Statistics and Its Interface Volume 13 (2020) 449–464

Statistical methods for quantifying between-study heterogeneity in meta-analysis with focus on rare binary events*

Chiyu Zhang, Min Chen, and Xinlei Wang†

Meta-analysis, the statistical procedure for combining results from multiple independent studies, has been widely used in medical research to evaluate intervention efficacy and drug safety. In many practical situations, treatment effects vary notably among the collected studies, and the variation, often modeled by the between-study variance parameter τ^2, can greatly affect the inference of the overall effect size. In the past, comparative studies have been conducted for both point and interval estimation of τ^2. However, most are incomplete, only including a limited subset of existing methods, and some are outdated. Further, none of the studies covers descriptive measures for assessing the level of heterogeneity. Nor are they focused on rare binary events that require special attention. We summarize by far the most comprehensive set including 11 descriptive measures, 23 estimators, and 16 confidence intervals. In addition to providing synthesized information, we further categorize these methods according to their key features. We then evaluate their performance based on simulation studies that examine various realistic scenarios for rare binary events, with an illustration using a data example of a gestational diabetes meta-analysis. We conclude that there is no uniformly "best" method. However, methods with consistently better performance do exist in the context of rare binary events, and we provide practical guidelines based on numerical evidences.

Keywords and phrases: Bias, Confidence interval, Coverage probability, DerSimonian and Laird, Fixed effect, Odds ratio, Mean squared error, Q statistic, Random effects.

1. INTRODUCTION

Meta-analysis, the statistical procedure for synthesizing information from multiple studies, has been widely used in many research areas including social, psychological and especially medical sciences. Meta-analysis is a powerful tool in drug safety evaluation, where the number of cases (adverse events) can be very limited in a single study. The U.S. Food and Drug Administration (FDA) released a draft guidance for industry titled "Meta-Analyses of Randomized Controlled Clinical Trials to Evaluate the Safety of Human Drugs or Biological Products" in November 2018, which demonstrates the importance of meta-analysis in the development of new drugs. Such meta-analysis often involves binary outcomes of rare events, which are the focus of this study.

The primary goal of a meta-analysis is usually to estimate and infer the overall effect size, where the variability in the effect estimates from component studies should be properly accounted for. Besides the within-study sampling errors, the variability may come from diverse characteristics of individual studies such as disparities in trial protocols, subjects' conditions, and population features, etc. When the study-wise differences exist, we call these studies (statistically) heterogeneous and the heterogeneity is typically measured by a between-study variance parameter τ^2. Also, descriptive measures have been widely used by clinicians to provide a more intuitive interpretation about the heterogeneity for ease of understanding.

For point estimation of τ^2, the DerSimonian and Laird (DL) estimator [10], one of the most widely used methods, has been frequently challenged for its default use in many software packages, largely due to its sizable negative bias when the heterogeneity level is high [2, 34, 35, 39, 47]. Many modifications over the DL estimator have been suggested based on the method of moments. Other approaches such as likelihood-based and other nonparametric methods can also be applied. For interval estimation of τ^2, different types of confidence intervals (CIs) have been constructed to gauge the estimation uncertainty. However, nearly all these methods were constructed without a special consideration of dichotomous data and their performance remains unclear in the context of *rare* binary events, in which some may produce large bias or even fail to work.

Comparative studies and review papers for descriptive measures, unlike for the point and interval estimation of τ^2, are scarce. For example, Veroniki et al. [46], Langan, Higgins and Simmonds [28], Petropoulou and Mavridis [37] reviewed and compared most of the existing estimators of τ^2,

*Wang's research was partially supported by NIH Grant R15GM131390.
†Corresponding author. ORCID: 0000-0002-8561-6511.

among which only Petropoulou and Mavridis [37] conducted simulation studies to evaluate their performance. Previous comparisons about CIs (e.g., [25, 45, 48]) were largely limited to several similar types of CIs. As detailed in Tables 3 and 5, none of these papers cover descriptive measures for quantifying the level of heterogeneity. Nor do they focus on rare binary events. Most of them are far from being complete, some even outdated, which motivates us to conduct this study to provide useful guidance to clinicians and biostatisticians.

