation for estimating the heterogeneity variance; the authors of these studies recommended the method of Paule and Mandel[6] and/or REML,[9] but only compared a subset of methods.

The tentative conclusions of that review provided motivation for a new simulation study, which we present in this paper. The limitations of previous simulation studies helped inform the design of this study. We consider the inclusion of all methods identified in recent reviews of heterogeneity variance methods,[5,11] compare methods comprehensively in a range of simulated scenarios representative of meta-analyses of health studies, and report a wide range of performance measures. Performance measures include those that relate directly to the heterogeneity variance estimates and those that measure the impact of heterogeneity variance estimates on the summary effect estimate and its confidence interval. Our recommendations are based on a subjective trade-off between many performance measures. To minimise any conflict of interest, we do not propose any new methods in this paper.

The aims of this simulation study are to (1) compare the relative performance of heterogeneity variance methods to establish which method(s) have the most reasonable properties and (2) find scenarios where the performance of all methods is poor, such that we cannot rely on a single method to provide an estimate. In scenarios where all methods perform poorly, we make wider recommendations for random-effects meta-analysis and dealing with between-study heterogeneity.

The outline of the paper is as follows. In Section 2, we introduce methods for estimating the heterogeneity variance and any other meta-analysis methods relevant to this simulation study. The design of the simulation study is given in Section 3, followed by the results of this study in Section 4. Results are discussed and conclusions are drawn in Sections 5 and 6, respectively.

2 | METHODS

2.1 | The heterogeneity variance parameter in a random-effects model

A random-effects model accounts for the possibility that underlying effects differ between studies in a meta-analysis. The random-effects model is defined as

$$\widehat{\theta}_i = \theta_i + \varepsilon_i \quad \theta_i = \theta + \delta_i, \tag{1}$$

where θ_i is the true effect size in study i, $\widehat{\theta}_i$ is the estimated effect size, and θ is the average effect across all studies. ε_i and δ_i are the within-study errors and the between-study heterogeneity, respectively. Meta-analysis methods typically assume that both are normally distributed, ie, $\varepsilon_i \sim N\left(0, \sigma_i^2\right)$ and $\delta_i \sim N(0, \tau^2)$. The heterogeneity variance parameter is a measure of the variance of θ_i around θ and is denoted by τ^2.

The inverse-variance method is most commonly used to estimate θ in this model; the estimate is given by

$$\widehat{\theta} = \sum_{i=1}^{k} w_i \widehat{\theta}_i / \sum_{i=1}^{k} w_i, \tag{2}$$

where k is the number of studies in the meta-analysis and w_i is the weight given to study i.

Under the random-effects model, using weights $w_i = 1/\left(\sigma_i^2 + \tau^2\right)$ provides the uniformly minimum variance unbiased estimator (UMVUE) of θ, which we denote by $\widehat{\theta}_{RE}$. When $\tau^2 = 0$, model (1) simplifies to what is commonly referred to as the fixed-effect model, where the true effects are homogeneous. In that case, the UMVUE of θ (which is now the common true effect for all k studies) is obtained with (2), but using weights $w_i = 1/\sigma_i^2$. We denote this estimator by $\widehat{\theta}_{FE}$. However, the variance parameters σ_i^2 and τ^2 are unknown in practice and must be estimated

from the data. Methods to estimate τ^2 are outlined in the next section.

2.2 | Heterogeneity variance estimators

Nine estimators were identified from two systematic reviews of heterogeneity variance methods.[5,11] Estimators proposed by Hunter and Schmidt,[20] Rukhin,[14] Malzahn et al,[2] and the maximum likelihood method proposed by Hardy and Thompson[21] are present in these reviews but excluded from the main results because preliminary analysis showed they are clearly inferior to other methods (as shown in Appendix S1). Furthermore, Bayesian methods that rely on a subjective choice of prior distribution are excluded because of difficulty in objectively comparing them to frequentist methods. The method proposed by Morris[22] is excluded because it is an approximation to REML. We excluded the positive DL estimator,[18] which truncates heterogeneity variance estimates below 0.01, because any positive cut-off value could be applied.

The included heterogeneity variance estimators are listed in Table 1. This table also includes acronyms for the estimators used throughout this paper. Their formulae are given as follows.

2.3 | Method of moments approach (estimators 1-5)

Five estimators included in this study can be derived from the method of moments approach, which is based on the generalised Q statistic[23]:

$$Q_{MM} = \sum_{i=1}^{k} a_i \left(\widehat{\theta}_i - \widehat{\theta}\right)^2,$$

TABLE 1 Nine heterogeneity variance estimators included in this study

	Estimator	Acronym
Method of moments estimators (truncated)		
1	DerSimonian-Laird	DL
2	Cochran ANOVA	CA
3	Paule-Mandel	PM
4	Two-step Cochran ANOVA	PM_{CA}
5	Two-step DerSimonian-Laird	PM_{DL}
Nontruncated estimators		
6	Hartung-Makambi	HM
7	Sidik-Jonkman	SJ
8	Alternative Sidik-Jonkman	SJ_{CA}
Maximum likelihood estimators		
9	Restricted maximum likelihood	REML

The weight assigned to study i is denoted by a_i and calculated differently depending on which of the five method of moments estimators is used. $\widehat{\theta}$ is given by Formula (2) with study weights $w_i = a_i$. By equating Q_{MM} to its expected value, the following general formula for the heterogeneity variance can be derived[23]:

$$\widehat{\tau}^2 = max\left\{0, \frac{Q_{MM} - \sum_{i=1}^{k} a_i \widehat{\sigma}_i^2 + \frac{\sum_{i=1}^{k} a_i^2 \widehat{\sigma}_i^2}{\sum_{i=1}^{k} a_i}}{\sum_{i=1}^{k} a_i - \frac{\sum_{i=1}^{k} a_i^2}{\sum_{i=1}^{k} a_i}}\right\}. \quad (3)$$

1. The DL estimator[1] uses the fixed-effect model weights $a_i = 1/\widehat{\sigma}_i^2$, which leads to the formula:

$$\widehat{\tau}^2_{DL} = max\left\{0, \frac{\sum_{i=1}^{k}\left(1/\widehat{\sigma}_i^2\right)\left(\widehat{\theta}_i - \widehat{\theta}_{FE}\right)^2 - (k-1)}{\sum_{i=1}^{k}\left(1/\widehat{\sigma}_i^2\right) - \frac{\sum_{i=1}^{k}\left(1/\widehat{\sigma}_i^2\right)^2}{\sum_{i=1}^{k}\left(1/\widehat{\sigma}_i^2\right)}}\right\}.$$

2. Cochran's ANOVA (CA) estimator uses equal study weights $a_i = 1/k$, leading to

$$\widehat{\tau}^2_{CA} = max\left\{0, \frac{1}{k-1}\sum_{i=1}^{k}\left(\widehat{\theta}_i - \widehat{\theta}_{CA}\right)^2 - \frac{1}{k}\sum_{i=1}^{k}\widehat{\sigma}_i^2\right\},$$

where $\widehat{\theta}_{CA}$ is calculated from Formula (2) with study weights $w_i = 1/k$.

3. The Paule-Mandel (PM) estimator uses the random-effects model study weights, defined by substituting $a_i = 1/\left(\widehat{\sigma}_i^2 + \widehat{\tau}^2_{PM}\right)$ into Formula (3). Since a_i is a function of $\widehat{\tau}^2_{PM}$, there is no closed-form expression for $\widehat{\tau}^2_{PM}$, and iteration is required to find the solution. Iterative algorithms including those suggested by Bowden et al[24] and Jackson et al[25] always converge. The same estimator has been derived independently of the methods of moments approach and is therefore often referred to as the empirical Bayes estimator in the literature.[26]
4. The two-step CA estimator also uses PM random-effects weights but restricts iteration to two steps (PM_{CA}). Cochran's ANOVA is used to initially estimate τ^2; thus, a closed-form expression can be derived by substituting $a_i = 1/\left(\widehat{\sigma}_i^2 + \widehat{\tau}^2_{CA}\right)$ into Formula (3).
5. The two-step DL estimator (PM_{DL}) has similar weights as PM_{CA} but uses the DL method to calculate an initial estimate of τ^2. Therefore, the study weights are $a_i = 1/\left(\widehat{\sigma}_i^2 + \widehat{\tau}^2_{DL}\right)$.

All five of these methods can produce negative variance estimates and are truncated to zero in such cases.

2.4 | Hartung-Makambi (estimator 6)

Hartung and Makambi[7] proposed a correction to the DL estimator so that $\widehat{\tau}^2$ is always positive and truncation is not required. The formula is given by

$$\widehat{\tau}^2_{HM} = \frac{\left(\sum_{i=1}^{k}\left(1/\widehat{\sigma}_i^2\right)\left(\widehat{\theta}_i-\widehat{\theta}_{FE}\right)^2\right)^2}{\left(\sum_{i=1}^{k}\left(1/\widehat{\sigma}_i^2\right)-\frac{\sum_{i=1}^{k}\left(1/\widehat{\sigma}_i^2\right)^2}{\sum_{i=1}^{k}\left(1/\widehat{\sigma}_i^2\right)}\right)\left(2(k-1)+\sum_{i=1}^{k}\left(1/\widehat{\sigma}_i^2\right)\left(\widehat{\theta}_i-\widehat{\theta}_{FE}\right)^2\right)}.$$

2.5 | Sidik-Jonkman (estimators 7 and 8)

Sidik and Jonkman[8] proposed the following two-step estimator that only produces positive τ^2 estimates:

$$\widehat{\tau}^2_{SJ} = \frac{1}{k-1}\sum_{i=1}^{k}\frac{1}{1+\left(\widehat{\sigma}_i^2/\widehat{\tau}_0^2\right)}\left(\widehat{\theta}_i-\widehat{\theta}_{SJ}\right)^2,$$

where $\widehat{\tau}_0^2 = \frac{1}{k-1}\sum_{i=1}^{k}\left(\widehat{\theta}_i-\widehat{\theta}_{CA}\right)^2$ is the initial heterogeneity variance estimate and $\widehat{\theta}_{SJ}$ is calculated from Formula (2) with weights $w_i = 1/\left(1+\left(\widehat{\sigma}_i^2/\widehat{\tau}_0^2\right)\right)$.

Sidik and Jonkman[8] noted that an alternative formula for $\widehat{\tau}_0^2$ may lead to an estimator with better properties. In a subsequent paper,[4] they proposed an alternative initial estimate $\widehat{\tau}_0^2 = \max\left\{0.01, \widehat{\tau}^2_{CA}\right\}$, where $\widehat{\tau}^2_{CA}$ is CA estimate of the heterogeneity variance (estimator 2).

