Methods for estimating between-study variance and overall effect in meta-analysis of odds-ratios

Ilyas Bakbergenuly*[1] | David C. Hoaglin[2] | Elena Kulinskaya[3]

[1]School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, UK
[2]Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, MA 01605, USA
[3]School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, UK

Correspondence
Ilyas Bakbergenuly, School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, UK. Email: i.bakbergenuly@uea.ac.uk

Abstract

In random-effects meta-analysis the between-study variance (τ^2) has a key role in assessing heterogeneity of study-level estimates and combining them to estimate an overall effect. For odds ratios the most common methods suffer from bias in estimating τ^2 and the overall effect and produce confidence intervals with below-nominal coverage. An improved approximation to the moments of Cochran's Q statistic, suggested by Kulinskaya and Dollinger, yields new point and interval estimators (KD) of τ^2 and of the overall log-odds-ratio. Another, simpler approach (SSW) uses weights based only on study-level sample sizes to estimate the overall effect.

In extensive simulations we compare our proposed estimators with established point and interval estimators for τ^2 and point and interval estimators for the overall log-odds-ratio (including the Hartung-Knapp-Sidik-Jonkman interval). Additional simulations included three estimators based on generalized linear mixed models and the Mantel-Haenszel fixed-effect estimator.

Results of our simulations show that no single point estimator of τ^2 can be recommended exclusively, but Mandel-Paule and KD provide better choices for small and large K, respectively. The KD estimator provides reliable coverage of τ^2. Inverse-variance-weighted estimators of the overall effect are substantially biased, as are the Mantel-Haenszel odds-ratio and the estimators from the generalized linear mixed models. The SSW estimator of the overall effect and a related confidence interval provide reliable point and interval estimation of the overall log-odds-ratio.

KEYWORDS:
between-study variance, heterogeneity, random-effects model, meta-analysis, binary outcome

1 | INTRODUCTION

Meta-analysis is broadly used for combining estimates of a measure of effect from a set of studies in order to estimate an overall (pooled) effect. In studies with binary individual-level outcomes, the most common measure of treatment effect is the odds

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/jrsm.1404

ratio[1]. Our primary interest lies in meta-analysis of odds ratios. The actual measure of effect is the logarithm of the odds ratio (LOR), and the summary data are the numbers of subjects and the numbers of events in the two arms of each study, from which the usual analysis calculates the logarithm of each study's sample odds ratio and the large-sample estimate of its variance. A fixed-effect (or common-effect) model (FEM) assumes that the studies share a single true effect. It is usually more likely that the true study-level effects differ. A random-effects model (REM) describes that variation via a distribution, whose mean serves as the overall effect and whose variance summarizes the heterogeneity of the true study-level effects. Higgins et al.[2] point out, "This variance explicitly describes the extent of the heterogeneity and has a crucial role in assessing the degree of consistency of effects across studies, which is an element of random-effects meta-analysis that often receives too little attention."

We focus mainly on two-stage approaches, which first calculate the studies' log-odds-ratios (and their estimated variances) and then combine those estimates; but we include, for limited comparisons, some one-stage approaches, which use the studies' numbers of events and subjects (e.g., in a binomial likelihood) and avoid calculating the sample log-odds-ratios. To estimate the overall effect, the most common methods use a weighted average of the study-level estimates in which the weight for a study's estimate is the reciprocal of an estimate of its variance. Under the REM such inverse-variance weights combine the variance of the study-level estimate and the variance of the distribution of true study-level effects (τ^2). Thus, they require an estimate of the between-study variance. Most of the common inverse-variance-weighted methods estimate τ^2 by using the theoretical moments of Cochran's Q or its generalization. However, Kulinskaya and Dollinger[3] and van Aert et al.[4] have shown that, for log-odds-ratio, the distributions assumed for those theoretical moments are incorrect. As a result, the moment-based point estimators of τ^2 are biased, and the interval estimators have coverage below the intended 95% level. Also, in combination with inverse-variance weighting, the departures from assumptions lead to biased point estimation of the overall effect and undercoverage of the associated confidence intervals (CIs). Therefore, for estimating between-study variance, we propose new point and interval estimators based on an improved approximation to the moments of Cochran's Q statistic, suggested by Kulinskaya and Dollinger[3]. For the overall effect, we propose a weighted average in which the weights depend only on the effective sample sizes.

We use simulation to compare bias of our proposed point estimator of τ^2 with that of three previous moment-based estimators (the popular estimators of DerSimonian and Laird[5] and Mandel and Paule[6] and the less-familiar estimator of Jackson[7]) and the restricted-maximum-likelihood estimator, and also to compare coverage of our proposed interval estimator of τ^2 with that of four previous estimators (profile likelihood[8], the Q-profile interval[9], and the generalized Q-profile intervals of Biggerstaff and Jackson[10] and Jackson[7]). We also compare bias and coverage of our proposed point estimator of the overall effect and a companion interval estimator with those of the related inverse-variance-based estimators. We extend the comparisons by including point estimators of τ^2 and point and interval estimators of overall effect obtained from logistic linear mixed-effects models, and also the Mantel-Haenszel estimator of the odds ratio.

Section 2 reviews estimation of study-level log-odds-ratio, and Section 3 briefly reviews random-effects models. Section 4 discusses previous point and interval estimators of between-study variance and introduces the proposed Kulinskaya-Dollinger

method. Section 5 describes the corresponding point and interval estimators of the overall effect. Section 6 presents our simulation study and summarizes its results. In Section 7 we apply the various methods to data on the effect of diuretics on pre-eclampsia. Section 8 offers a concluding summary.

The Web Appendix reviews the logistic linear mixed-effects models, tabulates the methods studied in our simulations, discusses the properties of the M-H estimator under the random-effects model, presents the result of the additional simulations that included the logistic linear mixed-effects estimators and the M-H estimator, and lists our R programs for calculating the proposed estimators.

2 | ESTIMATION OF STUDY-LEVEL LOG-ODDS-RATIO

Consider K studies that used a particular individual-level binary outcome. Each study i reports a pair of independent binomial variables, X_{iT} and X_{iC}, the numbers of events in n_{iT} subjects in the Treatment arm ($j = T$) and n_{iC} subjects in the Control arm ($j = C$); for $i = 1, \ldots, K$,

$$X_{iT} \sim Binom(n_{iT}, p_{iT}) \qquad \text{and} \qquad X_{iC} \sim Binom(n_{iC}, p_{iC}). \tag{2.1}$$

The log-odds-ratio for Study i is

$$\theta_i = \log_e\left(\frac{p_{iT}(1-p_{iC})}{p_{iC}(1-p_{iT})}\right) \qquad \text{estimated by} \qquad \hat{\theta}_i = \log_e\left(\frac{\hat{p}_{iT}(1-\hat{p}_{iC})}{\hat{p}_{iC}(1-\hat{p}_{iT})}\right). \tag{2.2}$$

The large-sample estimate of the variance of $\hat{\theta}_i$, derived by the delta method, is

$$\hat{\sigma}_i^2 = \widehat{\mathrm{Var}}(\hat{\theta}_i) = \frac{1}{n_{iT}\hat{p}_{iT}(1-\hat{p}_{iT})} + \frac{1}{n_{iC}\hat{p}_{iC}(1-\hat{p}_{iC})} \tag{2.3}$$

(in finite samples $\mathrm{Var}(\hat{\theta}_i)$ is not finite). Evaluation of $\hat{\theta}_i$ and $\hat{\sigma}_i^2$ requires the estimates $\hat{p}_{ij}$. The usual (and maximum-likelihood) estimate of p_{ij} is $\hat{p}_{ij} = x_{ij}/n_{ij}$, but an adjustment is necessary when either of the observed counts x_{ij} is 0 or n_{ij} (i.e., when the 2×2 table for Study i contains a 0 cell). The standard approach adds $1/2$ to x_{iT}, $n_{iT} - x_{iT}$, x_{iC}, and $n_{iC} - x_{iC}$ when the 2×2 table contains exactly one 0 cell, and it omits Study i when the 2×2 table contains two 0 cells. An alternative approach always adds a (> 0) to all four cells of the 2×2 table for each of the K studies; that is, it estimates p_{ij} by $\hat{p}_{ij(a)} = (x_{ij} + a)/(n_{ij} + 2a)$. The most common choice, $a = 1/2$, removes bias of order n^{-1} in $\hat{\theta}_i$ (Gart et al.[11]). It is convenient to denote the resulting estimate of θ_i by $\hat{\theta}_{i(a)}$.

Using $\hat{p}_{ij(a)}$ with $a = 1/2$ in Equation (2.3) yields an estimator of $\mathrm{Var}(\hat{\theta}_{i(a)})$ that is unbiased except for terms of order n^{-3}. When $n_{ij}p_{ij} < 3$, however, that estimator substantially overestimates $\mathrm{Var}(\hat{\theta}_{i(a)})$ (Gart and Zweifel[12]). As far as we are aware, the corresponding small-sample bias of the standard approach has not been calculated. However, using unbiased estimators of the θ_i and $\mathrm{Var}(\hat{\theta}_i)$ does not make the inverse-variance estimator of the combined LOR unbiased, because $1/\widehat{\mathrm{Var}}(\hat{\theta}_i)$ is a biased estimator of $1/\mathrm{Var}(\hat{\theta}_i)$ and the $\hat{\theta}_i$ and their estimated variances are not independent[13,14].

Double-zero studies

Meta-analysis of binary data is challenging when the event rates are low. Such situations may involve so-called double-zero studies (i.e., studies with zero events in both arms or, at the other extreme, $x_{iT} = n_{iT}$ and $x_{iC} = n_{iC}$). Actual practice varies, but often meta-analyses omit these studies. A popular argument is that such studies provide no information on the direction or magnitude of the effect[15].

Simulations that retain double-zero studies are rather scarce. Kuss[16] considers only methods that include double-zero studies without adjustment. Bhaumik et al.[13] refer to their extensive simulation study comparing inclusion (with $a = 1/2$) and exclusion of double-zero studies and claim that inclusion results in less bias in estimation of the overall effect, but negatively affects estimation of τ^2. Cheng et al.[17] provide a review and some limited simulations for $p \leq 0.01$ and $K = 5$. They argue that including double-zero studies is beneficial when θ is 0, but detrimental when a true treatment effect exists. We believe that this issue has no major practical consequences for our simulations (Section 6) because we use $\theta \geq 0$ and $p_{iC} \geq .1$.