The paper is organized as follows. In Section 2, we introduce notation and frequently used terms in meta-analysis. Section 3 reviews existing descriptive measures quantifying the level of heterogeneity. In Section 4, we list estimators for τ^2 and briefly summarize two recently developed ones that are not included in any of the existing review papers. In Section 5, different types of confidence intervals for τ^2 are described and categorized. In Section 6, we compare the performance, in terms of bias and mean squared error (MSE) for point estimators and empirical coverage probability and width for CIs, in a large collection of scenarios that are designed to mimic practical situations. In Section 7, we reanalyze the data from a meta-analysis of 20 trials of type 2 diabetes mellitus after gestational diabetes with focus on the heterogeneity among the component studies[1]. The final section provides recommendations in terms of choosing appropriate estimators and CIs in meta-analysis of rare binary events as well as a brief discussion.

2. NOTATION & FREQUENTLY USED TERMS

Suppose a meta-analysis includes K independent studies and the kth study contains n_k subjects ($k = 1, ..., K$). In study k, let θ_k be the true but unknown treatment effect and y_k be the observed treatment effect such that $\mathrm{E}[y_k|\theta_k] = \theta_k$ and $\mathrm{Var}[y_k|\theta_k] = \sigma_k^2$, the within-study variance. Typically s_k^2, an estimate of σ_k^2, is reported along with y_k in published studies and it is often treated as a known quantity in practice (i.e., indistinguishable from σ_k^2). When the study-specific effects θ_k's are treated as random variables rather than constants, we assume $\mathrm{E}[\theta_k] = \theta$ and $\mathrm{Var}[\theta_k] = \tau^2$, where θ, a parameter of main interest in the meta-analysis, represents the overall treatment effect across different studies, and τ^2 measures the between-study heterogeneity. There exist two main parametric models, namely *Re* (random effects) and *Fe* (fixed effect), to combine results from component studies. The *Re* model assumes that $y_k = \theta_k + \epsilon_k$, where $\theta_k \sim N(\theta, \tau^2)$ and $\epsilon_k \sim N(0, \sigma_k^2)$. When $\tau^2 = 0$, it is reduced to the *Fe* model $y_k = \theta + \epsilon_k$, where a common treatment effect θ is assumed for all component studies (i.e., $\theta_k \equiv \theta$). These models can be used with any effect measure, as long as the assumed normality is (approximately) valid.

For binary responses, we denote the number of events by $x_{k0}\,(x_{k1})$ and the number of subjects by $n_{k0}\,(n_{k1})$ in the control (treatment) group. The probability of having an event in the control (treatment) group is denoted by $p_{k0}\,(p_{k1})$. Effect measures for binary outcomes include risk difference (RD, $p_{k1}-p_{k0}$), risk ratio (RR, p_{k1}/p_{k0}) and odds ratio (OR, $[p_{k1}/(1-p_{k1})]/[p_{k0}/(1-p_{k0})]$). For rare binary events, RR≈OR. A logarithm transformation of the odds ratio (LOR) is often used in meta-analysis for a much faster convergence to asymptotic normality, and the within-study variance σ_k^2 is then estimated by $s_k^2 = \frac{1}{x_{k0}} + \frac{1}{n_{k0}-x_{k0}} + \frac{1}{x_{k1}} + \frac{1}{n_{k1}-x_{k1}}$. Gart [13] added a continuity correction factor of 0.5 to all the cells so that

$$y_k = \log \frac{x_{k1}+0.5}{n_{k1}-x_{k1}+0.5} - \log \frac{x_{k0}+0.5}{n_{k0}-x_{k0}+0.5},$$

and σ_k^2 is estimated by

$$s_k^2 = \frac{1}{x_{k0}+0.5} + \frac{1}{n_{k0}-x_{k0}+0.5} + \frac{1}{x_{k1}+0.5} + \frac{1}{n_{k1}-x_{k1}+0.5},$$

which will be used in our numerical evaluation of rare binary events.