2.6 | Restricted maximum likelihood (estimator 9)

To derive the REML estimator, the log-likelihood function from the random-effects model (1) derived from the maximum likelihood method[21] is transformed so that it excludes the parameter θ.[9] In doing so, REML avoids making the assumption that θ is known and is therefore thought to be an improvement on the original maximum likelihood estimator.[13] This results in the following modified log-likelihood function:

$$l = -\frac{k}{2}ln(2\pi)-\frac{1}{2}\sum_{i=1}^{k}ln\left(\sigma_i^2+\tau^2\right)-\frac{1}{2}\sum_{i=1}^{k}\frac{\left(\widehat{\theta}_i-\widehat{\theta}\right)}{\sigma_i^2+\tau^2}-\frac{1}{2}ln\left(\sum_{i=1}^{k}\frac{1}{\sigma_i^2+\tau^2}\right).$$

Maximising this modified log-likelihood function with respect to τ^2 (by differentiating and setting equal to zero) results in the following formula for the heterogeneity variance:

$$\widehat{\tau}^2_{REML} = max\left\{0, \frac{\sum_{i=1}^{k}a_i^2\left(\left(\widehat{\theta}_i-\widehat{\theta}_{RE}\right)^2-\widehat{\sigma}_i^2\right)}{\sum_{i=1}^{k}a_i^2}+\frac{1}{\sum_{i=1}^{k}a_i}\right\},$$

where $a_i = 1/\left(\widehat{\sigma}_i^2+\widehat{\tau}^2_{REML}\right)$.

The heterogeneity variance estimate is calculated through a process of iteration. Fisher scoring algorithm is used for iteration in this study, as implemented in the *metafor* package in R.[27]

2.7 | Confidence interval methods for the summary effect

In this study, we also investigate how choice of a particular heterogeneity variance estimation method may impact on the estimate of the summary effect θ and its confidence interval. As we described earlier, the inverse-variance method is typically used to estimate θ in a random-effects meta-analysis, so we calculate $\widehat{\theta}$ using this method throughout. The following are three methods to estimate a corresponding confidence interval.

A Wald-type confidence interval can be calculated as[1]

$$\widehat{\theta}\pm Z_{(1-C)/2}\sqrt{Var\left(\widehat{\theta}\right)} \qquad (4)$$
$$Var\left(\widehat{\theta}\right) = 1/\left(\sum_{i=1}^{k}1/\left(\widehat{\sigma}_i^2+\widehat{\tau}^2\right)\right),$$

where C is the coverage level of the confidence interval and $Z_{(1\ -\ C)/2}$ is the $(1-C)/2$ centile of the standard normal distribution (eg, $Z_{(1\ -\ 0.95)/2} = 1.96$).

Alternatively, a t distribution can be assumed for the summary effect with $k-1$ degrees of freedom[28]:

$$\widehat{\theta}\pm t_{k-1,(1-C)/2}\sqrt{Var\left(\widehat{\theta}\right)},$$

where $t_{k\ -\ 1,\ (1\ -\ C)/2}$ is the $(1-C)/2$ centile of the t distribution with $k-1$ degrees of freedom and $Var\left(\widehat{\theta}\right)$ is calculated from Formula (4).

The Hartung-Knapp-Sidik-Jonkman (HKSJ) method[29,30] also relies on a t-distribution and uses an alternative weighted variance for $\widehat{\theta}$:

$$\widehat{\theta}\pm t_{k-1,(1-C)/2}\sqrt{Var_{HKSJ}\left(\widehat{\theta}\right)}$$
$$Var_{HKSJ}\left(\widehat{\theta}\right) = \frac{\sum_{i=1}^{k}a_i\left(\widehat{\theta}_i-\widehat{\theta}\right)^2}{(k-1)\sum_{i=1}^{k}a_i},$$

where $a_i = 1/\left(\widehat{\sigma}_i^2 + \widehat{\tau}^2\right)$, $\widehat{\theta}$ is calculated from Formula (2) and $\widehat{\tau}^2$ can be estimated using any of the methods outlined in this paper.

This method is equivalent to the t distribution method, but its variance is multiplied by a scaling factor $\sum_{i=1}^{k} a_i\left(\widehat{\theta}_i - \widehat{\theta}\right)^2/(k-1)$.[30,31] In certain cases, this scaling factor can be less than one, which leads to a narrower confidence interval than the standard t distribution approach and can also lead to a narrower interval compared with the Wald-type method in few cases.[32] A variation of this method has been proposed to deal with this by constraining the scaling factor to be ≥ 1.[7] However, throughout this study, the HKSJ method without constraint is used.

3 | SIMULATION STUDY DESIGN

All simulations and analyses were carried out in R version 3.2.2. The package *metafor*[27] was used to run simulated meta-analyses and calculate heterogeneity variance estimates from methods coded in this package, bespoke code was used for those that are not. A study protocol was agreed by all authors before running these simulations and is available upon request from the first author.

3.1 | Simulation methods

For studies $i = 1, ..., k$ in each meta-analysis, true study effects θ_i are simulated from the distribution $N(\theta, \tau^2)$. Parameters θ, τ^2, and k take values as defined in Section 3.2. Study sample sizes N_i are generated from a distribution also detailed in Section 3.2 and are then split evenly between the two study groups n_{1i} and n_{2i}. Participant-level data are then simulated for both continuous and binary outcomes, and effect sizes and within-study variances (θ_i and σ_i^2) are estimated from these data. In continuous outcome meta-analyses, effects are measured as a standardised mean difference, and in binary outcome meta-analyses, effects are measured as a log-odds ratio.

For each study simulated from continuous outcome data, the following steps are carried out:

1. Generate n_{1i} observations from $N\left(0, \sigma_{1i}^2\right)$ and n_{2i} observations from $N\left(\theta_i, \sigma_{2i}^2\right)$. We assume variances σ_{1i}^2 and σ_{2i}^2 in the two groups are equal and, without loss of generality, set them equal to 1.
2. Calculate the sample means and standard deviations of these observations.
3. Calculate $\widehat{\theta}_i$ and $\widehat{\sigma}_i^2$ for standardised mean differences by Hedges g method, thus accounting for small sample bias of standardised mean differences (documented by Borenstein et al,[33] eqs 2.23 and 2.24).

For studies with an odds ratio outcome measure,

1. Generate an average event probability between the two study groups ($\overline{p}_i$) from one of the distributions as defined in Section 3.2. Although this simulation approach is not common, Smith et al[34] has previously defined a Bayesian meta-analysis model that included the same $\overline{p}_i$ parameter.
2. Derive underlying event probabilities for each study group (p_{1i} and p_{2i}) from the solutions to the following simultaneous equations:

$$\overline{p}_i = (p_{1i} + p_{2i})/2$$

$$\theta_i = \log[(p_{2i}(1-p_{1i}))/(p_{1i}(1-p_{2i}))].$$

3. Simulate cell counts of the 2 × 2 contingency table from the distributions $Bin(n_{1i}, p_{1i})$ and $Bin(n_{2i}, p_{2i})$. Apply a continuity correction of 0.5 to studies with zero cell counts.
4. Calculate $\widehat{\theta}_i$ and $\widehat{\sigma}_i^2$ for log odds ratios from the standard formulae in Borenstein et al.[33]

3.2 | Parameter values

Parameter values are chosen to represent the range of values observed in published meta-analyses in the Cochrane Database of Systematic Reviews[10] and based on parameter values from previous simulation studies.[11] For all combinations of parameter values as outlined in this section, 5000 meta-analyses are simulated. Binary outcome meta-analyses are simulated with log-odds ratios of $\theta = \{0, 0.5, 1.1, 2.3\}$ (corresponding to odds ratios of 1, 1.65, 3, and 10). Standardised mean difference meta-analyses are simulated with $\theta = 0.5$ only, because previous simulation studies suggest θ has no noticeable effect on any of the results.[13,17] Sample sizes are generated from the following five distributions to represent meta-analyses containing small, small-to-medium, medium, large, and small and large studies: (1) $N_i = 40$, (2) $N_i \sim U(40,400)$, (3) $N_i = 400$, (4) $N_i \sim U(2000, 4000)$, and (5) $N_i = 40$ (small) in half of studies and half selected from $N_i \sim U(2000, 4000)$ (large). If k is odd in the last scenario, one study is selected randomly (with probability 0.5) to be small or large. For odds ratio meta-analyses, the average event probability ($\overline{p}_i$) takes the values (1) 0.5, (2) 0.05, (3) 0.01, and (4) generated from the distribution $U(0.1,0.5)$. Simulated meta-analyses contain 2, 3, 5, 10, 20, 30, 50, and 100 studies.

Heterogeneity variance parameter values (τ^2) are defined such that the resulting meta-analyses span a wide range of levels of inconsistency between study effects. We measured inconsistency using the I^2 statistic,[32] an approximate measure of the relative size of the heterogeneity

variance to the total variability in effect estimates (the sum of within-study error variance and between-study heterogeneity). The chosen τ^2 values result in meta-analyses with average I^2 values of 0%, 15%, 30%, 45%, 60%, 75%, 90%, and 95% and are given in Appendix S2. I^2 values are calculated using the true τ^2 parameter estimates but still vary between simulated meta-analyses because of the simulated variation in the standard errors. Parameter values for τ^2 vary between scenarios with different distributions for N_i and $\overline{p}_i$ to maintain a consistent range of I^2. In each scenario, τ^2 is fixed, and I^2 varies between meta-analyses; therefore, we also present the range of I^2 next to the graphs in the results.

Simulating all combinations of parameter values leads to 320 scenarios for standardised mean difference meta-analyses ($8(k) \times 5(N_i) \times 8(\tau^2)$) and 5120 scenarios for odds ratio meta-analyses ($8(k) \times 5(N_i) \times 8(\tau^2) \times 4(\overline{p}_i) \times 4(\theta)$). Given the large number of simulated scenarios, this paper can only show results from a representative subset of these scenarios.

3.3 | Performance measures

Properties of heterogeneity variance estimators are measured in terms of bias and mean squared error. These two measures are plotted proportional to the heterogeneity variance parameter value, so that results can be compared more easily between scenarios with different τ^2. For example, a proportional bias of 100% means that $\widehat{\tau}^2$ is on average twice as large as the true τ^2. By the same token, a proportional bias of −50% means that $\widehat{\tau}^2$ is on average half as large as the true τ^2. Similarly, a proportional mean squared error of 100% implies that the mean squared error is equal to τ^2. We also report bias of $\widehat{\theta}$ and coverage of the three included methods to calculate 95% confidence intervals using estimates from the 11 included heterogeneity variance estimators.

4 | RESULTS

In Section 4.1, results are presented for performance measures that relate directly to the heterogeneity variance parameter: bias and mean squared error. In Section 4.2, we present bias of the summary effect. In Section 4.3, we present the coverage probability of the three confidence interval methods for the summary effect.

4.1 | Properties of heterogeneity variance parameter estimates

Estimators are compared in terms of bias in Figures 1 and 2 and in terms of mean squared error in Figures 3 and 4. The first figure in each case shows results from standardised mean difference meta-analyses, and the second shows results from odds ratio meta-analyses. We present selected scenarios containing small studies, small-to-medium studies, and small and large studies combined with scenarios where the average I^2 is either equal to 30% or 90% and for $\theta = 0.5$ only. For odds ratio meta-analyses, we present scenarios where the average event probability in each study is uniformly distributed between 0.1 and 0.5. In this section, results are summarised separately for each heterogeneity variance estimator.