3 | RANDOM-EFFECTS MODELS

3.1 | Standard random-effects model

The standard random-effects model assumes that each estimated study-level effect, $\hat{\theta}_i$, has an approximately normal distribution and that the true study-level effects, θ_i, follow a normal distribution:

$$\hat{\theta}_i \sim N(\theta_i, \sigma_i^2) \quad \text{and} \quad \theta_i \sim N(\theta, \tau^2). \tag{3.1}$$

Thus, the marginal distribution of $\hat{\theta}_i$ is $N(\theta, \sigma_i^2 + \tau^2)$. Although the σ_i^2 are generally unknown, they are routinely replaced by their estimates, $\hat{\sigma}_i^2$. A key step involves estimating the between-study variance, τ^2; the most popular random-effects method uses the DerSimonian-Laird estimate[5]. The estimate of the overall effect is then

$$\hat{\theta}_{RE} = \sum_{i=1}^{K} \hat{w}_i \hat{\theta}_i / \sum_{i=1}^{K} \hat{w}_i, \tag{3.2}$$

where $\hat{w}_i = \hat{w}_i(\hat{\tau}^2) = (\hat{\sigma}_i^2 + \hat{\tau}^2)^{-1}$ is the inverse-variance weight for Study i. If the σ_i^2 and τ^2 were known, the variance of $\hat{\theta}_{RE}$ would be $[\sum w_i]^{-1}$ with $w_i = (\sigma_i^2 + \tau^2)^{-1}$. In practice, the variance of $\hat{\theta}_{RE}$ is traditionally estimated by $[\sum \hat{w}_i(\hat{\tau}^2)]^{-1}$, and a confidence interval for θ uses critical values from the normal distribution.

The assumptions in this model (e.g., within-study normality, between-study normality, and known σ_i^2) have become familiar and seldom attract attention. Jackson and White[14], however, advocate careful examination; they conclude that methods that make fewer normality assumptions should be considered more often in practice.

3.2 | Logistic linear mixed-effects models

One alternative approach uses a binomial-normal likelihood; the resulting logistic linear mixed-effects model belongs to the class of generalized linear mixed models (GLMMs) (Turner et al.[18], Stijnen et.al[19]). Kuss[16], Jackson et al.[20], and Bakbergenuly and

Kulinskaya[21] review these GLMM methods. We include a fixed-intercept model (FIM) and a random-intercept model (RIM), equivalent to Models 4 and 5, respectively, of Jackson et al.[20] and to models FIM2 and RIM2 of Bakbergenuly and Kulinskaya[21]. Briefly, the FIM includes fixed control-group effects (log-odds for the control-group probabilities), and the RIM replaces these fixed effects with random effects. Web Appendix A.1 gives more details.

3.3 | Noncentral-hypergeometric-normal model (NCHGN)

When one conditions on the total number of events for Study i, $X_{iT}+X_{iC}=X_i$, only the number of events in the treatment group X_{iT} is random. Then, given the study-specific log-odds-ratio θ_i, X_{iT} has a noncentral hypergeometric distribution. If the θ_i are normally distributed, $\theta_i \sim N(\theta, \tau^2)$, the exact hypergeometric-normal likelihood function for Study i can be written as[22,19]:

$$l_{HGN}(x_{iT};\theta,\tau^2) = \int_{-\infty}^{\infty} \binom{n_{iT}}{x_{iT}}\binom{n_{iC}}{x_{iC}} \frac{\exp(x_{iT}\theta_i)}{P(\theta_i)} \phi(\theta_i|\theta,\tau^2)d\theta_i, \tag{3.3}$$

where the normalizing constant is defined as:

$$P(\theta_i) = \sum_{u=\max(0,X_i-n_{iC})}^{\min(n_{iT},X_i)} \binom{n_{iT}}{u}\binom{n_{iC}}{X_i-u}\exp(u\theta_i),$$

and $\phi(\cdot|\theta,\tau^2)$, is the probability density function of the normal distribution with mean θ and variance τ^2. Integrating out the unobserved study-specific effects produces the marginal distribution of X_{i1}. See Web Appendix A.1 for more details.

4 | METHODS OF ESTIMATING BETWEEN-STUDY VARIANCE

A number of methods provide point and interval estimates of between-study variance. In a comprehensive review of existing simulation and empirical studies, Veroniki et al.[23] focus on general-purpose estimators. Langan et al.[24] systematically review simulation studies that compared estimators of heterogeneity variance. They summarize performance in estimating heterogeneity and also in estimating the overall effect. The studies used a variety of effect measures, including the odds ratio. Langan et al.[25] use simulated data on standardized mean difference and odds ratio to compare nine estimators. We considered the recommendations of those three reports in choosing estimators to study. This section briefly reviews them; for reference, Web Appendix A.2 contains a list. More-detailed descriptions appear in Veroniki et al.[23], Langan et al.[24], and Langan et al.[25] and in Web Appendices A.1 and A.3.

4.1 | Point estimators

In applications the DerSimonian-Laird[5] method remains the most popular; its relative simplicity facilitated its early implementation in software. Accumulating evidence of its inferior performance has done little to dislodge it. Recommended alternative point estimators include restricted maximum likelihood (REML), the method of Mandel and Paule[6], and the method of Jackson[7]. These and other methods have been studied by many authors, including Viechtbauer[26] and Kosmidis et al.[27]. This section

briefly reviews these four methods and describes the Kulinskaya-Dollinger method. Information on the logistic linear mixed-effects models (FIM, RIM, and NCHGN) appears in Web Appendix A.1. All these methods replace negative values of $\hat{\tau}^2$ with zero.

4.1.1 | DerSimonian-Laird method (DL)

When $\tau^2 = 0$, the statistic $Q = \sum \hat{w}_i(\hat{\theta}_i - \hat{\theta})^2$, with $\hat{w}_i = \hat{w}_i(0) = 1/\hat{\sigma}_i^2$ and $\hat{\theta} = \sum \hat{w}_i\hat{\theta}_i / \sum \hat{w}_i$, is customarily assumed to have approximately the chi-squared distribution on $K - 1$ degrees of freedom. DerSimonian and Laird[5] substitute $w_i = 1/\sigma_i^2$ for $\hat{w}_i$, derive the corresponding expected value of Q when $\mathrm{Var}(\hat{\theta}_i) = \sigma_i^2 + \tau^2$, and estimate τ^2 by the method of moments. The resulting closed-form expression has made the DL estimator attractive.

4.1.2 | Restricted-maximum-likelihood method (REML)

Assuming that the $\hat{\theta}_i$ are distributed as $N(\theta, \hat{\sigma}_i^2 + \tau^2)$, the restricted-maximum-likelihood (REML) estimator $\hat{\tau}^2_{REML}$ maximizes the restricted (or residual) log-likelihood function $l_R(\theta, \tau^2)$, which differs from the ordinary likelihood function by the addition of $-\frac{1}{2}\ln(\sum \hat{w}_i(\tau^2))$. It is obtained iteratively, using $\theta = \hat{\theta}_{REML}$ from Equation (3.2) with weights $\hat{w}_i(\hat{\tau}^2_{REML})$. REML is superior to DL because of its balance between unbiasedness and efficiency (Viechtbauer[26]). However, like DL, using the $\hat{\sigma}_i^2$ as if they were the σ_i^2 may undermine its performance.

One can also obtain the REML estimator of τ^2 by maximizing the penalized log-likelihood developed by Kosmidis et al.[27] to reduce the bias of maximum-likelihood estimation.

4.1.3 | Mandel-Paule method (MP)

The Mandel-Paule (MP) estimator, $\hat{\tau}^2_{MP}$, is another iterative moment-based estimator of the between-study variance[6,28].

As in Section 3.1, let the random-effects weights and $\hat{\theta}_{RE}$ depend on τ^2; denote the resulting Q by $Q(\tau^2)$. The Mandel-Paule estimator $\hat{\tau}^2_{MP}$ is obtained by iteratively solving the equation

$$Q(\tau^2) = \sum_{i=1}^{K} \hat{w}_i(\tau^2)(\hat{\theta}_i - \hat{\theta}_{RE})^2 = K - 1 \tag{4.1}$$

and requiring $\hat{\tau}^2_{MP} > 0$.

This method is equivalent to the empirical Bayes methods of Carter and Rolph[29] and Morris[30], as noted by Rukhin and Vangel[31] and Rukhin et al.[32].

4.1.4 | Jackson method (J)

DerSimonian and Kacker[33] generalize Q, replacing the $\hat{w}_i$ by arbitrary fixed positive constants, a_i, to obtain $Q_a = \sum a_i(\hat{\theta}_i - \hat{\theta}_a)^2$, from which they derive a general method-of-moments estimator of τ^2. They discuss several special cases, including DL (with $a_i = 1/\hat{\sigma}_i^2$, treating the $\hat{\sigma}_i^2$ as fixed).

As an option when some heterogeneity is anticipated but there is little prior knowledge about its extent, Jackson[7] uses Q_a with $a_i = 1/\sigma_i$. Although that choice yields a point estimator of τ^2, he focuses on the interval estimator. However, the R function *inference* in the supplementary materials of Jackson[7] returns the point estimate. (His computational procedure avoids negative $\hat{\tau}^2$.) We abbreviate the point and interval estimators as J. In practice, meta-analyses would use the $\hat{\sigma}_i$, so the a_i in Q_a are not fixed.

4.1.5 | Kulinskaya-Dollinger method (KD)

The chi-squared approximation for Q is inaccurate, and the actual distribution depends on the effect measure. Under the null hypothesis of homogeneity of the log-odds-ratio, Kulinskaya and Dollinger[3] obtain corrected approximations for the mean and variance of Q and match those corrected moments to obtain a gamma distribution that (as their simulations confirm) closely fits the null distribution of Q. These approximations blend theoretical derivations with simulation results. Let $E_{KD}(Q)$ denote the corrected expected value of Q under the null hypothesis $\tau^2 = 0$. This corrected first moment has the form $E_{KD}(Q) = K - 1 - 0.687[K - 1 - E_{th}(Q)]$, where $E_{th}(Q)$ is a theoretical moment obtained from their general expansion for the mean of Q for arbitrary effect measures[34]. The corrected variance of Q is a quadratic function of the corrected mean $E_{KD}(Q)$. The expression for $E_{th}(Q)$ involved in specifying the corrected distribution of Q is not simple; Kulinskaya and Dollinger[3] give the details. For large sample sizes, $E_{th}(Q) \to K - 1$.