Next, we introduce the (generalized) Q statistic [9] and related terms, which will frequently appear in the paper. For any parameter of interest, we use the corresponding letter/symbol with a hat to denote its estimate. For example, we use $\hat{\theta}$ to denote the estimate of the overall treatment effect θ. The Q statistic is defined as the weighted sum of squared deviations between the estimated overall treatment effect and observed treatment effect in each individual study, namely

$$Q = \sum_{k=1}^{K} w_k (y_k - \hat{\theta})^2, \tag{1}$$

where w_k is a positive weight assigned to study k, and $\hat{\theta} = \sum_{k=1}^{K} w_k y_k / \sum_{k=1}^{K} w_k$, the weighted average of the estimated study-specific effects. A commonly used weighting scheme is to set $w_k = [\widehat{\mathrm{Var}}(y_k)]^{-1}$, i.e., the inverse of the estimated variance of y_k. Under this inverse-variance weighing scheme, the variance of $\hat{\theta}$ can be given by $1/\sum_{k=1}^{K} w_k$ if we treat w_k's as known constants (i.e., indistinguishable from $[\mathrm{Var}(y_k)]^{-1}$). Further, this scheme yields $w_k = 1/s_k^2$ for the *Fe* model, and $w_k = 1/(s_k^2 + \hat{\tau}^2)$ for the *Re* model, where $\hat{\tau}^2$ can be any estimator discussed in Section 4. Under the *Fe* (*Re*) model with the inverse-variance weights, we denote the corresponding Q statistic by Q_{Fe} (Q_{Re}) and the corresponding $\hat{\theta}$ by $\hat{\theta}_{Fe}$ ($\hat{\theta}_{Re}$) with variance v_{Fe} (v_{Re}). Note that Q_{Fe} is also known as the DerSimonian and Laird's Q test statistic [10].

Throughout this paper, we use χ^2_{df} to denote a chi-squared distribution with df degrees of freedom, and use $\chi^2_{df,\alpha}$ to denote its 100α-th percentile.

3. DESCRIPTIVE MEASURES QUANTIFYING BETWEEN-STUDY HETEROGENEITY

As mentioned in the introduction, (statistical) heterogeneity exists when true effects being evaluated differ among studies in a meta-analysis. Assessing the extent of heterogeneity is essential for validating model assumptions and decision making. An obvious choice is by estimating the variance parameter τ^2, as is typically done in a random-effects meta-analysis. As pointed out by Higgins and Thompson [19], this measure does not facilitate comparison of heterogeneity across meta-analyses of different types of outcomes (e.g., the survival time can be either continuous or discrete). Also, its scale is specific to a chosen effect metric and the interpretation can be difficult. For example, odds ratio is a commonly used effect measure for binary data. Still, the variance of log-odds ratio is not easy to understand for many non-statisticians. Alternatively, one may test the presence of the between-study heterogeneity and use the corresponding test statistic or p-value to indicate the extent of heterogeneity. However, such measures depend on the scale of effect sizes or the number of component studies K. To overcome these limitations, effort has been devoted to development of various descriptive measures that can provide more intuitive information about the heterogeneity.

Table 1 summarizes 11 descriptive heterogeneity measures in the literature. Note that all these measures are general-purpose and none is specifically designed for binary outcomes. Takkouche, Cadarso-Suarez and Spiegelman [42] proposed two measures, R_I and CV_B, to quantify the level of heterogeneity in five published meta-analyses. The statistic R_I was developed to estimate $\tau^2/(\tau^2+\sigma^2)$, the proportion of total variation in the effect estimates that is due to between-study heterogeneity. This quantity is also known as the intra-class correlation in the context of cluster sampling. Here, the within-study variances σ_k^2's are assumed to be constant, i.e., $\sigma_k^2 \equiv \sigma^2$, which is estimated by $1/\sum_{k=1}^K 1/s_k^2$, making $R_I = \frac{\hat\tau^2}{\hat\tau^2+K/\sum_{k=1}^K 1/s_k^2}$. The other statistic CV_B estimates the between-study coefficient of variation $\tau/|\theta|$ by $\sqrt{\hat\tau^2}/|\hat\theta|$. Obviously, CV_B is affected by the overall treatment effect θ and is undefined when $\theta = 0$.