4.1.1 | DerSimonian-Laird

In standardised mean difference meta-analyses, DL is negatively biased when I^2 is large and study sample sizes are small (as shown in Figure 1, bottom left). The estimator is more negatively biased in the equivalent odds ratio meta-analyses, even with event rates between 0.1 and 0.5 (Figure 2). Additionally, DL is negatively biased in odds ratio meta-analyses when sample sizes are small to medium (Figure 2, middle left). In all other scenarios presented in Figures 1 and 2, DL is positively biased in meta-analyses containing fewer than 10 to 20 studies and roughly unbiased for those with more studies. DL has similar bias to many estimators including PM_{CA}, PM_{DL}, and REML in scenarios with small-to-medium studies. In meta-analyses with a mix of small and large studies (Figures 1 and 2, third column), DL is one of the least positively biased estimators —distinctly lower than PM and PM_{CA}.

DerSimonian-Laird has a relatively low mean squared error in the same scenarios where negative bias is also observed (Figures 3 and 4). However, this is not necessarily a good property because only underestimates can be truncated to zero and truncation reduces the error of the estimate. Low mean squared error is also observed in scenarios with small and large studies where DL has low bias (Figures 3 and 4, third column).

4.1.2 | Cochran ANOVA

Cochran ANOVA tends to produce higher estimates of the heterogeneity variance than most other estimators for both standardised mean difference and odds ratio meta-analyses. As such, CA is roughly unbiased in scenarios with high I^2 when most other estimators are negatively biased. However, CA is one of the most positively biased estimators for low-to-moderate I^2. CA's positive bias is particularly prominent in scenarios with small and large studies (Figures 1 and 2, third column); it is counterintuitive to assign equal study weights (as the CA estimator does) in these scenarios with large differences in study size. CA also has higher mean

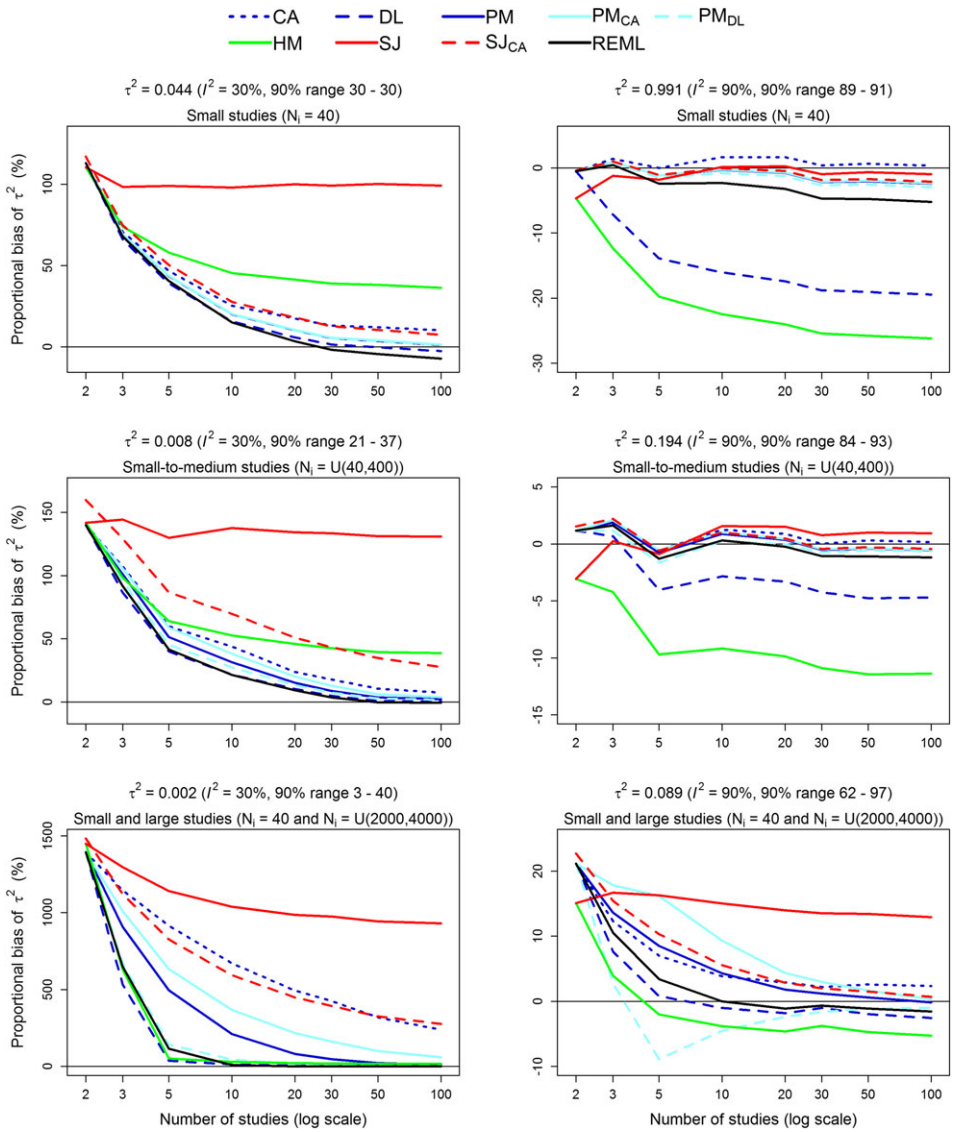


FIGURE 1 Bias of heterogeneity variance estimates in standardised mean difference outcome meta-analyses. Scenarios containing small studies (first row), small-to-medium studies (second row), and small and large studies (third row). Effect size $\theta = 0.5$. Note: the y-axis limits differ between plots. CA, Cochran ANOVA; DL, DerSimonian-Laird; HM, Hartung-Makambi; PM, Paule-Mandel; PM_{CA}, two-step Cochran ANOVA; PM_{DL}, two-step DerSimonian-Laird; REML, restricted maximum likelihood; SJ, Sidik-Jonkman; SJ_{CA}, alternative Sidik-Jonkman [Colour figure can be viewed at wileyonlinelibrary.com]

squared error than most other estimators when the estimator is positively biased (Figures 3 and 4).

4.1.3 | Paule-Mandel

Paule-Mandel has properties similar to DL in scenarios of standardised mean difference meta-analyses that contain small or small-to-medium sized studies (Figure 1, first and second column). In these scenarios, PM is roughly unbiased when I^2 is typically high or the meta-analysis has more than 20 studies and positively biased otherwise. In scenarios where DL is negatively biased, PM often has less negative bias, except in scenarios with highly sparse data where all estimators perform poorly (Figure 2, bottom left). In scenarios with a mix of small and large studies (Figures 1 and 2, third column), PM has a higher mean squared error and higher positive bias than DL, PM_{DL}, Hartung-Makambi (HM), and REML (Figures 1–4, third column).

4.1.4 | Two-step CA (PM_{CA})

Two-step CA uses CA as an initial estimate of heterogeneity. Two-step CA's bias and mean squared error are equal to, or somewhere between, CA and PM in all scenarios. Given that CA and PM have high positive bias and large

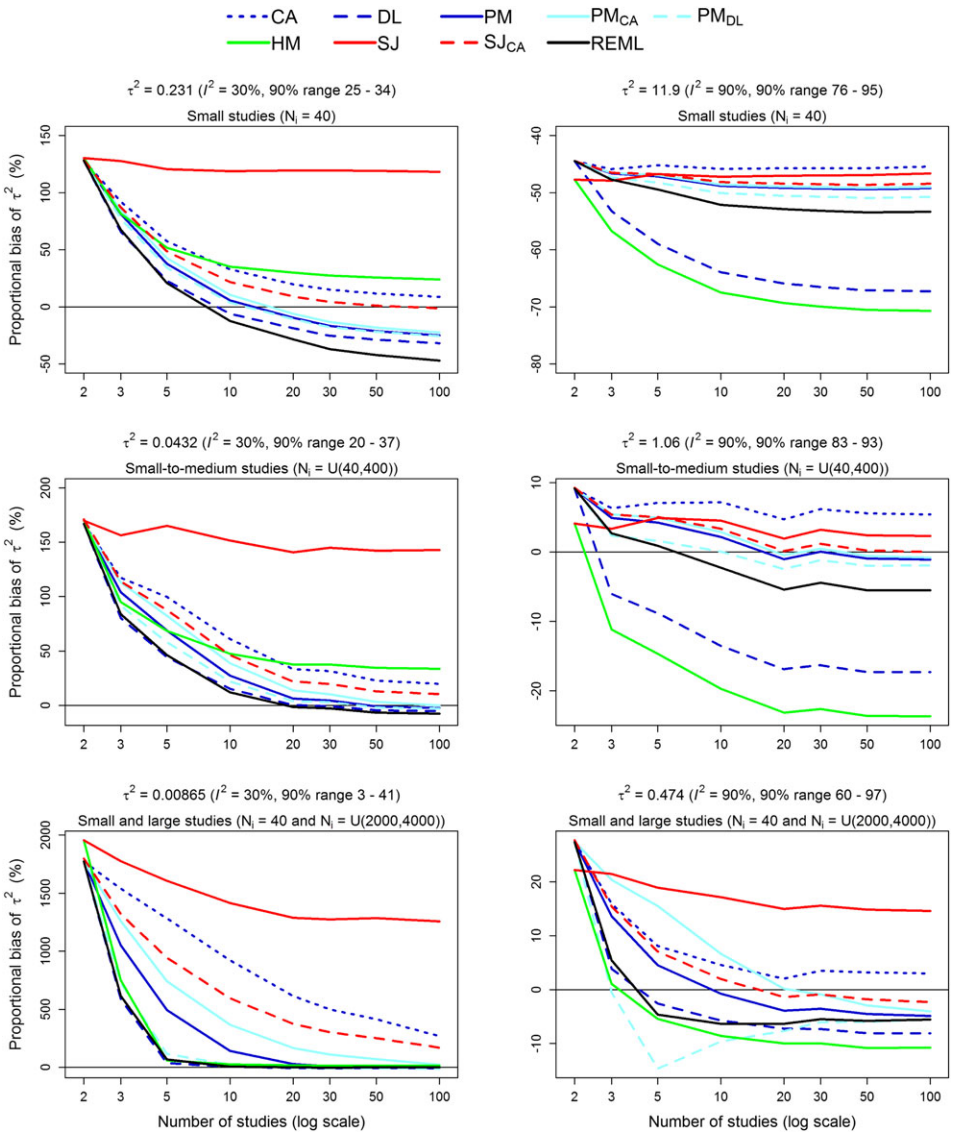


FIGURE 2 Bias of heterogeneity variance estimates in odds ratio meta-analyses with underlying summary odds ratio 1.65 and an average event probability between 0.1 and 0.5. Scenarios containing small studies (first row), small-to-medium studies (second row), and small and large studies (third row). Effect size $\theta = 0.5$. Note: the y-axis limits differ between plots. CA, Cochran ANOVA; DL, DerSimonian-Laird; HM, Hartung-Makambi; PM, Paule-Mandel; PM_{CA}, two-step Cochran ANOVA; PM_{DL}, two-step DerSimonian-Laird; REML, restricted maximum likelihood; SJ, Sidik-Jonkman; SJ_{CA}, alternative Sidik-Jonkman [Colour figure can be viewed at wileyonlinelibrary.com]

mean squared error in scenarios with small and large studies, so too does PM_{CA} (Figures 1–4, third column).