We propose a new estimator of τ^2 based on this improved approximation. One obtains the KD estimate $\hat{\tau}^2_{KD}$ by iteratively solving

$$Q(\tau^2) = \sum_{i=1}^{K} \frac{(\theta_i - \hat{\theta}_{RE})^2}{\hat{\sigma}_i^2 + \tau^2} = E_{KD}(Q). \tag{4.2}$$

This estimator closely resembles the Mandel-Paule estimator; both assume that adding τ^2 to $\hat{\sigma}_i^2$ in the IV weights makes the non-null distribution of Q (or at least, its mean) close to its null distribution. This assumption needs to be verified by simulation.

4.2 | Interval estimators

Viechtbauer[9] and Jackson and Bowden[35] compare confidence-interval estimators of the between-study variance. Interval estimators recommended by Veroniki et al.[23] include profile likelihood[8], the Q-profile interval[9], and the generalized Q-profile intervals of Biggerstaff and Jackson[10] and Jackson[7]. Quality of estimation varies with the effect measure; for odds ratio van Aert et al.[4] found that coverage of the last three methods can deviate substantially from the nominal 95% level. If the lower confidence limit is not defined or is negative, all these methods set it to zero. The logistic linear mixed-effects methods (FIM, RIM, and NCHGN) as implemented in the *rma.glmm* function in *metafor*[36], used in our simulations, do not produce confidence intervals for τ^2.

4.2.1 | Profile-likelihood interval (PL)

The 95% profile-likelihood confidence interval for τ^2 consists of the values that are not rejected by the likelihood-ratio test with τ^2 as the null hypothesis (Hardy and Thompson[8]). Here the other parameter in the likelihood, $\hat{\theta}$, is a function of τ^2, as in

Equation (3.2). The values of τ^2 in the confidence interval satisfy

$$\left\{\tau^2 : \quad l_R(\hat{\theta}(\tau^2), \tau^2) > l_R(\hat{\theta}_{REML}, \hat{\tau}^2_{REML}) - \frac{1}{2}\chi^2_{1;0.95}\right\}, \tag{4.3}$$

where $\chi^2_{1;0.95} = 3.841$ is the 0.95 quantile of the χ^2_1 distribution, and $l_R(\hat{\theta}(\tau^2), \tau^2)$ is the restricted log-likelihood function evaluated at $(\hat{\theta}(\tau^2), \tau^2)$.

4.2.2 | Q-profile confidence interval (QP)

If the weight for Study i is $1/(\sigma_i^2 + \tau^2)$, the generalized Q-statistic

$$Q(\tau^2) = \sum_{i=1}^{K} \frac{(\hat{\theta}_i - \hat{\theta}(\tau^2))^2}{\sigma_i^2 + \tau^2} \tag{4.4}$$

follows the chi-squared distribution with $K - 1$ degrees of freedom. To obtain the Q-profile confidence interval, Viechtbauer[9] finds the lower and upper confidence limits by iteratively solving $Q(\tilde{\tau}^2_L) = \chi^2_{K-1;0.975}$ and $Q(\tilde{\tau}^2_U) = \chi^2_{K-1;0.025}$. In practice it is necessary to use the $\hat{\sigma}_i^2$ instead of the σ_i^2, and then the generalized Q-statistic no longer follows the assumed chi-squared distribution.

4.2.3 | Biggerstaff and Jackson interval (BJ)

For a generic effect measure, Biggerstaff and Jackson[10] derive the exact distribution of the statistic

$$Q = \sum_{i=1}^{K} w_i(\hat{\theta}_i - \hat{\theta})^2, \tag{4.5}$$

where $w_i = 1/\sigma_i^2$ and $\hat{\theta} = (\sum w_i\hat{\theta}_i)/(\sum w_i)$. They show that the distribution is that of a linear combination of mutually independent chi-squared random variables, each with 1 degree of freedom, and they take advantage of available software for evaluating the cumulative distribution function F_Q of such a distribution.

That distribution yields a generalized Q-profile confidence interval, whose lower and upper limits are the solutions to the equations

$$Q(\tilde{\tau}^2_L) = F_{Q;0.975}, \qquad Q(\tilde{\tau}^2_U) = F_{Q;0.025}, \tag{4.6}$$

in which $F_{Q;0.025}$ and $F_{Q;0.975}$ are, respectively, the 0.025 and 0.975 quantiles. If the equation for $\tilde{\tau}^2_L$ has no solution, they set $\tilde{\tau}^2_L = 0$. We refer to this interval as the BJ confidence interval.

Despite the title of Biggerstaff and Jackson[10], Q in (4.5) is not Cochran's heterogeneity statistic. In the definition of Q, Cochran[37] used $w_i = 1/\hat{\sigma}_i^2$.

4.2.4 | Jackson interval (J)

As mentioned in Section 4.1.4, Jackson[7] proposes another generalized Q-profile confidence interval for τ^2. The approach is the same as for the BJ interval, but with $a_i = 1/\sigma_i$ in Q_a.

4.2.5 | Kulinskaya-Dollinger interval (KD)

For the log-odds-ratio, we propose a new confidence interval for the between-study variance. The KD confidence interval for τ^2 combines the Q-profile approach and the improved approximation by Kulinskaya and Dollinger[3]. This corrected Q-profile confidence interval can be estimated from the lower and upper quantiles of F_Q, the cumulative distribution function for the corrected distribution of Q, as in equation (4.6). The upper and lower confidence limits for τ^2 can be calculated iteratively.

5 | METHODS OF ESTIMATING OVERALL EFFECT

Most of the point estimators of the overall effect have corresponding interval estimators, but some do not. Therefore, we describe point estimators and interval estimators in separate sections.

5.1 | Point estimators

A random-effects method that estimates θ by a weighted mean with inverse-variance weights, as in Equation (3.2), is determined by the particular $\hat{\tau}^2$ that it uses in $\hat{w}_i(\hat{\tau}^2)$. The best-known and most widely used estimator, $\hat{\theta}_{DL}$, was introduced by DerSimonian and Laird[5]; it uses $\hat{\tau}^2_{DL}$. Its shortcomings, in particular bias and below-nominal coverage of the companion confidence interval, have led numerous authors to propose alternative estimators of τ^2. Some of those shortcomings arise from the derivation underlying $\hat{\tau}^2_{DL}$, which uses the σ_i^2 and τ^2 and then substitutes the $\hat{\sigma}_i^2$ and $\hat{\tau}^2$. Unfortunately, the alternative methods (REML, J, and MP) generally rely on that same unsupported substitution. In our simulations, we add one more inverse-variance-weighted estimator, KD, to this list.

In an attempt to avoid the bias in the inverse-variance-weighted estimators, we include a point estimator whose weights depend only on the studies' effective sample sizes (Hedges and Olkin[38], Hunter and Schmidt[39]). For this estimator (SSW) $\hat{\theta}_i$ uses $\hat{p}_{ij(a)}$ with $a = 1/2$ (as discussed in Section 2), and $w_i = \tilde{n}_i = n_{iT}n_{iC}/(n_{iT} + n_{iC})$; $\tilde{n}_i$ is the effective sample size in Study i. These weights would be equivalent to the inverse-variance weights if all the probabilities across studies were equal (i.e., $p_{iT} = p_{iC} \equiv p$ for $i = 1, \ldots, K$).

As we mentioned in Section 1, we also include estimators obtained from logistic linear mixed-effects models, namely FIM, RIM, and NCHGN.

A reviewer pointed out that the weights in SSW are the same as those in the Mantel-Haenszel estimator of a common risk difference[40], and suggested that we include the Mantel-Haenszel estimator (MH) of a common odds ratio. That fixed-effect estimator applies the weight $(n_{iT} - x_{iT})x_{iC}/(n_{iT} + n_{iC})$ to the sample odds ratio for Study i. As we discuss in Web Appendix A.3, we expect MH to be biased under the REM.

In summary, the point estimators that we study are DL, REML, J, MP, KD, SSW, FIM, RIM, NCHGN, and MH.

5.2 | Interval estimators

The point estimators DL, REML, J, MP, and KD have companion interval estimators of θ. The customary approach estimates the variance of $\hat{\theta}_{RE}$ by $[\sum \hat{w}_i(\hat{\tau}^2)]^{-1}$ and bases the width of the interval on the normal distribution. That expression for the variance of $\hat{\theta}_{RE}$ would be correct if it were based on $w_i = (\sigma_i^2 + \tau^2)^{-1}$. In practice, however, using $\hat{w}_i(\hat{\tau}^2)$ may not yield a satisfactory approximation. Also, we have not seen empirical evidence that the sampling distributions of $\hat{\theta}_{RE}$ for the various choices of estimator for τ^2 are adequately approximated by a normal distribution.

Hartung and Knapp[41] and, independently, Sidik and Jonkman[42] developed an estimator for the variance of $\hat{\theta}_{DL}$ that takes into account the variability of the $\hat{\sigma}_i^2$ and $\hat{\tau}^2$. The Hartung-Knapp-Sidik-Jonkman (HKSJ) confidence interval uses the estimator

$$\widehat{\text{Var}}_{HKSJ}(\hat{\theta}_{DL}) = \sum_{i=1}^{K} \hat{w}_i(\hat{\tau}_{DL}^2)(\hat{\theta}_i - \hat{\theta}_{DL})^2 / [(K-1)\sum_{i=1}^{K} \hat{w}_i(\hat{\tau}_{DL}^2)], \tag{5.1}$$

together with critical values from the t distribution on $K-1$ degrees of freedom. A potential weakness is that the derivation of the variance estimator and the t distribution uses the σ_i^2 and τ^2 and then substitutes the $\hat{\sigma}_i^2$ and $\hat{\tau}_{DL}^2$. Also, the HKSJ interval uses $\hat{\theta}_{DL}$ as its midpoint, so it will have any bias that is present in $\hat{\theta}_{DL}$. We study a modification of HKSJ (HKSJ KD) that uses the KD estimator of τ^2 and uses $\hat{\theta}_{KD}$ as the midpoint.