Under the assumption of a common within-study variance σ^2, Higgins and Thompson [19] formulated a general heterogeneity measure as a function of the overall treatment effect θ, the between-study variance τ^2, the within-study variance σ^2, and the number of component studies, namely, $f(\theta, \tau^2, \sigma^2, K)$. They proposed three criteria that such a measure should satisfy in general in order to facilitate its comparability and interpretability, including (i) dependence on the extent of heterogeneity, (ii) scale invariance, i.e. $f(\theta, \tau^2, \sigma^2, K) = f(a+b\theta, b^2\tau^2, b^2\sigma^2, K)$ for any a and b, and (iii) size invariance, i.e. $f(\theta, \tau^2, \sigma^2, K_1) = f(\theta, \tau^2, \sigma^2, K_2)$ for any positive integers K_1 and K_2. Criterion (i) implies that the function f should increase monotonically with τ^2. Criterion (ii) implies that f should be a function of the ratio $\rho \equiv \frac{\tau^2}{\sigma^2}$ and that θ should not be involved. Criterion (iii) implies that f does not depend on K. It can be shown that any monotonically increasing function of ρ satisfies the three criteria. Based on this, three statistics, H^2, R^2 and I^2 were proposed. The first, H^2, estimates the quantity $\rho + 1$ by equating the observed value of Q_{Fe} to its expectation so that $H^2 = \frac{Q_{Fe}}{K-1}$ can be interpreted as relative excess in Q_{Fe} over its expected value, the degrees of freedom $K-1$. The second, R^2, attempts to estimate $\rho + 1$ as well; but here, $\rho + 1$ is approximated by v_{Re}/v_{Fe} so that $R^2 = \hat v_{Re}/\hat v_{Fe} = \sum_{k=1}^K \frac{1}{s_k^2}/\sum_{k=1}^K \frac{1}{s_k^2+\hat\tau^2}$, which can be interpreted as the inflation in the confidence interval for $\hat\theta_{Re}$ under the Re model compared with $\hat\theta_{Fe}$ under the Fe model. Both H^2 and R^2 should be at least 1, where 1 means perfect homogeneity; and the larger the value, the more heterogeneous the studies. In practice, the authors suggested to use H and R because clinicians may be more familiar with standard deviations than variances. The third statistic, I^2, estimates a different function of ρ, i.e. $\frac{\rho}{1+\rho} = \frac{\tau^2}{\tau^2+\sigma^2}$, which represents the proportion of total variance that is due to between-study variation. Higgins and Thompson [19] suggested to compute I^2 by $I^2_{HT} = 1 - \frac{K-1}{Q_{Fe}}$, which leads to a convenient relationship $I^2_{HT} = 1 - \frac{1}{H^2}$. Jackson, White and Riley [24] suggested to compute I^2 by $I^2_R = 1 - \frac{\hat v_{Fe}}{\hat v_{Re}} = 1 - \sum_{k=1}^K \frac{1}{s_k^2+\hat\tau^2}/\sum_{k=1}^K \frac{1}{s_k^2}$, which leads to another convenient relationship $I^2_R = 1 - \frac{1}{R^2}$. Both I^2_{HT} and I^2_R are usually expressed as percentages between 0% and 100%, where a value of 0% corresponds to no observed heterogeneity, while larger values indicate increasing levels of heterogeneity. They estimate the same quantity as R_I does, but with different within-study variance estimates. Among these measures (i.e. H^2, R^2, I^2_{HT} or I^2_R), I^2_{HT} is most popular and in the literature, I^2 typically represents I^2_{HT} as I^2_R is much less known. Higgins and Green [18] empirically provided a rough guide to the interpretation of I^2 using overlapping intervals: a value in [0,0.4] suggests that heterogeneity may not be that important; [0.3, 0.6] may represent moderate heterogeneity; [0.5,0.9] may represent substantial heterogeneity; and [0.75,1] implies considerable heterogeneity.