4.1.5 | Two-step DerSimonian-Laird (PM_{DL})

In a similar fashion to PM_{CA}, PM_{DL} has bias and mean squared error that is equal to, or somewhere between, DL and PM in all scenarios. Two-step DL has properties similar to the best performing of the two estimators in all simulated scenarios. In scenarios with large and small studies, PM_{DL} has low positive bias and mean squared error similar to DL, and in scenarios where DL is negatively biased, PM_{DL} and PM have comparable properties. There is little difference in the properties of PM_{DL} and REML in all scenarios.

4.1.6 | Hartung-Makambi

In meta-analyses with small or small-to-medium study sizes and zero or low I^2, HM tends to produce relatively high estimates of heterogeneity and therefore has relatively high positive bias (Figures 1 and 2, top left). This is perhaps because HM is a transformation of the DL estimator that only produces positive estimates. HM tends to produce comparatively low estimates when I^2 is

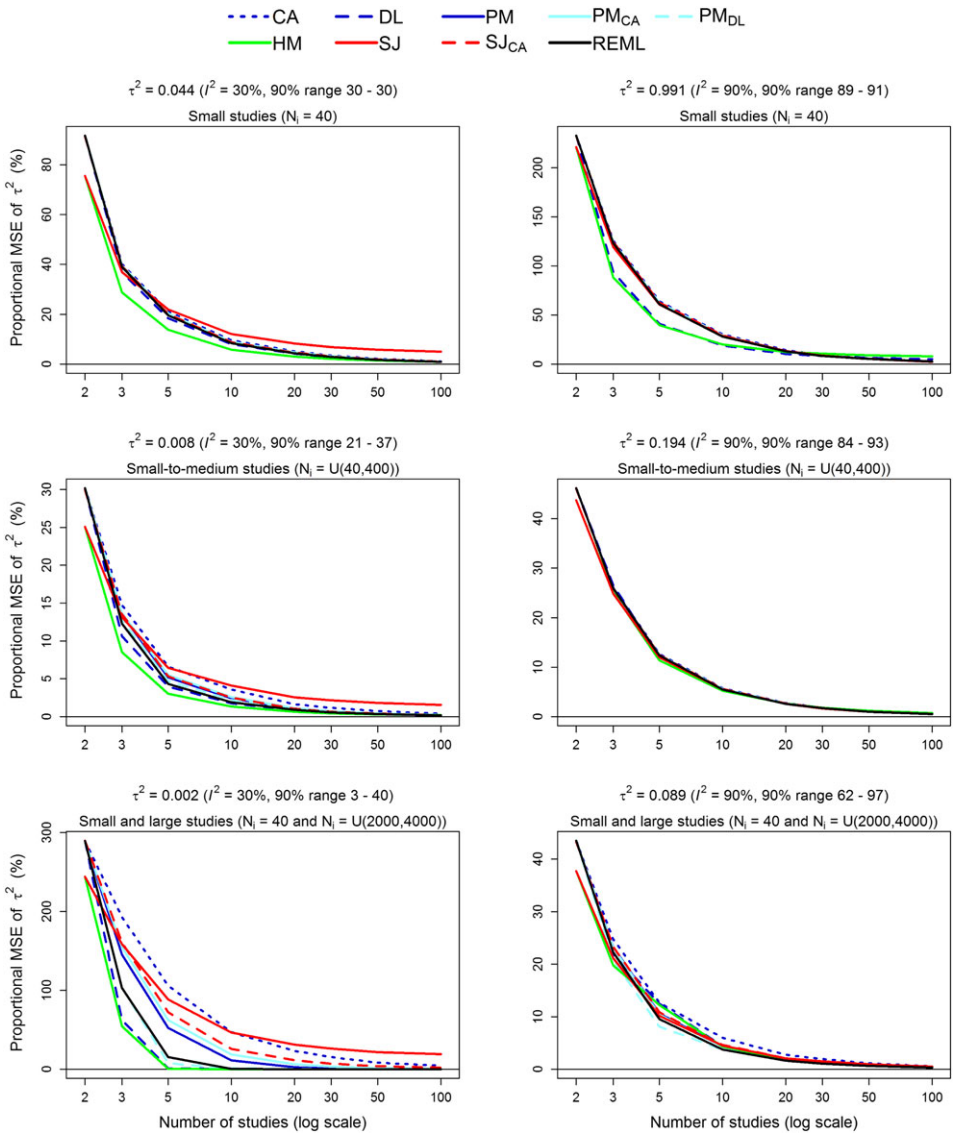


FIGURE 3 Mean squared error of heterogeneity variance estimates in standardised mean difference outcome meta-analyses. Scenarios containing small studies (first row), small-to-medium studies (second row), and small and large studies (third row). Effect size $\theta = 0.5$. Note: the y-axis limits differ between plots. CA, Cochran ANOVA; DL, DerSimonian-Laird; HM, Hartung-Makambi; PM, Paule-Mandel; PM_{CA}, two-step Cochran ANOVA; PM_{DL}, two-step DerSimonian-Laird; REML, restricted maximum likelihood; SJ, Sidik-Jonkman; SJ_{CA}, alternative Sidik-Jonkman [Colour figure can be viewed at wileyonlinelibrary.com]

moderate or high and has more negative bias than DL in these scenarios. HM has a comparatively low mean squared error in all scenarios presented (Figures 3 and 4), including scenarios where HM has high positive bias. Hartung-Makambi is one of the best performing estimators in meta-analyses containing small and large studies (Figures 1–4, third column), with properties comparable with DL, PM_{DL}, and REML.

4.1.7 | Sidik-Jonkman (SJ)

Sidik-Jonkman typically produces one of the highest estimates of the heterogeneity variance in both standardised mean difference and odds ratio meta-analyses, even higher than the other estimators that only produce positive estimates (HM and SJ_{CA}). As such, SJ has considerable positive bias and high mean squared error in meta-analyses with up to moderate I^2. For example, in standardised mean difference meta-analyses containing small-to-medium sized studies and low I^2 (Figure 1, top middle), SJ has bias of more than 100% when almost all other estimators are roughly unbiased.

4.1.8 | Alternative Sidik-Jonkman (SJ_{CA})

SJ_{CA} generally has improved properties over the original SJ estimator. In meta-analyses with small studies (as shown in Figures 1 and 2, first column), SJ_{CA} is one of the least

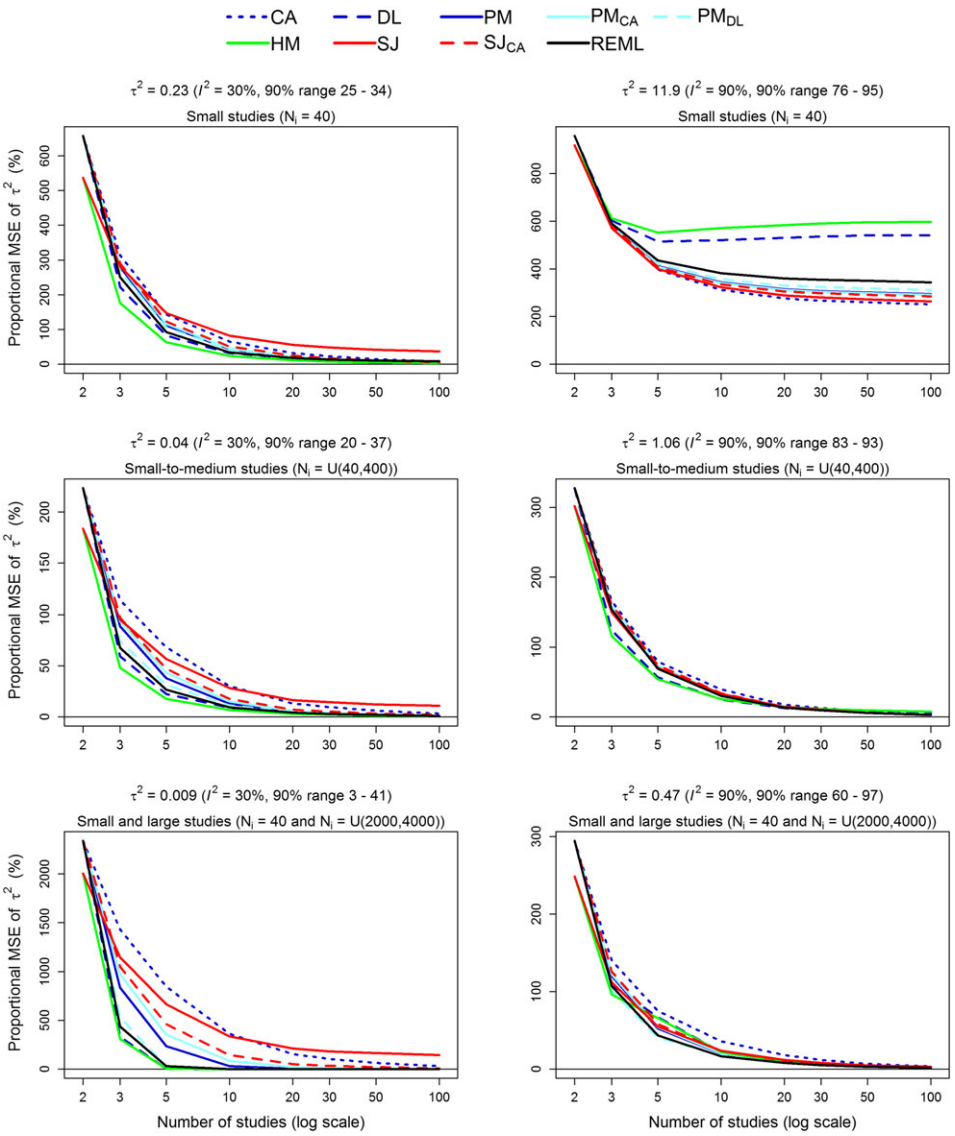


FIGURE 4 Mean squared error of heterogeneity variance estimates in odds ratio meta-analyses with underlying summary odds ratio 1.65 and an average event probability between 0.1 and 0.5. Scenarios containing small studies (first row), small-to-medium studies (second row), and small and large studies (third row). Effect size $\theta = 0.5$. Note: the y-axis limits differ between plots. CA, Cochran ANOVA; DL, DerSimonian-Laird; HM, Hartung-Makambi; PM, Paule-Mandel; PM_{CA}, two-step Cochran ANOVA; PM_{DL}, two-step DerSimonian-Laird; REML, restricted maximum likelihood; SJ, Sidik-Jonkman; SJ_{CA}, alternative Sidik-Jonkman [Colour figure can be viewed at wileyonlinelibrary.com]

biased estimators, with bias similar to many of the truncated methods including DL, PM, and REML. As the typical study size increases, the extent of SJ_{CA}'s positive bias also increases, such that it becomes one of the most positively biased estimators in meta-analyses with small and large studies (Figures 1 and 2, third column). In scenarios where SJ_{CA} has positive bias, it also has relatively high mean squared error (i.e., in meta-analyses with large studies, see Figures 3 and 4, third column).