The interval estimator corresponding to SSW (SSW KD) uses the SSW point estimator as its center, and its width equals the estimated standard deviation of SSW under the random-effects model times twice the critical value from the t distribution on $K-1$ degrees of freedom. The estimator of the variance of SSW is

$$\widehat{\text{Var}}(\hat{\theta}_{SSW}) = \frac{\sum \tilde{n}_i^2(v_i^2 + \hat{\tau}^2)}{(\sum \tilde{n}_i)^2}, \tag{5.2}$$

in which v_i^2 comes from Equation (2.3) and $\hat{\tau}^2 = \hat{\tau}_{KD}^2$.

In summary, the interval estimators that we study are DL, REML, J, MP, KD, HKSJ, HKSJ KD, SSW KD, FIM, RIM, and NCHGN.

6 | SIMULATION STUDY

In a simulation study with log-odds-ratio as the effect measure, we varied six parameters: the number of studies K, the total sample size of each study n, the proportion of observations in the Control arm q, the overall true LOR θ, the between-study variance τ^2, and the probability of an event in the Control arm p_C. The number of studies $K \in \{5, 10, 30\}$. We included sample sizes that were equal for all K studies and sample sizes that varied among studies. The total sample sizes were $n \in \{40, 100, 250, 1000\}$ for equal sample sizes and the average total sample sizes were $\bar{n} \in \{30, 60, 100, 160\}$ for unequal sample sizes. In choosing sample sizes that varied among studies, we followed a suggestion of Sánchez-Meca and Marín-Martínez[43], who selected study sizes having skewness 1.464, which they considered typical in behavioral and health sciences. For $K = 5$, Table 1 lists the sets of five sample sizes, which have the chosen skewness and average equal to 30, 60, 100, and 160. The simulations for $K = 10$ and $K = 30$ used each set of unequal sample sizes twice and six times, respectively. The values of q were .5 and .75. The sample sizes of the Treatment and Control arms were $n_{iT} = \lceil (1-q)n_i \rceil$ and $n_{iC} = n_i - n_{iT}$, $i = 1, \ldots, K$.

TABLE 1 Unequal sample sizes for simulations with $K = 5$

$\bar{n} \setminus i$	1	2	3	4	5
30	12	16	18	20	84
60	24	32	36	40	168
100	64	72	76	80	208
160	124	132	136	140	268

The values of the overall true LOR were $\theta = 0(0.5)2$ (that is, from 0 to 2 in steps of 0.5). The probability in the Control arm was $p_C = .1,\ .2,\ .4$. The values of the between-study variance were $\tau^2 = 0(0.1)1$, corresponding to small to moderate heterogeneity. This interval of τ^2 values is similar to or, for smaller sample sizes, somewhat shorter than that for the meta-analyses of LOR in the Cochrane database, Langan et al.[25] Appendix 2.

Altogether, the simulations comprised 7,920 combinations of the six parameters. We generated 10,000 meta-analyses for each combination. The true values of LOR (θ_i) were generated from a normal distribution with mean θ and variance τ^2. For a given p_{iC}, the number of events in the control group, X_{iC}, was generated from the Binomial(n_{iC}, p_{iC}) distribution. The number of events in the treatment group, X_{iT}, was generated from the Binomial(n_{iT}, p_{iT}) distribution with $p_{iT} = p_{iC}\exp(\theta_i)/(1 - p_{iC} + p_{iC}\exp(\theta_i))$. The estimate $\hat{\theta}_i$ was calculated as in Equation (2.2), and its sampling variance was estimated by substituting $\hat{p}_{iT}$ and $\hat{p}_{iC}$ in Equation (2.3). The methods differed however, in the way they obtained $\hat{p}_{ij}$ from x_{ij} and n_{ij}. In all standard methods, we added $1/2$ to each cell of the 2×2 table only when the table had at least one cell equal to 0. This approach corresponds to the default values of the arguments **add**, **to**, and **drop00** of the *escalc* procedure in *metafor*[36]. In the KD methods, and for estimation of τ^2 in SSW, we corrected for bias by adding $1/2$ to each cell of all K tables. We also tried always adding $1/2$ in the standard methods, but that made the biases for $\hat{\tau}^2$ worse.

Expanding our comparative study, we included the MH estimator of θ and the estimators from the FIM, RIM, and NCHGN models in simulations for selected values of the parameters: $p_C = 0.1$, $q = 0.5$ and equal sample sizes with $n = 40$ and $n = 100$. The three logistic linear mixed-effects methods provide point but not interval estimators of τ^2 and both point and interval estimators of θ. For the MH point estimator of θ, we studied two versions: the usual version (MH), which does not modify the cell counts, and a version that always adds $1/2$ to each cell (MH with $1/2$). The results of these additional simulations are plotted in Web Appendix A.4.

6.1 | Results of simulation studies

Our full simulation results are available as an arXiv e-print (Bakbergenuly et al.[44]). They comprise 300 figures, each presenting a plot of bias or coverage versus τ^2 for the four values of n or $\bar{n}$ and the three values of K. A detailed summary is given below and illustrated by Figures 1 to 3.

Bias in estimation of τ^2 (Figure 1)

All the estimators have bias that varies with τ^2, often roughly linearly. The sign and magnitude of the bias and the slope of that relation depend on p_{i2}, θ, n, K, and q. For example, when $p_{i2} = .1$, $\theta = 0$, $q = .5$, $n = 40$, and $K = 5$, the bias of KD goes from +0.32 when $\tau^2 = 0$ to −0.08 when $\tau^2 = 1$, and the traces for the other estimators, close together, go from around +0.12 to around −0.47. Among these, MP appears to be the least biased. As K increases, the pattern shifts down; and as n increases, the traces tend to flatten (when $n = 1000$, most of the estimators are unbiased, but the bias when $\tau^2 = 1$ is −0.08 for J and −0.17 for DL). As θ increases, the patterns shift down. When all studies are unbalanced (in favor of the control arm), $q = .75$, the patterns often shift down, and the slopes become steeper.

Figure 1 shows these patterns for KD and MP in the balanced case ($q = .5$). Both estimators have positive bias at zero, but for larger values of τ^2, the bias of MP is mostly negative, whereas for KD it may be positive for larger values of θ. MP is considerably worse than KD (apart from $\tau^2 = 0$) for $K = 30$. For $K = 5$ and 10, KD is less biased than MP for large values of τ^2, but it may be worse for small values.

The effect of increasing p_{i2} is not simple. As p_{i2} increases from .1 to .2 to .4, the (positive) bias of KD at $\tau^2 = 0$ decreases, and its bias at $\tau^2 = 1$ approaches 0; at $\tau^2 = 0$ the (positive) bias of the other estimators changes little, but at $\tau^2 = 1$ the magnitude of the (negative) bias decreases when $\theta = 0$ but decreases and then increases when $\theta = 2$.

None of the point estimators of τ^2 has bias consistently close enough to 0 to be recommended; but among the existing estimators, MP and KD provide better choices for small and large K, respectively.

Bias in estimation of θ (Figure 2)

In the results for bias of the point estimators of θ, a common pattern is that the bias is roughly linearly related to τ^2 with a positive slope. The varied positions of the estimators' traces relative to the horizontal line of zero bias, however, complicate the process of summarizing. The situation with $p_{iC} = .1$ and $\theta = 0$ is straightforward: When $n = 40$ and $K = 5$, all estimators have no bias when $\tau^2 = 0$; when $\tau^2 = 1$, SSW has bias 0.14, and the other estimators' biases range from 0.23 to 0.26. Increasing K (to 10 and 30) has little effect on the pattern, and increasing n (to 100, 250, and 1000) flattens the pattern until little bias remains. (The plots for $n = 100$ show that the bias of SSW decreases more rapidly.) When $\theta = 0.5$, the pattern splits into three: SSW has much smaller slope and flattens to essentially zero bias; the bias of KD changes from negative to positive around $\tau^2 = 0.5$; and the common trace for the other estimators parallels that for KD and is about 0.06 units above it. Again, by $n = 1000$ the traces have flattened and merged. As θ increases (to 1.0, 1.5, and 2.0), the traces for all estimators except SSW shift down further, and the gap between KD and the others widens. When $p_{iC} = .2$ and $\theta > 0$, slopes of the non-SSW traces decrease as θ increases (the traces are flat when $\theta = 2$). When $p_{iC} = .4$ and $\theta > 0$, the non-SSW traces go from flat to having negative slope as θ increases. Also, increasing K tends to shift those traces down slightly.

When all K studies are unbalanced ($q = .75$), $p_{iC} = .1$, $\theta = 0$, and $n = 40$, the estimators have larger positive bias, even when $\tau^2 = 0$. This effect decreases as θ and p_{iC} increase, consistent with the behavior when $q = .5$, and it is absent when $n \geq 100$.

As expected, in the vast majority of situations, SSW avoids most, if not all, of the bias in the IV-weighted estimators. The bias of the IV-weighted estimators affects their efficiency, so SSW tends to have smaller mean squared error than MP as τ^2 and K increase, but larger MSE than KD when $K = 5$ and $K = 10$ and when $K = 30$ and τ^2 is small.

Coverage in estimation of τ^2 (Figure 3)

Coverage of τ^2 is generally good for $K = 5$, but is considerably worse for larger numbers of studies, especially so for large values of θ. All methods are somewhat conservative at $\tau^2 = 0$. When $K = 5$, PL is very conservative, whereas KD provides close to nominal coverage for $\tau^2 > 0$, though it may become a bit liberal for large θ. The other methods are between these two, being somewhat conservative for small sample sizes n. For $K = 10$, PL is still mostly conservative, though it may become somewhat liberal for larger τ^2. KD is almost perfect, though in one instance, for unequal sample sizes with $\bar{n} = 30$, $p_C = .4$, and $\theta = 2$, its coverage drops to 90%. The other intervals are too liberal for small n. The large number of studies K presents the greatest challenge for the standard methods. PL is the most affected, with considerable undercoverage up to $n = 100$ for medium to large values of τ^2. The other methods also have low coverage for small n, but they improve faster with increasing n. KD provides reliable coverage except for small sample sizes combined with $p_C = .4$ and $\theta \geq 1.5$, where its undecoverage worsens with increasing τ^2, though it is still considerably better than all the competitors.