The assumption of a constant within-study variance is probably untrue in many real life data. Thus, Crippa et al. [8] lifted this assumption and proposed a new measure R_b, defined as $R_b = \frac{1}{K}\sum_{k=1}^K \frac{\hat\tau^2}{s_k^2+\hat\tau^2}$, to assess the contribution of the between-study variance τ^2 to v_{Re} (i.e., the variance of the pooled random-effects estimate $\hat\theta_{Re}$). It can be viewed as an average of the study-specific proportions of the study-specific variances due to between-study heterogeneity. They showed that the quantity τ^2/v_{Re} underlying R_b is a strictly increasing function of τ^2 and is scale-invariant. However, this quantity depends on K and so is not size-invariant. They further showed that $R_I \geq \max(R_b, I^2_{HT})$.

Table 1. Descriptive measures quantifying the between-study heterogeneity

Name	$f(\theta,\tau^2,\sigma^2,K)$	Formula	Ref.	Interpretation	Assume $\sigma_k^2 \equiv \sigma^2$?
R_I	$\frac{\tau^2}{\tau^2+\sigma^2}$	$\frac{\hat{\tau}^2}{\hat{\tau}^2+K/\sum_{k=1}^K s_k^{-2}}$	[42]	Proportion of total variation in the estimates of treatment effect due to between-study heterogeneity	Yes
CV_B	$\frac{\tau}{\lvert\theta\rvert}$	$\frac{\sqrt{\hat{\tau}^2}}{\lvert\hat{\theta}\rvert}$	[42]	Between-study coefficient of variation	No
H^2	$\frac{\tau^2+\sigma^2}{\sigma^2}$	$\frac{Q_{Fe}}{K-1}$	[19]	Relative excess in Q_{Fe} over its degrees of freedom	Yes, but can be used for different σ_k^2.
R^2	$\frac{\tau^2+\sigma^2}{\sigma^2} \approx \frac{v_{Re}}{v_{Fe}}$	$\frac{\sum_{k=1}^K s_k^{-2}}{\sum_{k=1}^K (s_k^2+\hat{\tau}^2)^{-1}}$	[19]	Inflation in the confidence interval for a single summary estimate under Re model compared with Fe model	Yes, but can be used for different σ_k^2.
I^2_{HT}	$\frac{\tau^2}{\tau^2+\sigma^2}$	$1-\frac{K-1}{Q_{Fe}}$	[19]	Same as R_I	Yes
I^2_R	$\frac{\tau^2}{\tau^2+\sigma^2}$	$1-\frac{\sum_{k=1}^K (s_k^2+\hat{\tau}^2)^{-1}}{\sum_{k=1}^K s_k^{-2}}$	[24]	Same as R_I	Yes
R_b	$\frac{\tau^2}{v_{Re}} \approx \frac{1}{K}\sum_{k=1}^K \frac{\tau^2}{\sigma_k^2+\tau^2}$	$\frac{1}{K}\sum_{k=1}^K \frac{\hat{\tau}^2}{s_k^2+\hat{\tau}^2}$	[8]	Proportion of the between-study heterogeneity τ^2 relative to v_{Re}, the variance of $\hat{\theta}_{Re}$.	No
H^2_r	$\frac{\tau^2+\sigma^2}{\sigma^2}$	$\frac{\pi Q_r^2}{2K(K-1)}$, $Q_r=\sum_{k=1}^K \frac{1}{s_k}\lvert y_k-\hat{\theta}_{Fe}\rvert$	[31]	Same as H^2	Yes
I^2_r	$\frac{\tau^2}{\tau^2+\sigma^2}$	$1-\frac{2K(K-1)}{\pi Q_r^2}$	[31]	Same as R_I	Yes
H^2_m	$\frac{\tau^2+\sigma^2}{\sigma^2}$	$\frac{\pi Q_m^2}{2K^2}$, $Q_m=\sum_{k=1}^K \frac{1}{s_k}\lvert y_k-\hat{\theta}_m\rvert$, $\hat{\theta}_m$ is weighted median estimate	[31]	Same as H^2	Yes
I^2_m	$\frac{\tau^2}{\tau^2+\sigma^2}$	$\frac{Q_m^2-2K^2/\pi}{Q_m^2}$	[31]	Same as R_I	Yes