4.1.9 | Restricted maximum likelihood

REML has similar properties to PM_{DL} and DL in most scenarios. In a small number of scenarios where DL is negatively biased, REML is also negatively biased but often to a much lesser extent (observed most prominently in Figure 2, bottom left). REML has relatively low bias and low mean squared error comparable with DL, HM, and PM_{DL} in scenarios containing small and large studies.

4.2 | Summary effect estimates

Results show that summary effect estimates $(\widehat{\theta})$ are almost unbiased in all scenarios of standardised mean difference meta-analyses ($\theta = 0.5$) and odds ratio meta-analyses with common events. However, summary effect estimates

are biased towards the null value of zero in odds ratio meta-analyses with rare events. This is likely to be partly a consequence of the choice of continuity correction (we added 0.5 to zero cell counts), and the degree of bias was similar across all heterogeneity variance estimators. We present bias of the summary effect in Supporting Information only.

4.3 | Coverage of 95% summary effect confidence intervals

Coverage is presented in Figure 5 for all combinations of heterogeneity variance estimators and (95%) Wald-type, *t* distribution, and HKSJ confidence interval methods for the summary effect. Results are presented for standardised mean difference meta-analyses only, but results are consistent with the equivalent scenarios of odds ratio meta-analyses with common events (event probabilities 0.1 to 0.5, see Appendix S3).

4.3.1 | Wald-type method

Coverage of the 95% Wald-type confidence interval can differ by up to 5% between heterogeneity variance estimators, up to 30% between numbers of studies, and up to 20% between heterogeneity values. Coverage varies between 96% and 100% when studies are homogeneous and can be as low as 65% when the typical I^2 is 90%

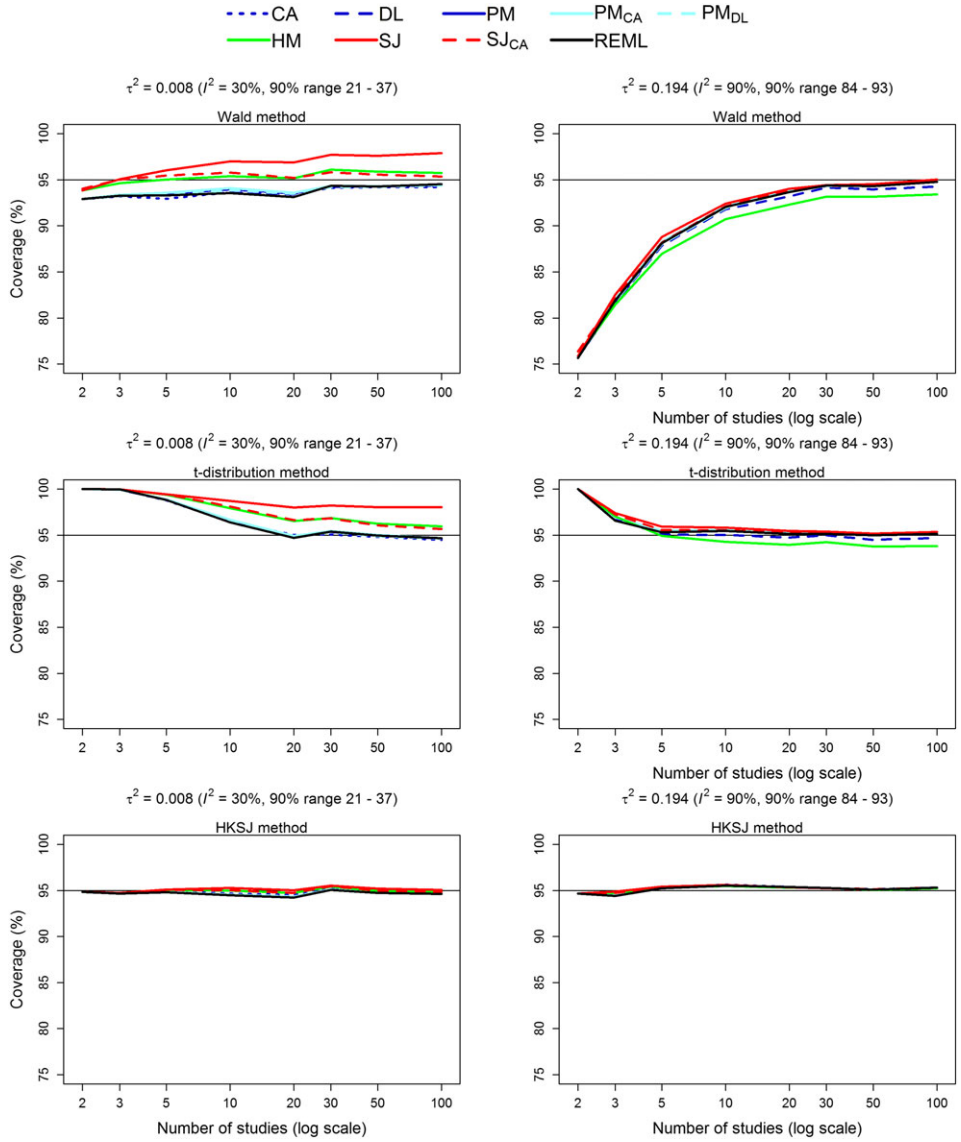


FIGURE 5 Coverage of 95% confidence intervals of the summary effect in standardised mean difference meta-analyses with small-to-medium studies (N_i = U(40,400)). Coverage of Wald-type (first row), *t* distribution (second row), and HKSJ (third row) confidence intervals presented. CA, Cochran ANOVA; DL, DerSimonian-Laird; HM, Hartung-Makambi; PM, Paule-Mandel; PM_{CA}, two-step Cochran ANOVA; PM_{DL}, two-step DerSimonian-Laird; REML, restricted maximum likelihood; SJ, Sidik-Jonkman; SJ_{CA}, alternative Sidik-Jonkman [Colour figure can be viewed at wileyonlinelibrary.com]

($\tau^2 = 0.187$) and meta-analyses have two or three studies. When heterogeneity is present, the confidence interval's coverage tends towards the nominal value of 95% as the number of studies increases.

4.3.2 | Standard *t* distribution method

Coverage of the *t* distribution 95% confidence interval is generally more robust to changes in the mean I^2, as shown in Figure 5. In these scenarios, however, coverage can differ by up to 5% depending on the heterogeneity variance estimator used and the number of studies. When there are 20 studies or more, 95% *t* distribution confidence intervals have a coverage of 94% to 97% but perform conservatively with coverages close to 100% when there are fewer than 20 studies. The heterogeneity variance estimator that works best with this confidence interval method varies considerably between scenarios, so it is difficult to select one overall.

4.3.3 | HKSJ method

The HKSJ confidence interval for the summary effect has better coverage than the other two methods in all scenarios. This method has a coverage of 94% to 96% in standardised mean difference meta-analyses presented in Figure 5 and is insensitive to the choice of heterogeneity variance estimator. The method can produce confidence intervals with suboptimal coverage in odds ratio meta-analyses with rare events, where all meta-analysis methods perform poorly (as demonstrated in Appendix S4).

4.4 | Generalisability of presented results

The results presented so far come from a subset of all simulation scenarios, but these results can be generalised to some extent. All results are presented in Supporting Information.

First, all results presented in the main paper come from scenarios with standardised mean difference and log-odds ratio summary effects of 0.5 (odds ratio = 1.65), but results were consistent with more extreme odds ratio effects in most scenarios. The exception is in odds ratio meta-analyses containing only small studies with rare events (average event probability = 0.05), where a larger effect size (odds ratio = 10) produced heterogeneity variance estimates with more negative bias across all methods. Results from other effect sizes are found in Supporting Information.

Second, results are not presented in the main paper from scenarios where all heterogeneity variance methods failed with considerable negative bias. This occurred in all scenarios of odds ratio meta-analyses with rare events (event probability = 0.05 and 0.01) except where study sizes were large (sample size > 4000 per study). In these scenarios, summary effects were considerably biased, and confidence interval methods also failed to produce reasonable coverage. For example, simulation results show that the HKSJ method can have coverage as low as 85% in odds ratio meta-analyses with small-to-medium sized studies with an underlying event probability of 0.05 (see Appendix S4). Poor properties were perhaps observed in these scenarios because many studies contained zero events and a continuity correction was applied (0.5 was added to all 2×2 cell counts in these simulations). An alternative continuity correction may have produced different results.

Finally, results presented thus far are from meta-analyses with typical I^2 values of 0%, 30%, 60%, and 90% (corresponding to four heterogeneity variance parameter values). Meta-analyses with other typical I^2 values were simulated, but the four presented gave an adequate description of the properties of methods across all levels of inconsistency.

5 | DISCUSSION

The DL heterogeneity variance estimator is not recommended for widespread use in two-stage random-effects meta-analysis and, therefore, should not be the default method for meta-analysis in statistical software packages; it produces estimates with more negative bias than most other methods in odds ratio meta-analyses with small studies or rare events and to a lesser extent in standardised mean difference meta-analyses with small studies. This finding can perhaps be explained by DL's fixed-effect study weights that are based solely on estimated within-study variances; these variances are imprecise and likely to be biased under such conditions. This observation is in agreement with previous simulation studies,[4,12] as identified in a systematic review.[11] Viechtbauer[13] and Böhning et al[35] stated that DL is unbiased when within-study variances are known. However, DL is one of the better performing estimators in meta-analyses with large differences in study size.

This simulation study identified three heterogeneity variance estimators with more reasonable properties: REML,[9] PM,[6] and the two-step PM that uses a DL initial estimate.[23] Paule-Mandel is often approximately unbiased when DL is negatively biased. However, results also show PM has high positive bias when there are large differences in study size. This can perhaps be attributed to the random-effects study weights used in this method,

which can lead to small studies being given a relatively large weight under heterogeneous conditions. A similar issue regarding the use of random-effects study weights for summary effect estimation has been noted elsewhere.[36] The two-step DL estimator (PM_{DL}) inherits most of the best properties of DL and PM methods and is simple to compute. Restricted maximum likelihood has very similar properties to this two-step estimator and is already widely known, recommended in two previous simulation studies for meta-analyses with continuous[3,13] and binary[13] outcomes. Furthermore, REML is already available in many statistical software packages.[27,37] Of those with reasonable properties, REML is the only estimator that assumes normality of effect sizes, but a previous simulation study[38,39] showed all these methods are reasonably robust under non-normal conditions.