Coverage in estimation of θ (Figure 3)

Interval estimators of θ respond in a variety of ways to the variables in the simulations. No simple description adequately summarizes the patterns. In one common pattern, coverage decreases as τ^2 increases, often falling substantially below the nominal 95% for the IV-weighted estimators. For a given value of θ and $K = 10$ and $K = 30$, undercoverage tends to decrease as n increases. For $K = 5$, however, the undercoverage of the IV-weighted estimators generally increases as n goes from 40 to 100 to 250 to 1000; when $n = 1000$, coverage is around 95% when $\tau^2 = 0$ and roughly constant, at several percentage points below 95%, for $0.1 \leq \tau^2 \leq 1$ (the decrease is greater for $p_{iC} = .2$ and $p_{iC} = .4$ than for $p_{iC} = .1$). Because HKSJ, HKSJ KD, and SSW KD do not exhibit such undercoverage in these situations, the explanation is likely to lie in the use of the normal distribution as the basis for the CI.

On the other hand, for given values of θ, $n = 40$ and $n = 100$, and $\tau^2 > 0$, coverage tends to decrease as K increases. This effect is small for SSW KD (which moves from overcoverage to coverage close to 95%) and larger (by varying amounts) for all the other estimators. Thus, counterintuitively, when more than a small amount of heterogeneity is present and $n \leq 100$, increasing the number of studies decreases coverage. A likely contributor is bias in estimating θ, which (for $n = 40$ and $n = 100$) is positive and increasing as τ^2 increases, and changes little with K.

A different pattern arises when $\theta \geq 1$, $n = 40$ and $n = 100$ (and $\bar{n} = 30$ and 100), and $K = 30$. Coverage of HKSJ KD and KD is below 95% when $\tau^2 = 0$ and increases toward 95% as τ^2 increases. For KD the explanation probably lies in its bias in estimating θ, which is negative and rises toward 0 (but remains < 0) as τ^2 increases. For HKSJ KD (which has greater undercoverage), the reason is less clear. Undercoverage of both KD and HKSJ KD at $\tau^2 = 0$ increases as θ increases. This

pattern arises when $p_{iC} = .1$ and $p_{iC} = .2$. When $p_{iC} = .4$, however, it is not evident when $\theta = 1$. When $\theta \geq 1.5$, coverage of KD and HKSJ KD decreases as τ^2 increases and then stabilizes.

We do not recommend standard confidence intervals based on IV-weighted estimators of θ, because of their undercoverage. HKSJ and HKSJ KD often have coverage close to 95%, but they sometimes have serious undercoverage. All problems are typically worse for the unbalanced sample sizes. The SSW KD interval often has coverage somewhat greater than 95%, but its coverage is at least 93% (except for a few cases involving $K = 30$ and unequal sample sizes with $\bar{n} = 30$).

Additional results: FIM, RIM, NCHGN, and MH

In estimating τ^2, FIM and RIM (Figure A4.1) often have bias that is between those of $\hat{\tau}^2_{KD}$ and $\hat{\tau}^2_{MP}$ (Figure 1) and is generally not small, going from positive near $\tau^2 = 0$ to negative at larger τ^2. The size of their bias tends to decrease as K increases. As θ increases, the bias of RIM tends to decrease, whereas the bias of FIM remains roughly constant. The pattern of NCHGN is more complicated: positive and decreasing as K increases and θ increases when $n = 40$; but roughly linear (+ to −) in τ^2, increasing as θ increases, and flattening as K increases when $n = 100$, where NCHGN is almost unbiased for larger τ^2 when $K = 30$. However, convergence rates of NCHGN are rather low, especially so for low values of τ^2 and K; they improve somewhat for larger values of n (Figure A4.4).

For point estimation of θ (Figure A4.2), the biases of FIM, RIM, NCHGN, and MH follow patterns that resemble those of KD and MP and are quite unlike the (generally more favorable) patterns of SSW. The bias of MH increases with τ^2, starting at 0 when $\theta = 0$ but at around −0.1 to −0.05 when $\theta = 1$. MH with 1/2 is less positively biased than MH when $\theta = -1$ or 0, but more negatively biased than MH for low values of τ^2 when $\theta = 1$ (Figures A3.1 and A3.2).

For interval estimation of θ (Figure A4.3), FIM, RIM, and NCHGN generally have lower coverage than the other estimators, decreasing as θ increases. When $n = 100$, the coverage of RIM decreases rapidly as τ^2 increases, and that pattern becomes more pronounced as K increases.

In summary, we do not recommend MH or the GLMMs for point or interval estimation of θ.

7 | EXAMPLE: EFFECTS OF DIURETICS ON PRE-ECLAMPSIA

Data from nine trials that reported the effect of diuretics on pre-eclampsia Collins et al.[45] were studied by Hardy and Thompson[8], Biggerstaff and Tweedie[46], Turner et al.[18], Viechtbauer[9], Kulinskaya and Olkin[47], and Bakbergenuly and Kulinskaya[48]. The data are shown in Table 2 and are re-analyzed here in order to compare the methods of point and interval estimation of between-study variance and the log-odds-ratio. estimation in interval estimation of between study variance, eight methods in point estimation of overall effect measure and ten methods in confidence interval estimation of overall effect measure. For comparison we include results from three generalized linear mixed models available in the *metafor* package[36]: the fixed-intercept model (FIM), the random-intercept model (RIM), and the exact method based on the noncentral hypergeometric distribution (NCHGN). Bakbergenuly and Kulinskaya[21] give more details on those methods.

TABLE 2 Data for meta-analysis on effects of diuretics on pre-eclampsia,[45]

Study	y_{iT}	y_{iC}	n_{iT}	n_{iC}	$\hat{p}_{iT}$	$\hat{p}_{iC}$	$\hat{\theta}_i$	$\tilde{n}_i$
1	14	14	131	136	0.1068	0.1029	0.042	66.727
2	21	17	385	134	0.0545	0.1268	-0.924	99.403
3	14	24	57	48	0.2456	0.5000	-1.122	26.057
4	6	18	38	40	0.1579	0.4500	-1.473	19.487
5	12	35	1011	760	0.0118	0.0460	-1.391	433.857
6	138	175	1370	1336	0.1007	0.1310	-0.297	676.393
7	15	20	506	524	0.0296	0.0382	-0.262	257.421
8	6	2	108	103	0.0555	0.0194	1.089	52.720
9	65	40	153	102	0.4248	0.3921	0.135	61.200

Table 3 provides the point estimates of the between-study variance and the point estimates and confidence intervals for the overall log-odds-ratio and the overall odds ratio; and Table 4 shows the point estimates and confidence intervals for the between-study variance. DL has the lowest estimate of τ^2, 0.230, followed by the GLMM estimates at 0.254 to 0.264, and KD gives the highest estimate, 0.392. MP is second highest at 0.386. QP provides the longest confidence interval for τ^2, with length 2.130, and KD the second longest at 1.875, whereas BJ is considerably shorter at 1.384, and NCHGN has a very short interval with a length of just 0.667.

In estimating θ, all inverse-variance-weighted methods give similar values, ranging from -0.518 to -0.517 apart from KD which is -0.507, and the GLMM methods also give similar values ranging from -0.516 to -0.513. By contrast the fixed-effect model produces the highest estimate, -0.398, and SSW produces the lowest, -0.558. All the standard inverse-variance-weighted methods and the GLMMs show a significant effect of diuretics on pre-eclampsia, whereas all methods using t quantiles (HKSJ DL, HKSJ KD, and SSW KD) do not find a significant effect.

It is rather difficult to decide, from our simulation results, which method gives the best estimates, as the sample sizes, even though rather balanced, vary greatly, from 38 to 1370 in the Treatment arm. Therefore we ran additional simulations, where we kept the sample sizes and the prevalence in the Control arm as in the actual nine trials, and varied the value of θ from -0.4 to -0.6 and the value of $\tau^2 = 0.20(0.05)0.45$ to cover the range of possible values of these parameters. We used 10, 000 repetitions at each combination of θ and τ^2. Results of these simulations are shown in Figure 4.

From these simulations, MP and KD are the least biased estimates of τ^2; the other methods have considerable negative bias, especially DL and the GLMMs, RIM being the most biased. KD provides the best coverage of τ^2, though the coverage of all methods appears to be reasonable. All methods but SSW considerably overestimate θ, though here NCHGN and FIM are the least biased, with positive biases of 0.01 to 0.03. Coverage of θ is best for SSW KD and somewhat too low for the other methods based on t quantiles. The coverage of the standard IV-weighted methods based on normal quantiles is clearly not acceptable, and the GLMMs provide even worse coverage, probably because of their underestimation of τ^2.

TABLE 3 Meta-analysis of diuretics in pre-eclampsia. Point estimates of the between-study variance τ^2 and point estimates and confidence intervals for the overall log-odds-ratio (θ) and the overall odds ratio (OR); REM is the random-effects model, and FEM is the fixed-effect model. L and U are the lower and upper limits of the 95% confidence intervals.

Model	Method	$\hat{\tau}^2$	$\hat{\theta}$	L	U	Length	OR	L	U
FEM			−0.398	−0.573	−0.223	0.530	0.672	0.564	0.800
REM	DL	0.230	−0.517	−0.916	−0.117	0.799	0.596	0.400	0.889
REM	HKSJ DL		−0.517	−1.061	0.028	1.089	0.596	0.346	1.028
REM	REML	0.300	−0.518	−0.956	−0.080	0.876	0.596	0.384	0.923
REM	J	0.329	−0.518	−0.971	−0.065	0.906	0.596	0.379	0.937
REM	MP	0.386	−0.518	−0.998	−0.037	0.961	0.596	0.369	0.963
REM	KD	0.392	−0.507	−0.987	−0.027	0.960	0.602	0.373	0.973
REM	HKSJ KD	0.392	−0.507	−1.054	0.040	1.094	0.602	0.348	1.040
REM	SSW KD	0.392	−0.558	−1.337	0.221	1.558	0.572	0.263	1.247
GLMM	FIM	0.254	−0.513	−0.923	−0.104	0.819	0.599	0.398	0.901
GLMM	RIM	0.264	−0.516	−0.930	−0.102	0.828	0.597	0.395	0.903
GLMM	NCHGN	0.260	−0.513	−0.927	−0.100	0.827	0.599	0.396	0.905

TABLE 4 Meta-analysis of diuretics in pre-eclampsia. Point estimates and confidence intervals for the between-study variance EM is the random-effects model, and GLMM is a generalized linear mixed model. L and U are the lower and upper limits of the 95% confidence intervals. Methods are DL with QP and BJ confidence intervals, J, MP with QP interval, REML with PL interval, and KD. The GLMM estimate using the NCHGN distribution is included for comparison.