When $\sigma_k^2 \equiv \sigma^2$ and σ^2 is estimated by s^2, R_b, R_I, I^2_{HT} and I^2_R all yield the same quantity $\frac{\hat{\tau}^2}{s^2+\hat{\tau}^2}$. The authors conducted a simulation study to examine the performance of R_I, I^2_{HT} and R_b. Both R_I and I^2_{HT} tend to be positively biased and this overestimation increases as K increases. Confidence intervals based on R_I and I^2_{HT} give lower coverage probabilities compared to those based on R_b and the difference becomes more obvious when the within-study variances vary more and when the heterogeneity level increases.

To reduce the impact of outlying studies, Lin, Chu and Hodges [31] proposed new robust measures H^2_r, H^2_m, I^2_r and I^2_m, which are analogous to and have the same interpretations as H^2 and I^2, respectively. These methods were developed upon the absolute deviation measures Q_r and Q_m rather than the usual squared deviation measure Q, as defined in Table 1 and will be described in more detail in Section 4.

All the measures except for CV_B depend on the precision of the study-specific effects. As the sample sizes of the component studies increase, σ_k^2's would decrease to zero so that R_I, R_B and all I^2's would increase to 1 and all H^2's and R^2 would become arbitrarily large, even when there is little between-study heterogeneity. The measure CV_B avoids this drawback but has its own limitation: it would approach $+\infty$ as θ goes to 0. Finally, we mention that some of the measures involve the estimated value $\hat{\tau}^2$. In principle, $\hat{\tau}^2$ can be any estimator of τ^2, but most software uses the DL estimator $\hat{\tau}^2_{DL}$ as the default choice.

4. ESTIMATORS

We summarize 23 estimators for τ^2 in Table 2, among which most can be applied to all kinds of effect measures except for the improved Paule and Mandel estimator (IPM, [2]) and Malzahn, Böhning, and Holling (MBH, [32]). IPM is specifically designed to work with OR for binary outcomes, and MBH can be only used for standardized mean difference (SMD). All estimators can be divided into five groups: method of moments, likelihood-based, model error variance (least squares), Bayes, and other nonparametric estimators. Some have closed form expressions while the others require numerical solutions. Some produce only positive estimates while the others require truncation to zero when a negative value occurs. Some properties of the estimators are summarized in Table 2.

Table 3 shows previous studies that reviewed and compared (large) subsets of these estimators. Recommendations were made either based on their own simulations or conclusions from the literature. Among them, Veroniki et al. [46], Langan, Higgins and Simmonds [28] and Petropoulou and Mavridis [37] are the most comprehensive. Veroniki et al. [46] reviewed 17 estimators as listed in Table 3, including all the method of moments estimators except for the IPM, multistep DL and LCH estimators, all three likelihood-based estimators, the SJ estimator, all the Bayesian estimators, and DL_b. Langan, Higgins and Simmonds [28] and Petropoulou and Mavridis [37] added IPM, MBH, and SJ_{HO} into the comparison. Note that IPM was briefly summarized but not compared with other estimators in Veroniki et al. [46]. Also, EB mentioned in [37] has been shown to be equivalent to PM. Langan, Higgins and Simmonds [28] also added RB estimators with different priors, RB_u and RB_a.

Two newly proposed estimators, the LCH estimators [31] and the multistep DL estimator DL_M [44], are included in our pool. We mark them in bold in Table 2 and provide a brief description for each in below. The IPM estimator [2] is described as well because it is the only method specifically designed for rare binary events. More details about other estimators can be found in [46] and references therein.