One of the aims of this simulation study was to investigate when it is appropriate to rely on one estimate of the heterogeneity variance. Results show all estimators are imprecise and often fail to detect high levels of heterogeneity in meta-analyses containing fewer than 10 studies. Furthermore, only 14% of meta-analyses in the Cochrane Database of Systematic Reviews contain 10 studies or more,[10] so it is rarely appropriate to rely on one estimate of heterogeneity in this setting. All estimators have poor properties even in meta-analyses containing high numbers of studies when study sizes are small or the event of interest is rare.

Estimates of the summary effect and its HKSJ confidence interval are of less cause for concern and perform well even in meta-analyses with only two studies. In particular, the HKSJ confidence interval offers a large improvement in coverage over the commonly used Wald-type confidence interval. However, caution must still be applied when dealing with meta-analysis datasets with rare events, where summary effects are biased and the HKSJ confidence interval method can have coverage as low as 85%. Summary effect estimates in this study were calculated using the inverse-variance approach, although the use of the Mantel-Haenszel method has been recommended for rare events[18,40] and may have improved properties in these scenarios. These findings agree with a previous simulation study,[41] in which the HKSJ method was compared with other confidence interval methods for both continuous and binary outcome measures. The results presented in this paper show the HKSJ method is robust to changes in the heterogeneity variance estimate.

Our findings do not concur with some previous simulation studies. In all cases, this can be attributed to differences in parameter values and other differences in simulation study design. The original estimator proposed by Sidik and Jonkman[8] performed well in the author's own simulations, yet simulations in this study show they have considerable positive bias in meta-analyses of up to moderate I^2. This was not observed by Sidik and Jonkman[8] because simulated meta-analyses were only presented with high I^2.[11] The method of PM has been recommended based on the results of three previous simulation studies,[3,12,15] but these studies did not simulate meta-analyses with moderate-to-large differences in study size, where PM has considerable positive bias. Novianti et al[3] only recommended REML for continuous outcome meta-analyses and observed a small negative bias when the outcome is binary and high I^2; this bias was less pronounced in our simulations with low-to-moderate I^2 that Novianti et al[3] did not include in their simulations.[11]

The limitations of this simulation study are as follows. First, only a subset of all confidence interval methods for the summary effect are included. Results show the HKSJ method is more robust than the Wald method to the choice of heterogeneity variance estimator, but no confidence interval method can be recommended solely from the results of this study. Other methods include the profile likelihood method,[21] which has also been shown as a better alternative to the Wald method in simulated meta-analysis data[42] and recommended elsewhere.[43] Second, a continuity correction of 0.5 was applied whenever simulated studies with a binary outcome contained zero events, but other methods with a better performance are available.[44] This choice may have affected the results in scenarios where the event is rare (ie, 0.05), but alternative continuity corrections are unlikely to have led to meaningful improvements where the event rate is extremely rare (ie, 0.01) and all random-effects methods fail in terms of all performance measures. We assumed effects to be normally distributed, and although this is a limitation, it has been shown that most of the investigated methods are robust even in extreme non-normal distributions.[38] Third, our analyses assume that all studies provide unbiased estimates of the true effects underlying them. In practice, results of studies may be biased if the studies are performed suboptimally, and meta-analyses may be biased if studies are missing for reasons related to their results (eg, due to publication bias). These biases can affect estimation of heterogeneity (both upwardly or downwardly) and lead to inappropriate conclusions. Finally, although the study aimed to simulate a comprehensive range of scenarios, this range could never be complete given how diverse meta-analyses are in practice; not all outcome measures were included (eg, hazard ratios), and the distributions from which sample sizes were drawn in this study cannot be considered representative of all observed distributions because study sample sizes are unlikely to conform to a defined distribution.

TABLE 2 A summary of results and recommendations (considering only REML, PM and PM_{DL} heterogeneity variance methods, and HKSJ confidence interval)

		OR outcome with average event probability:		Standardised Mean Difference outcome
		0.05	0.1 to 0.5	
Study sizes:	**Small**	All estimators have substantial negative bias in the presence of heterogeneity. HKSJ confidence interval can have coverage too high/low for >20 studies (Appendix S4).	REML/PM/PM_{DL} recommended, but all estimators biased/imprecise for <10 studies. HKSJ confidence interval yields the nominal coverage.	
	Small-to-medium			
	Small and large		REML/PM_{DL} and HKSJ confidence interval recommended (as above), but all heterogeneity variance estimators biased/imprecise for <10 studies. PM positively biased.	

Abbreviations: HKSJ, Hartung-Knapp-Sidik-Jonkman; PM, Paule-Mandel; PM_{DL}, two-step DerSimonian-Laird; REML, restricted maximum likelihood.

We compared methods in the context of a classical two-stage meta-analysis where study effect estimates and their standard errors are calculated first and then combined at the second final stage. Alternatively, one-stage meta-analyses can be undertaken using individual participant data using mixed modelling techniques; these raw data can be derived trivially from study-level 2×2 contingency tables for binary outcome meta-analyses.[45,46] Stijnen et al[45] explain that this approach makes random-effects meta-analyses more feasible with sparse data and does not require a continuity correction in case of zero events. Jackson et al[47] reviewed modelling approaches for this type of meta-analysis data and suggest these models can offer improved statistical inference on the summary effect. However, these models can present additional numerical issues given their complexity. Future work comparing the properties of heterogeneity variance methods between one-stage and two-stage binary outcome meta-analyses would be informative.

The HKSJ method is generally preferred over the Wald-type method. However, Wiksten et al[31] showed it can occasionally lead to less conservative results, even when the Wald method uses a fixed-effect variance structure. Sidik and Jonkman[4] proposed a modification to the HKSJ method to ensure the resulting confidence interval is at least as wide as the Wald-type fixed-effect confidence interval. We did not apply this modification in our study. A simulation study by Rover et al[48] found the modified method provides coverage closer to the nominal level when differences in study size were large.

Summarising the properties of a comprehensive list of heterogeneity variance estimators, compared over many combinations of parameter values, was the biggest challenge of this study. By simulating meta-analyses from a wide range parameter values, inevitably, there are scenarios that reflect meta-analyses rarely observed in practice. For example, most meta-analyses contain very few studies,[10,49] but meta-analyses with up to 100 studies were simulated in order to show results over the full range of possible meta-analysis sizes. An attempt was made to focus more on the scenarios representative of real meta-analyses when interpreting results, but this was inevitably subjective.

6 | CONCLUSION

A summary of our recommendations are given in Table 2. The two-step DL estimator (PM_{DL}) and REML can often be biased but overall have the most reasonable properties in standardised mean difference and odds ratio meta-analyses. Of these two estimators, REML is recommended on the basis of these results because it is already widely known, available in most statistical software packages, and consistent with the method commonly used for one-stage meta-analyses using individual participant data.[50] The two-step DL estimator is recommended as an alternative if a simpler, noniterative method is required.

The HKSJ confidence interval for the summary effect is generally recommended over the standard t distribution and Wald-type methods, particularly in binary outcome meta-analyses with rare events and the number of studies included is less than 20. To be consistent, we recommend the same REML estimate of the heterogeneity variance to calculate this confidence interval. However,

this is inconsequential given how robust this confidence interval is to changes in the heterogeneity variance method in most scenarios.

A REML point estimate, or indeed any other single estimate of heterogeneity, should not be relied on to gauge the extent of heterogeneity in most meta-analyses. Confidence intervals should always be reported to express imprecision of the heterogeneity variance estimate. However, a point estimate can usually be used reliably to calculate a summary effect with an HKSJ confidence interval.

ORCID

Dean Langan http://orcid.org/0000-0002-5414-8882
Dan Jackson http://orcid.org/0000-0002-4963-8123
Areti Angeliki Veroniki http://orcid.org/0000-0001-6388-4825
Evangelos Kontopantelis http://orcid.org/0000-0001-6450-5815

REFERENCES

1. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials*. 1986;7(3):177-188.
2. Malzahn U, Böhning D, Holling H. Nonparametric estimation of heterogeneity variance for the standardised difference used in meta-analysis. *Biometrika*. 2000;87(3):619-632. https://doi.org/10.1093/biomet/87.3.619
3. Novianti PW, Roes KCB, van der Tweel I. Estimation of between-trial variance in sequential meta-analyses: a simulation study. *Contemp Clin Trials*. 2014;37(1):129-138. https://doi.org/10.1016/j.cct.2013.11.012
4. Sidik K, Jonkman JN. A comparison of heterogeneity variance estimators in combining results of studies. *Stat Med*. 2007;26(9):1964-1981. https://doi.org/10.1002/sim.2688
5. Veroniki AA, Jackson D, Viechtbauer W, et al. Methods to estimate heterogeneity variance and its uncertainty in meta-analysis. *Res Syn Meth*. 2015;7(1):55-79. https://doi.org/10.1002/jrsm.1164
6. Paule RC, Mandel J. Consensus values and weighting factors. *J Res Natl Bur Stand*. 1982;87(5):377-385.
7. Hartung J, Makambi KH. Reducing the number of unjustified significant results in meta-analysis. *Commun Stat Simul Comput*. 2003;32(4):1179-1190. https://doi.org/10.1081/SAC-120023884
8. Sidik K, Jonkman JN. Simple heterogeneity variance estimation for meta-analysis. *J R Stat Soc Ser B Appl Stat*. 2005;54(2):367-384. https://doi.org/10.1111/j.1467-9876.2005.00489.x
9. Harville DA. Maximum likelihood approaches to variance component estimation and to related problems. *J Am Stat Assoc*. 1977;72(358):320-338. https://doi.org/10.2307/2286796
10. Langan D, Higgins JPT, Simmonds M. An empirical comparison of heterogeneity variance estimators in 12,894 meta-analyses. *Res Syn Meth*. 2015;6(2):195-205. https://doi.org/10.1002/jrsm.1140
11. Langan D, Higgins JPT, Simmonds M. Comparative performance of heterogeneity variance estimators in meta-analysis: a review of simulation studies. *Res Syn Meth*. 2016;8(2):181-198. https://doi.org/10.1002/jrsm.1198
12. Panityakul T, Bumrungsup C, Knapp G. On estimating residual heterogeneity in random-effects meta-regression: a comparative study. *J Stat Theory Appl*. 2013;12(3):253-265.
13. Viechtbauer W. Bias and efficiency of meta-analytic variance estimators in the random-effects model. *J Educ Behav Stat*. 2005;30(3):261-293. https://doi.org/10.3102/10769986030003261
14. Rukhin AL, Biggerstaff BJ, Vangel MG. Restricted maximum likelihood estimation of a common mean and the Mandel-Paule algorithm. *J Stat Plan Inference*. 2000;83(2):319-330. https://doi.org/10.1016/S0378-3758(99)00098-1
15. Bhaumik DK, Amatya A, Normand SLT, et al. Meta-analysis of rare binary adverse event data. *J Am Stat Assoc*. 2012;107(498):555-567.
16. Knapp G, Hartung J. Improved tests for a random effects meta-regression with a single covariate. *Stat Med*. 2003;22(17):2693-2710.
17. Sanchez-Meca J, Marín-Martínez F. Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychol Methods*. 2008;13(1):31-48.
18. Kontopantelis E, Springate DA, Reeves D. A re-analysis of the Cochrane library data: the dangers of unobserved heterogeneity in meta-analyses. *PLoS One*. 2013;8(7):e69930.
19. Chung Y, Rabe-Hesketh S, Choi IH. Avoiding zero between-study variance estimates in random-effects meta-analysis. *Stat Med*. 2013;32(23):4071-4089.
20. Hunter J, Schmidt F. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. Thousand Oaks, Calif.: SAGE Publications; 2004.
21. Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. *Stat Med*. 1996;15(6):619-629.
22. Morris CN. Parametric empirical Bayes inference: theory and applications. *J Am Stat Assoc*. 1983;78(381):47-55.
23. DerSimonian R, Kacker R. Random-effects model for meta-analysis of clinical trials: an update. *Contemp Clin Trials*. 2007;28(2):105-114. https://doi.org/10.1016/j.cct.2006.04.004
24. Bowden J, Tierney JF, Copas AJ, Burdett S. Quantifying, displaying and accounting for heterogeneity in the meta-analysis of RCTs using standard and generalised Q statistics. *BMC Med Res Methodol*. 2011;11(1):41.
25. Jackson D, Turner R, Rhodes K, Viechtbauer W. Methods for calculating confidence and credible intervals for the residual between-study variance in random effects meta-regression models. *BMC Med Res Methodol*. 2014;14(1):103.
26. Rukhin AL. Estimating heterogeneity variance in meta-analysis. *J R Stat Soc Series B Stat Methodology*. 2013;75(3):451-469.
27. Viechtbauer W. Conducting meta-analyses in R with the metaphor package. *J Stat Softw*. 2010;36(3):1-48.