Model	Method	$\hat{\tau}^2$	L	U	Length
REM	DL (QP)	0.230	0.072	2.202	2.130
REM	DL (BJ)	0.230	0.047	1.431	1.384
REM	J	0.329	0.074	1.678	1.604
REM	MP (QP)	0.386	0.072	2.202	2.130
REM	REML (PL)	0.300	0.043	1.475	1.432
REM	KD	0.392	0.087	1.962	1.875
GLMM	NCHGN	0.260			

8 | SUMMARY

Our extensive simulations demonstrate that the existing methods of meta-analysis of (log) odds ratio often present a biased view of both the heterogeneity and the overall effect. In brief: small sample sizes are rather problematic, and meta-analyses that involve numerous small studies are especially challenging. Because the study-level effects and their variances are related, estimates of the overall effects are biased, and the coverage of confidence intervals is too low, especially for small sample sizes and larger numbers of studies.

The between-study variance, τ^2, is typically estimated by generic methods which assume normality of the estimated effects $\hat{\theta}_i$. It is usually overestimated near zero, but the standard methods are negatively biased for larger values of τ^2. Our findings agree with those by van Aert et al.[4] that the standard interval estimators of τ^2 are often too liberal. The behavior of the profile-likelihood method is especially erratic.

Therefore, we proposed and studied a new method of estimating τ^2 based on the corrected approximation to the null distribution of Cochran's Q for log-odds-ratio developed by Kulinskaya and Dollinger[3]. The KD method provides reliable interval estimation of τ^2 across all values of τ^2, n, and K. Point estimation of τ^2 is more challenging; even though KD is better for $K = 30$, for small values of K it has positive bias and MP is better.

Arguably, the main purpose of a meta-analysis is to provide point and interval estimates of an overall effect.

Usually, after estimating the between-study variance τ^2, inverse-variance weights are used in estimating the overall effect and its variance. This approach relies on the theoretical result that, for known variances, and given unbiased estimates $\hat{\theta}_i$, it yields a uniformly minimum-variance unbiased estimate (UMVUE) of θ.

In practice, however, the true within-study variances are unknown, and use of the estimated variances makes the inverse-variance-weighted estimate of the overall effect biased. These biases (and even their sign) depend on τ^2 and the true value of θ, worsen for unbalanced studies, and may be considerable, even for reasonably large sample sizes such as $n = 250$. The coverage of the overall effect follows the same patterns because the centering of the confidence intervals is biased. Additionally, traditional intervals using normal quantiles are too narrow; and the use of t-quantiles, as in the HKSJ method, brings noticeable though not sufficient improvement.

Our additional simulations showed that the MH method and the GLMMs also do not perform well for point or interval estimation of θ.

A pragmatic approach to unbiased estimation of θ uses weights that do not involve estimated variances of study-level estimates, for example, weights proportional to the study size n_i. Hedges and Olkin[38], Hunter and Schmidt[39], and Shuster[49], among others, have proposed such weights. We use weights proportional to an effective sample size, $\tilde{n}_i = n_{iT}n_{iC}/n_i$; these are equivalent to the optimal inverse-variance weights for LOR when all the probabilities are equal. Importantly, because inverse-variance-weighted estimators have considerable biases, little, if any efficiency is lost by using the sample-size-based weights.

A reasonable estimator of τ^2, such as MP or KD, can be used as $\hat{\tau}^2$. Further, confidence intervals for θ centered at $\hat{\theta}_{SSW}$ with $\hat{\tau}^2_{KD}$ in Equation (5.2) can be used. In our simulations, this is by far the best interval estimator of θ, providing near-nominal coverage under all studied conditions.

HIGHLIGHTS

What is already known?

- In combining estimates from studies that had a binary individual-level outcome, the most common methods of meta-analysis use a weighted average of the studies' odds ratios (on the logarithmic scale), under a random-effects model; but their required estimators of the between-study or heterogeneity variance suffer from bias and below-nominal coverage, and produce bias and undercoverage in estimates of the overall log-odds-ratio.

What is new?

- Our extensive simulations confirm that the usual methods of meta-analysis produce biased estimates of the overall effect and confidence intervals whose coverage is too low. Estimates of heterogeneity variance have similar shortcomings. Small sample sizes are rather problematic, and meta-analyses that involve numerous small studies are especially challenging.
- For estimating between-study variance, a new method (KD), based on an improved approximation to the null distribution of Cochran's Q, provides reliable interval estimates. The KD point estimator is inferior to another estimator (Mandel-Paule) when the number of studies is small, but is better otherwise.
- A new, pragmatic point estimator of the overall effect (SSW) uses a weighted average in which a study's weight is proportional to its effective sample size. It has less bias than the popular inverse-variance-weighted estimators and three estimators obtained from generalized linear mixed models.
- The best interval estimator of the overall log-odds-ratio is centered on SSW and bases its endpoints on a t distribution and the KD point estimator of the between-study variance.

ential impact for RSM readers outside the authors' field

- The methods in common use for random-effects meta-analysis of odds ratios can advantageously be replaced by the new estimators, which have better performance.
- Meta-analysis software should include the new estimators.

FUNDING

The work by E. Kulinskaya was supported by the Economic and Social Research Council [grant number ES/L011859/1].

DATA AVAILABILITY STATEMENT

Our full simulation results are available as an arXiv e-print (Bakbergenuly et al.[44])

References

1. Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. *Methods for Meta-analysis in Medical Research*. John Wiley & Sons, Ltd . 2000.

2. Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society, Series A* 2009; 172: 137–159.

3. Kulinskaya E, Dollinger MB. An accurate test for homogeneity of odds ratios based on Cochran's Q-statistic. *BMC Medical Research Methodology* 2015; 15(1): 49.

4. van Aert RC, van Assen MA, Viechtbauer W. Statistical properties of methods based on the Q-statistic for constructing a confidence interval for the between-study variance in meta-analysis. *Research Synthesis Methods* 2019; 0(ja). doi: 10.1002/jrsm.1336

5. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; 7(3): 177–188.

6. Mandel J, Paule RC. Interlaboratory evaluation of a material with unequal numbers of replicates. *Analytical Chemistry* 1970; 42(11): 1194–1197.

7. Jackson D. Confidence intervals for the between-study variance in random effects meta-analysis using generalised Cochran heterogeneity statistics. *Research Synthesis Methods* 2013; 4(3): 220–229.

8. Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. *Statistics in Medicine* 1996; 15(6): 619–629.

9. Viechtbauer W. Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine* 2007; 26(1): 37–52.

10. Biggerstaff BJ, Jackson D. The exact distribution of Cochran's heterogeneity statistic in one-way random effects meta-analysis. *Statistics in Medicine* 2008; 27(29): 6093–6110.

11. Gart JJ, Pettigrew HM, Thomas DG. The effect of bias, variance estimation, skewness and kurtosis of the empirical logit on weighted least squares analyses. *Biometrika* 1985; 72(1): 179–190.

12. Gart JJ, Zweifel JR. On the bias of various estimators of the logit and its variance with application to quantal bioassay. *Biometrika* 1967; 54: 181–187.

13. Bhaumik DK, Amatya A, Normand SLT, et al. Meta-Analysis of Rare Binary Adverse Event Data. *Journal of the American Statistical Association* 2012; 107(498): 555–567.

14. Jackson D, White IR. When should meta-analysis avoid making hidden normality assumptions?. *Biometrical Journal* 2018; 60: 1040–1058.

15. 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. *Statistics in Medicine* 2007; 26(1): 53–77.

16. Kuss O. Statistical methods for meta-analyses including information from studies without any eventsâĂŤadd nothing to nothing and succeed nevertheless. *Statistics in Medicine* 2015; 34(7): 1097–1116.

17. Cheng J, Pullenayegum E, Marshall JK, Iorio A, Thabane L. Impact of including or excluding both-armed zero-event studies on using standard meta-analysis methods for rare event outcome: a simulation study. *BMJ Open* 2016; 6(8): e010983.

18. Turner RM, Omar RZ, Yang M, Goldstein H, Thompson SG. A multilevel model framework for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine* 2000; 19(24): 3417–3432.

19. Stijnen T, Hamza TH, Özdemir P. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. *Statistics in Medicine* 2010; 29(29): 3046–3067.

20. Jackson D, Law M, Stijnen T, Viechtbauer W, White IR. A comparison of seven random-effects models for meta-analyses that estimate the summary odds ratio. *Statistics in Medicine* 2018; 37: 1059–1085.

21. Bakbergenuly I, Kulinskaya E. Meta-analysis of binary outcomes via generalized linear mixed models: a simulation study. *BMC Medical Research Methodology* 2018; 18(70).

22. Van Houwelingen HC, Zwinderman KH, Stijnen T. A bivariate approach to meta-analysis. *Statistics in Medicine* 1993; 12: 2273–2284.

23. Veroniki AA, Jackson D, Viechtbauer W, et al. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods* 2016; 7(1): 55–79.

24. Langan D, Higgins JP, Simmonds M. Comparative performance of heterogeneity variance estimators in meta-analysis: a review of simulation studies. *Research Synthesis Methods* 2017; 8(2): 181–198. doi: 10.1002/jrsm.1198

25. Langan D, Higgins JP, Jackson D, et al. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Research Synthesis Methods* 2019; 10(1): 83–98. doi: 10.1002/jrsm.1316

26. Viechtbauer W. Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics* 2005; 30(3): 261–293.

27. Kosmidis I, Guolo A, Varin C. Improving the accuracy of likelihood-based inference in meta-analysis and meta-regression. *Biometrika* 2017; 104(2): 489–496.