28. Follmann DA, Proschan MA. Valid inference in random-effects meta-analysis. *Biometrics*. 1999;55(3):732-737.

29. Hartung J, Knapp G. A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Stat Med*. 2001;20(24):3875-3889.

30. Sidik K, Jonkman JN. A simple confidence interval for meta-analysis. *Stat Med*. 2002;21(21):3153-3159. https://doi.org/10.1002/sim.1262

31. Wiksten A, Rücker G, Schwarzer G. Hartung-Knapp method is not always conservative compared with fixed-effect meta-analysis. *Stat Med*. 2016;35(15):2503-2515. https://doi.org/10.1002/sim.6879

32. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med*. 2002;21(11):1539-1558. https://doi.org/10.1002/sim.1186

33. Borenstein M, Hedges LV, Higgins JPT. *Introduction to Meta-Analysis*. Hoboken, NJ, USA: Wiley; 1999.

34. Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: a comparative study. *Stat Med*. 1995;14(24):2685-2699.

35. Böhning D, Malzahn U, Dietz E, Schlattmann P, Viwatwongkasem C, Biggeri A. Some general points in estimating heterogeneity variance with the DerSimonian-Laird estimator. *Biostatistics*. 2002;3(4):445-457. https://doi.org/10.1093/biostatistics/3.4.445

36. Higgins JPT, Spiegelhalter DJ. Being sceptical about meta-analyses: a Bayesian perspective on magnesium trials in myocardial infarction. *Int J Epidemiol*. 2002;31(1):96-104. https://doi.org/10.1093/ije/31.1.96

37. Kontopantelis E, Reeves D. metaan: random-effects meta-analysis. *Stata J*. 2010;10(3):395.

38. Kontopantelis E, Reeves D. Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: a simulation study. *Stat Methods Med Res*. 2012;21(4):409-426.

39. Kontopantelis E, Reeves D. Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: a comparison between DerSimonian-Laird and restricted maximum likelihood. *Stat Methods Med Res*. 2012;21(6):657-659.

40. Bradburn MJ, Deeks JJ, Berlin JA, Russell Localio A. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Stat Med*. 2007;26(1):53-77.

41. IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Med Res Methodol*. 2014;14(1):25.

42. Henmi M, Copas JB. Confidence intervals for random effects meta-analysis and robustness to publication bias. *Stat Med*. 2010;29(29):2969-2983. https://doi.org/10.1002/sim.4029

43. Cornell JE. Random-effects meta-analysis of inconsistent effects: a time for change. *Ann Intern Med*. 2014;160(4):267-270. https://doi.org/10.7326/M13-2886

44. Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Stat Med*. 2004;23(9):1351-1375. https://doi.org/10.1002/sim.1761

45. Stijnen T, Hamza TH, Ozdemir P. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. *Stat Med*. 2010;29(29):3046-3067.

46. Simmonds M, Higgins JPT. A general framework for the use of logistic regression models in meta-analysis. *Stat Methods Med Res*. 2016;25(6):2858-2877.

47. Jackson D, Law M, Stijnen T, Viechtbauer W, White IR. A comparison of 7 random-effects models for meta-analyses that estimate the summary odds ratio. *Stat Med*. 2018 (in press). doi: https://doi.org/10.1002/sim.7588;37(7):1059-1085.

48. Röver C, Knapp G, Friede T. Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. *BMC Med Res Methodol*. 2015;15(1):99.

49. Davey J, Turner RM, Clarke MJ, Higgins JPT. Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: a cross-sectional, descriptive analysis. *BMC Med Res Methodol*. 2011;11(1):160. https://doi.org/10.1186/1471-2288-11-160

50. Simmonds M, Stewart G, Stewart L. A decade of individual participant data meta-analyses: a review of current practice. *Contemp Clin Trials*. 2015;45:76-83.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Langan D, Higgins JPT, Jackson D, et al. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Res Syn Meth*. 2019;10:83–98. https://doi.org/10.1002/jrsm.1316

Appendix 1: Proportional bias (left-hand-side) and proportional mean squared error (right-hand-side) in selected scenarios with estimators proposed by Rukhin (B0, BP) and Malzahn, Böhning and Holling (MBH) included

Scenarios containing standardised mean difference meta-analyses ($\theta = 0.5$) with small-to-medium study sizes ($N_i = 40 - 400$) and an average I^2 of 60%.

See separate file for figure.

Appendix 2: Heterogeneity variance parameter values for each simulated scenario.

Study sizes	Avg. event probability	$I^2 = 15\%$	$I^2 = 30\%$	$I^2 = 45\%$	$I^2 = 60\%$	$I^2 = 75\%$	$I^2 = 90\%$	$I^2 = 95\%$
odds ratio meta-analyses ($\theta = 0.5$)								
small	0.5	0.0670	0.1780	0.3440	0.6330	1.330	4.500	15.60
small-to-medium		0.0144	0.0333	0.0655	0.1220	0.2440	0.7800	1.670
medium		0.0067	0.0174	0.0333	0.0560	0.1220	0.3670	0.7800
small and large		0.0025	0.0066	0.0144	0.0230	0.0756	0.3560	0.7800
large		0.0001	0.0023	0.0046	0.0082	0.0166	0.0450	0.0100
small	0.1 to 0.5	0.0944	0.2330	0.4450	0.8560	1.89	20.00	*
small-to-medium		0.0178	0.0433	0.0855	0.1545	0.3220	1.110	2.300
medium		0.0089	0.0233	0.0433	0.0780	0.1560	0.4500	1.110
small and large		0.0036	0.0084	0.0178	0.0356	0.0945	0.4560	1.220
large		0.0012	0.0023	0.0058	0.0107	0.0222	0.0645	0.1340
small	0.05	0.4220	1.156	2.560	7.560	*	*	*
small-to-medium		0.0755	0.1890	0.3780	0.7450	1.780	*	*
medium		0.0340	0.0967	0.1890	0.3560	0.7560	3.440	*
small and large		0.0144	0.0345	0.0745	0.1670	0.4330	2.300	*
large		0.0053	0.0133	0.0255	0.0445	0.0890	0.2300	0.5600
small	0.01	2.780	14.50	*	*	*	*	*
small-to-medium		0.3780	1.110	2.450	6.700	*	*	*
medium		0.1200	0.4500	1.067	2.440	7.800	*	*
small and large		0.0656	0.1780	0.3400	0.1000	3.670	*	*
large		0.0245	0.0622	0.1220	0.2330	0.4780	1.780	*
standardised mean difference meta-analyses ($\theta = 0.5$)								
small	-	0.0178	0.0444	0.0845	0.156	0.322	0.1	2.440
small-to-medium	-	0.00345	0.00856	0.0156	0.023	0.056	0.12	0.3400
medium	-	0.00178	0.00444	0.00844	0.01545	0.0311	0.089	0.1200
small and large	-	0.000656	0.00156	0.00344	0.00744	0.0189	0.089	0.1200
large	-	0.000244	0.00056	0.001133	0.00211	0.00422	0.0133	0.0256

τ^2 consistent between numbers of studies and distributions of study effects. $I^2 = 0\%$ always corresponds to $\tau^2 = 0$ so these scenarios are not included in the table.

** the given average I^2 could not be attained for any τ^2 value, so meta-analyses were not simulated.*

Appendix 3: Coverage of 95% confidence intervals of the summary effect in odds ratio meta-analyses with small-to-medium studies ($N_i = U(40, 400)$) and an average event probability between 0.1 and 0.5

Coverage of Wald-type (first row), t-distribution (second row), and HKSJ (third row) confidence intervals presented.

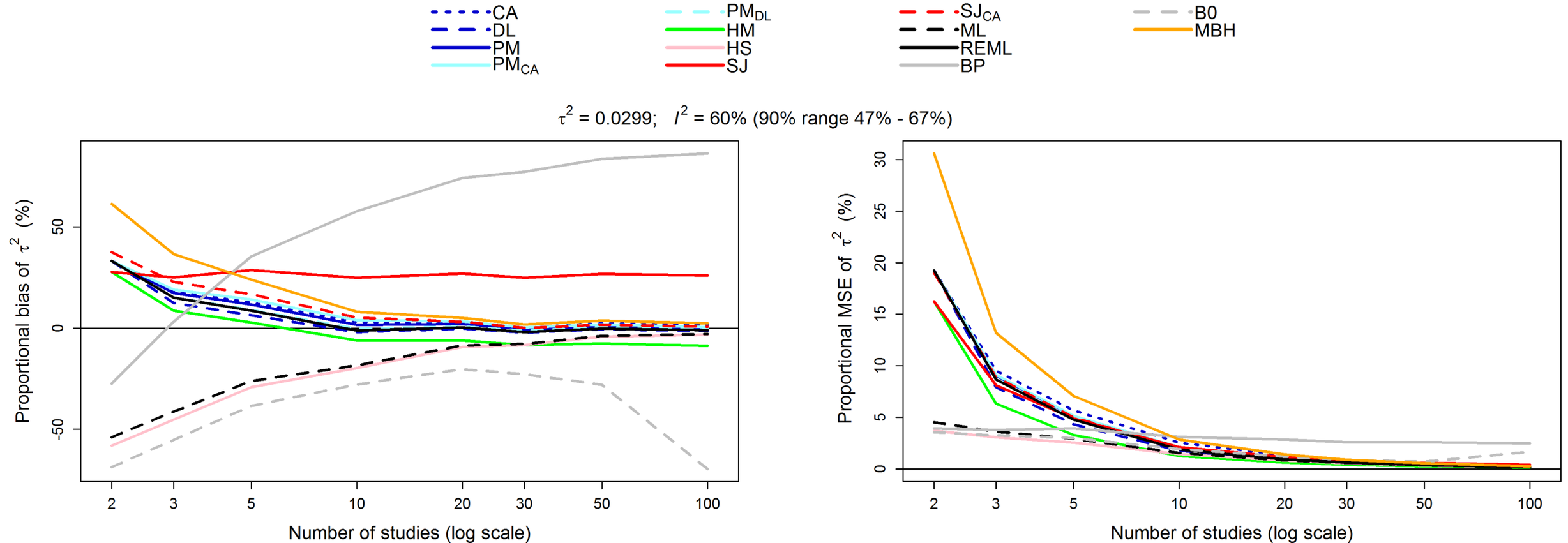
See separate file for figure.

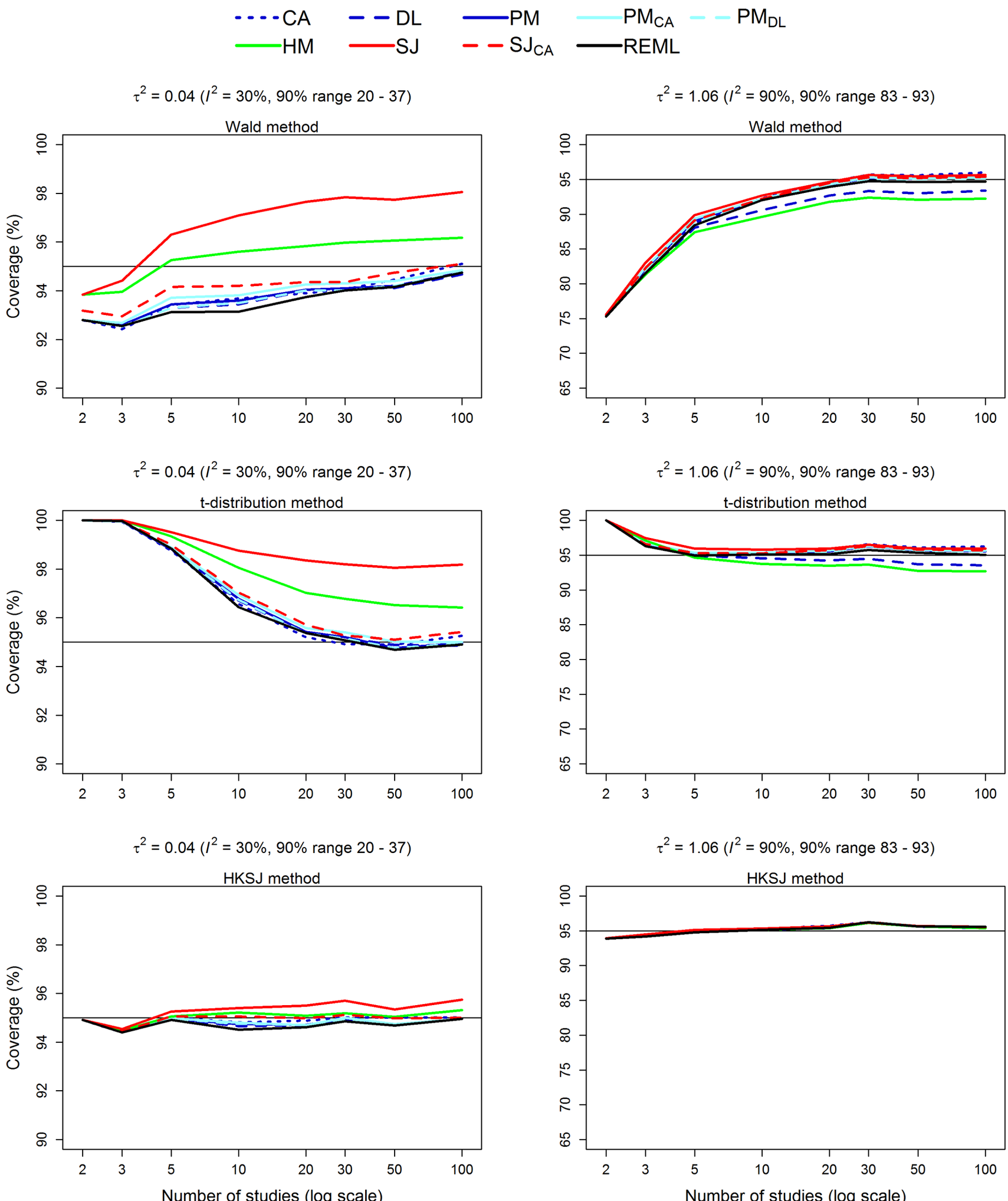
Appendix 4: Coverage of 95% confidence intervals of the summary effect in odds ratio meta-analyses with small-to-medium studies ($N_i = 40 - 400$) and an average event probability of 0.05.

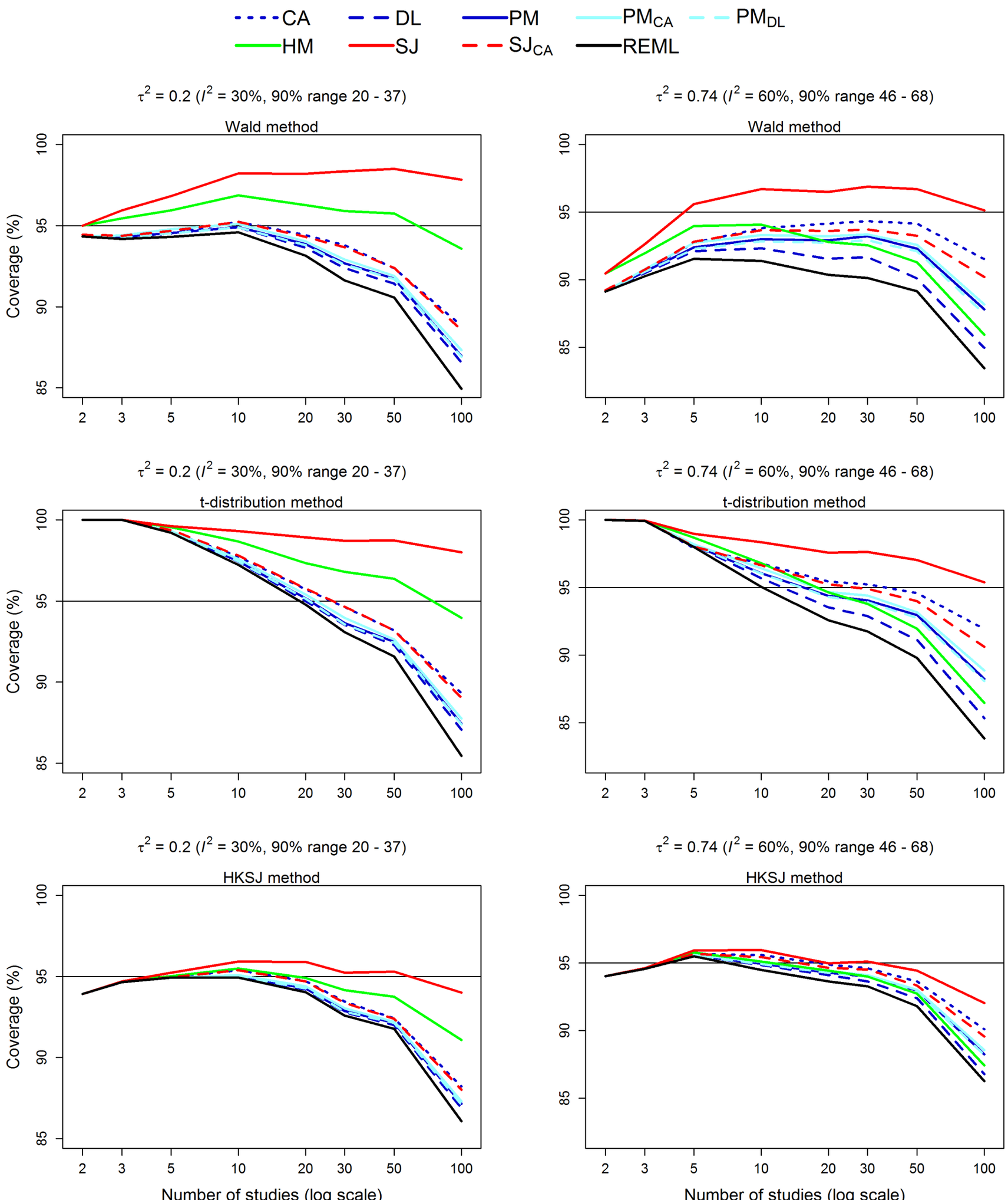
Coverage of Wald-type (first row), t-distribution (second row) and Hartung-Knapp (third row) confidence intervals presented.

There was no such τ^2 that produced a mean I^2 of 90% so scenarios where $I^2 = 60\%$ are presented instead. Effect size $\theta = 0.5$.

See separate file for figure.


CA
DL
PM
PM_CA
PM_DL
HM
HS
SJ
SJ_CA
ML
REML
BP
B0
MBH
τ² = 0.0299; I² = 60% (90% range 47% - 67%)
Proportional bias of τ² (%)
Proportional MSE of τ² (%)
Number of studies (log scale)
Number of studies (log scale)


CA
DL
PM
PM$_{CA}$
PM$_{DL}$
HM
SJ
SJ$_{CA}$
REML
τ^2 = 0.04 (I^2 = 30%, 90% range 20 - 37)
Wald method
τ^2 = 1.06 (I^2 = 90%, 90% range 83 - 93)
Wald method
τ^2 = 0.04 (I^2 = 30%, 90% range 20 - 37)
t-distribution method
τ^2 = 1.06 (I^2 = 90%, 90% range 83 - 93)
t-distribution method
τ^2 = 0.04 (I^2 = 30%, 90% range 20 - 37)
HKSJ method
τ^2 = 1.06 (I^2 = 90%, 90% range 83 - 93)
HKSJ method
Coverage (%)
Number of studies (log scale)


CA
DL
PM
PM_CA
PM_DL
HM
SJ
SJ_CA
REML
τ² = 0.2 (I² = 30%, 90% range 20 - 37)
τ² = 0.74 (I² = 60%, 90% range 46 - 68)
Wald method
t-distribution method
HKSJ method
Coverage (%)
Number of studies (log scale)
100
95
90
85
2
3
5
10
20
30
50
100