28. Paule RC, Mandel J. Consensus values and weighting factors. *Journal of Research of the National Bureau of Standards* 1982; 87(5): 377–385.

29. Carter G, Rolph J. Emprical Bayes Methods Applied to Estimating Fire Alarm Probabilities. *Journal of the American Statistical Association* 1974; 69: 880–885.

30. Morris C. Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association* 1983; 78: 47–55.

31. Rukhin A, Vangel M. Estimation of a common mean and weighted means statistics. *Journal of the American Statistical Association* 1998; 93: 303–308.

32. Rukhin A, Biggerstaff B, Vangel M. Restricted maximum likelihood estimation of a common mean and the Mandel–Paule algorithm. *Journal of Statistical Planning and Inference* 2000; 83: 319–330.

33. DerSimonian R, Kacker R. Random-effects model for meta-analysis of clinical trials: an update. *Contemporary Clinical Trials* 2007; 28(2): 105–114.

34. Kulinskaya E, Dollinger M, Bjørkestøl K. On the moments of Cochran's Q statistic under the null hypothesis, with application to the meta-analysis of risk difference. *Research Synthesis Methods* 2011; 2: 254-270. doi: 10.1002/jrsm.54

35. Jackson D, Bowden J. Confidence intervals for the between-study variance in random-effects meta-analysis using generalised heterogeneity statistics: should we use unequal tails?. *BMC Medical Research Methodology* 2016; 16(1): 118.

36. Viechtbauer W. Package *metafor*. *The Comprehensive R Archive Network* 2015.

37. Cochran WG. The combination of estimates from different experiments. *Biometrics* 1954; 10(1): 101–129.

38. Hedges LV, Olkin I. *Statistical Methods for Meta-Analysis*. San Diego, California: Academic Press . 1985.

39. Hunter JE, Schmidt FL. *Methods of Meta-analysis: Correcting Error and Bias in Research Findings*. Newbury Park, CA: Sage Publications, Inc. . 1990.

40. Greenland S, Robins JM. Estimation of a common effect parameter from sparse follow-up data. *Biometrics* 1985; 41: 55–68.

41. Hartung J, Knapp G. A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Statistics in Medicine* 2001; 20(24): 3875–3889.

42. Sidik K, Jonkman JN. A simple confidence interval for meta-analysis. *Statistics in Medicine* 2002; 21(21): 3153–3159.

43. Sánchez-Meca J, Marín-Martínez F. Testing the significance of a common risk difference in meta-analysis. *Computational Statistics & Data Analysis* 2000; 33(3): 299–313.

44. Bakbergenuly I, Hoaglin DC, Kulinskaya E. Simulation study of estimation of between-study variance and overall effect in meta-analysis of odds ratio. In: eprint arXiv:1902.07154v1 [stat.ME]. ; 2019.

45. Collins R, Yusuf S, Peto R. Overview of randomised trials of diuretics in pregnancy.. *British Medical Journal (Clinical Research Edition)* 1985; 290(6461): 17–23.

46. Biggerstaff BJ, Tweedie RL. Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in Medicine* 1997; 16(7): 753–768.

47. Kulinskaya E, Olkin I. An overdispersion model in meta-analysis. *Statistical Modelling* 2014; 14(1): 49–76.

48. Bakbergenuly I, Kulinskaya E. Beta-binomial model for meta-analysis of odds ratios. *Statistics in Medicine* 2017; 36(11): 1715–1734.

49. Shuster JJ. Empirical vs natural weighting in random effects meta-analysis. *Statistics in Medicine* 2010; 29(12): 1259–1265. doi: 10.1002/sim.3607

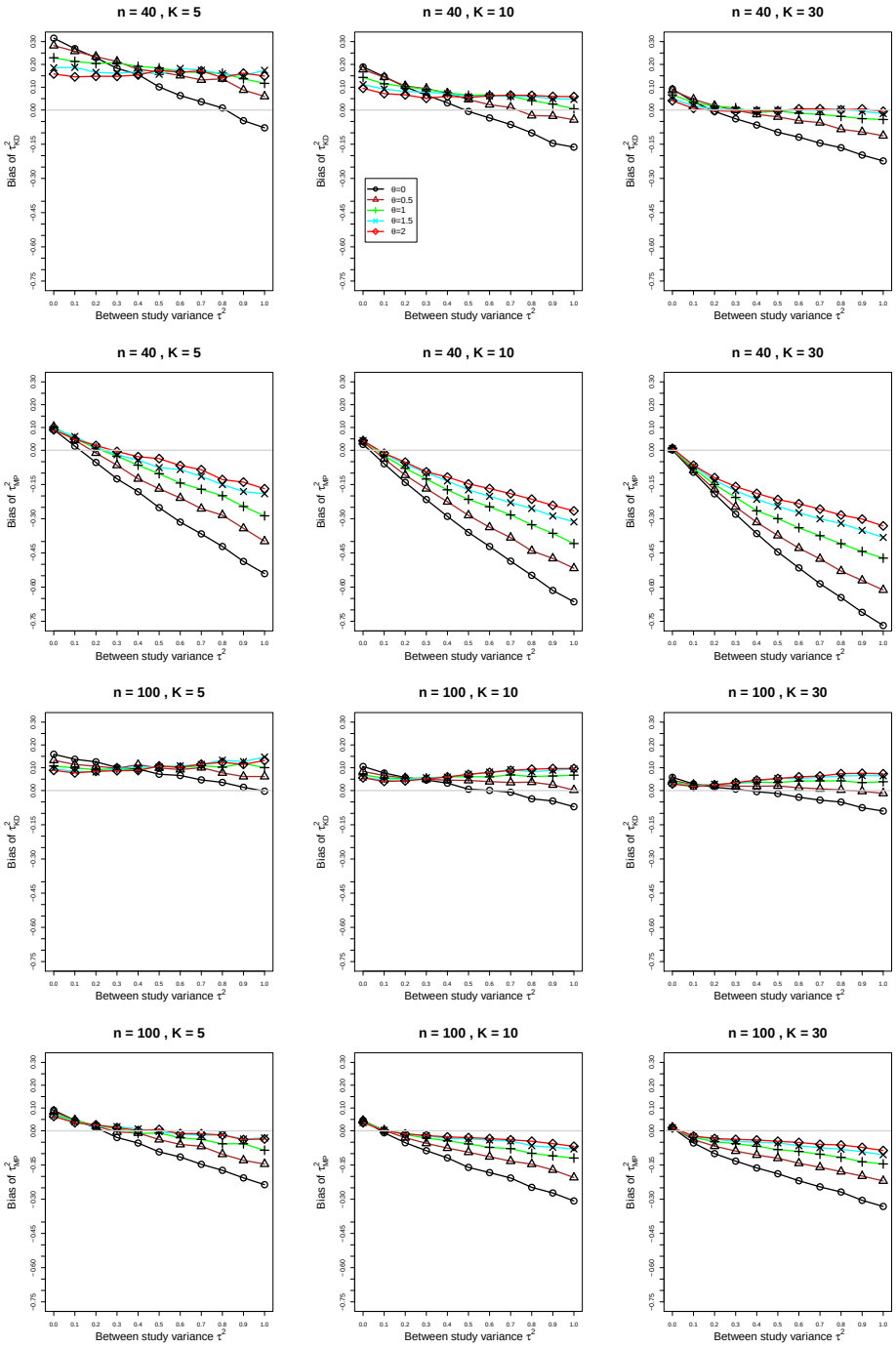

FIGURE 1 Bias of $\hat{\tau}^2_{KD}$ and $\hat{\tau}^2_{MP}$ in estimating the between-study variance τ^2 for $\theta = 0(0.5)2$, $p_{iC} = 0.1$, $q = .5$, $n = 40,\ 100$. The symbols for the values of θ are $\theta = 0$, black $\circ$; $\theta = 0.5$, brown $\triangle$; $\theta = 1$, green $+$; $\theta = 1.5$, blue $\times$; and $\theta = 2$, red $\diamond$. Light grey line at 0.

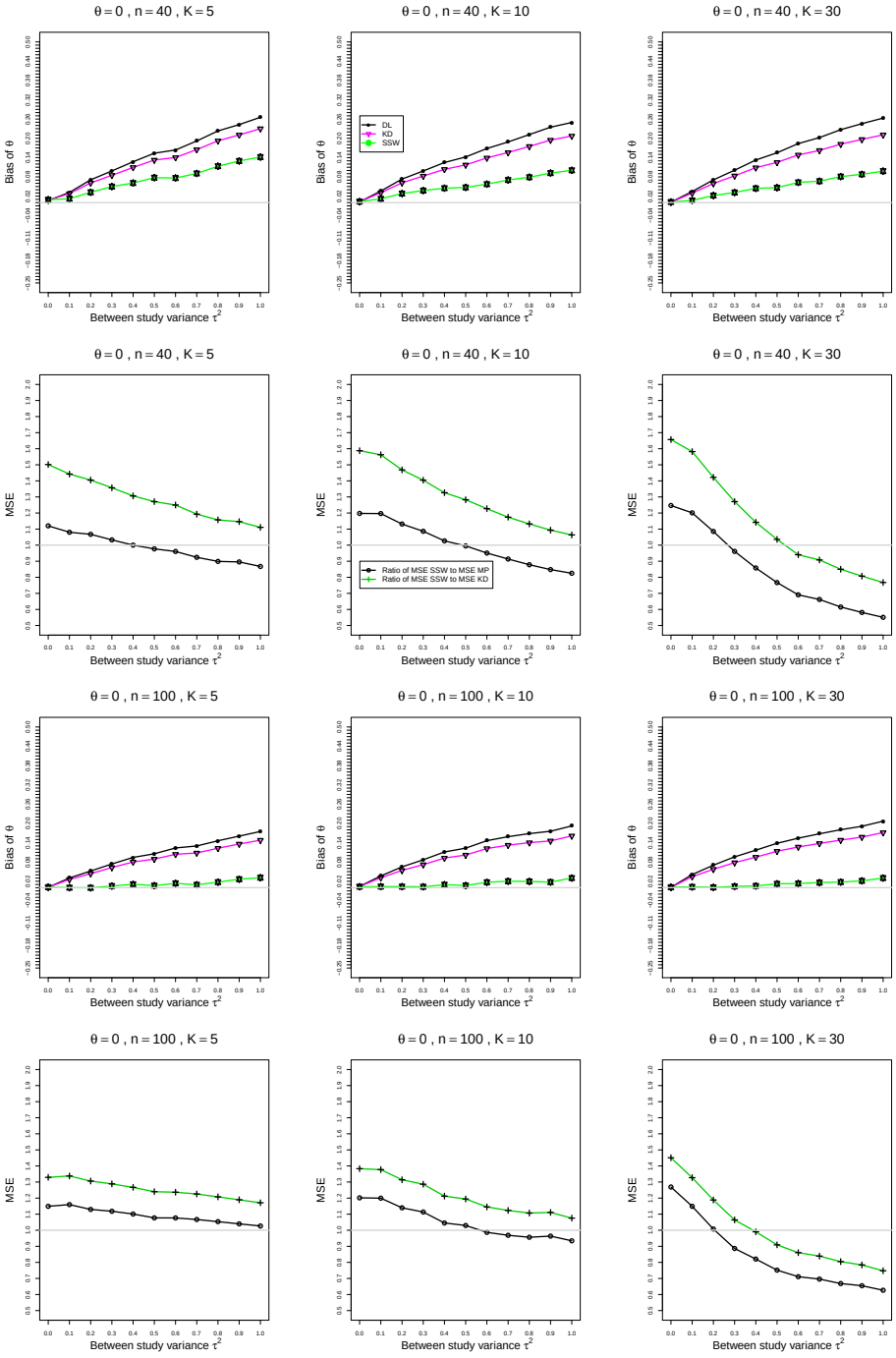

FIGURE 2 Bias and ratio of MSEs for estimators of the overall effect θ for $\theta = 0$, $p_{iC} = .1$, $q = .5$, and equal sample sizes $n = 40,\ 100$. Light grey line at 0 and 1, respectively.

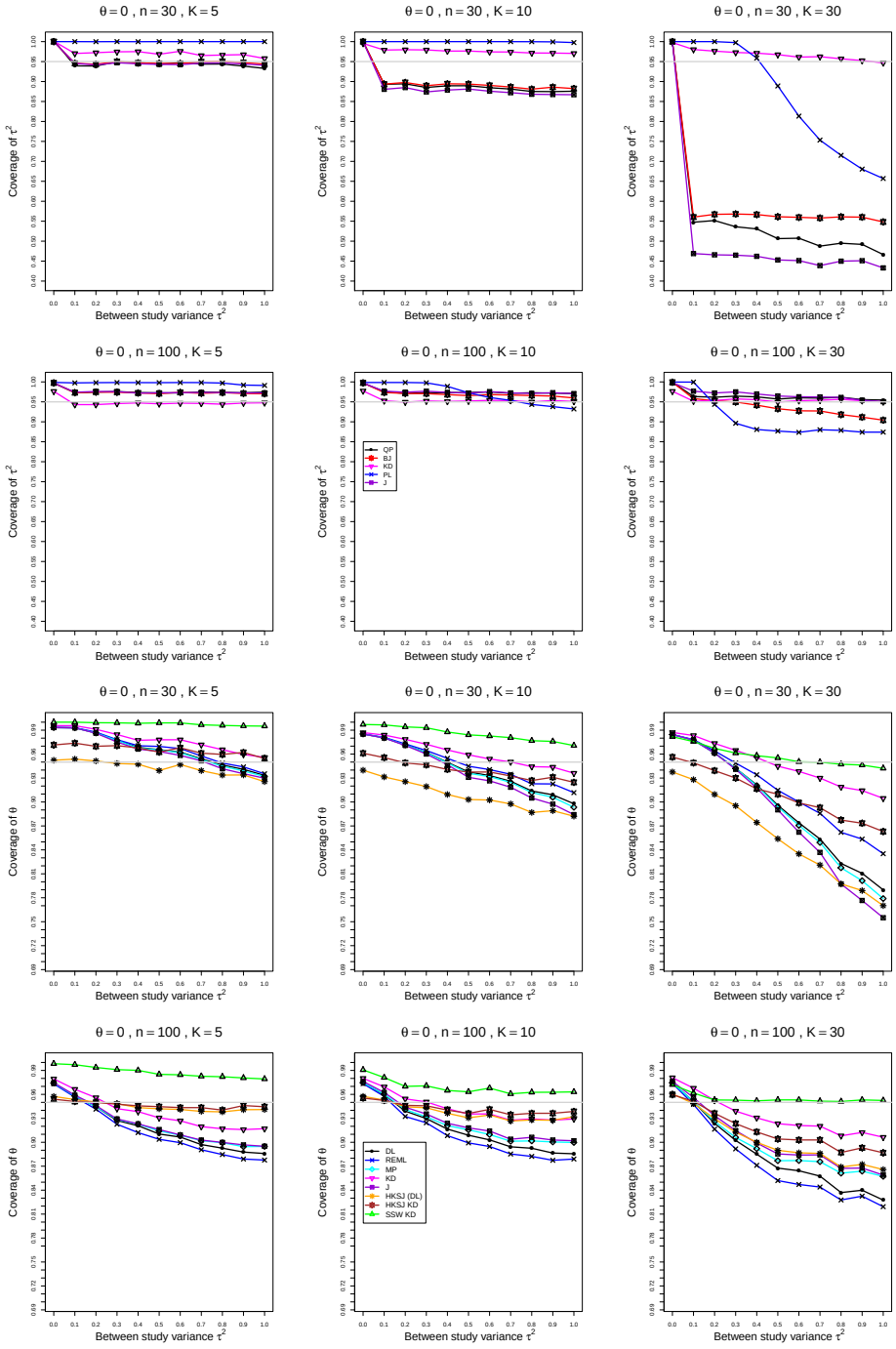

FIGURE 3 Coverage of between-studies variance τ^2 (top two rows) and overall effect θ (bottom two rows) for $\theta = 0$, $p_{iC} = .1$, $q = .5$, and unequal sample sizes $\bar{n} = 30,\ 100$. Light grey line at 0.95.

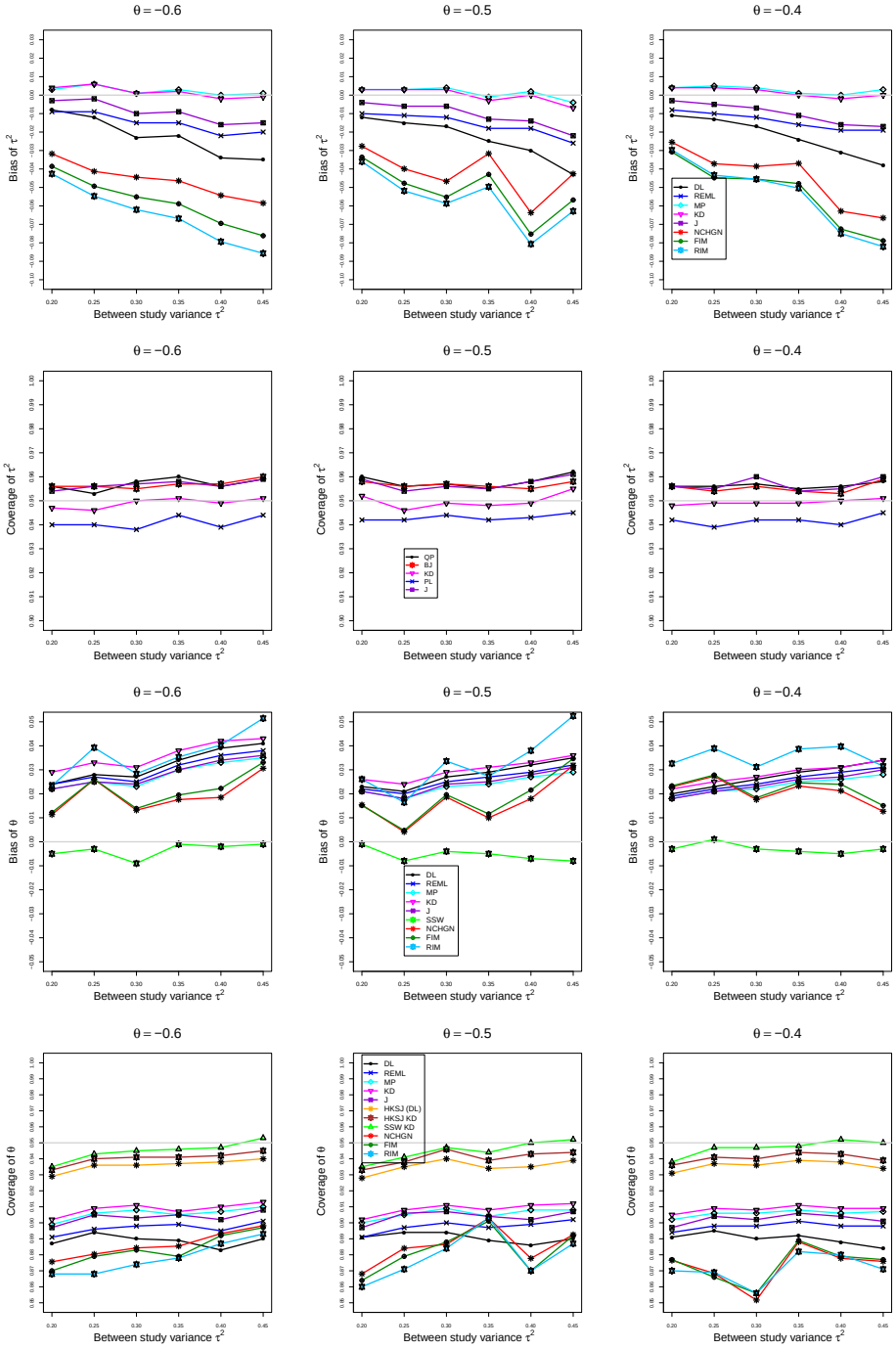

FIGURE 4 Bias and coverage of estimators of the between-study variance τ^2 and of the LOR θ for the sample sizes and the $\hat{p}_{iC}$ in the pre-eclampsia data of Collins et al.[45], $\theta = -0.6,\ -0.5,\ -0.4$, and $\tau^2 = 0.20(0.05)0.45$.