Lin, Chu and Hodges (LCH) Lin, Chu and Hodges [31] proposed two alternative estimators, $\hat{\tau}^2_r$ and $\hat{\tau}^2_m$, designed to be

Table 2. Overview of 23 estimators for the between-study variance τ^2

Estimators	Abbreviation	Reference	Iterative?	Sign	Effect Measure
Method of Moments					
Hedges and Olkin	HO	[17]	No	$>= 0$	
Two-step Hedges and Olkin	HO_2	DerSimonian and Kacker [9]	No	$>= 0$	
DerSimonian and Laird	DL	DerSimonian and Laird [10]	No	$>= 0$	
Positive DerSimonian and Laird	DL_p	Kontopantelis, Springate and Reeves [27]	No	> 0	
Two-step DerSimonian and Laird	DL_2	DerSimonian and Kacker [9]	No	$>= 0$	
Multistep Dersimonian and Laird	DL_M	van Aert and Jackson [44]	No	$>= 0$	
Paule and Mandel	PM	Paule and Mandel [36]	Yes	$>= 0$	
Improved Paule and Mandel	IPM	Bhaumik et al. [2]	Yes	$>= 0$	OR
Hartung and Makambi	HM	Hartung and Makambi [16]	No	> 0	
Hunter and Schmidt	HS	Hunter and Schmidt [20]	No	$>= 0$	
Lin, Chu and Hodges	LCH	Lin, Chu and Hodges [31]	No	$>= 0$	
Likelihood-based					
Maximum Likelihood	ML	Hardy and Thompson [14]	Yes	$>= 0$	
Restricted maximum likelihood	$REML$	Viechtbauer [47]	Yes	$>= 0$	
Approximate restricted maximum likelihood	$AREML$	Morris [33]	Yes	$>= 0$	
Model error variance (Least squares)					
Sidik and Jonkman	SJ	Sidik and Jonkman [39]	No	> 0	
Sidik and Jonkman (HO prior)	SJ_{HO}	Sidik and Jonkman [40]	No	> 0	
Bayesian					
Rukhin Bayes	RB_0	Rukhin [38]	Yes	$>= 0$	
Positive Rukhin Bayes	RB_p	Rukhin [38]	Yes	> 0	
Empirical Bayes (Equivalent to PM)	EB	Morris [33]	Yes	$>= 0$	
Fully Bayes	FB	Smith, Spiegelhalter and Thomas [41]	Yes	> 0	
Bayes Modal	BM	Chung et al. [6], Chung, Rabe-Hesketh and Choi [5]	Yes	> 0	
Other nonparametric					
Malzahn, Böhning, and Holling	MBH	Malzahn, Böhning and Holling [32]	No	$>= 0$	SMD
Non-parametric bootstrap DerSimonian and Laird	DL_b	Kontopantelis, Springate and Reeves [27]	No	$>= 0$	

Table 3. Existing comparative studies for various estimators of the between-study variance τ^2

Review paper	Estimators compared	Effect measure	Recommendations
Viechtbauer [47]	HO, DL, HS, ML, $REML$	SMD and MD	$REML$
Sidik and Jonkman [40]	HO, DL, SJ, SJ_{HO}, ML, $REML$, EB	OR	SJ_{HO} when τ^2 is expected to be small or moderate; SJ when τ^2 is expected to be large.
Kontopantelis, Springate and Reeves [27]	HO, HO_2, DL, DL_2, DL_b, DL_p, SJ, SJ_{HO}, ML, RB, RB_p	Generic	DL_b
Veroniki et al. [46]	HO, HO_2, DL, DL_2, DL_p, DL_b, PM, HM, HS, ML, $REML$, $AREML$, SJ, RB, RB_p, FB, BM	Generic	PM
Langan, Higgins and Simmonds [28]	Estimators in Veroniki et al. [46] except for FB plus IPM, SJ_{HO}, RB_u, RB_a, MBH	RR, OR, SMD, MD and Generic	PM
Petropoulou and Mavridis [37]	Estimators in Langan, Higgins and Simmonds [28] except for RB_u, RB_a	OR and MD	DL_b and DL_p
Langan et al. [29]	DL, HO, PM, PM_{HO}, PM_{DL}, HM, SJ, SJ_{HO}, $REML$	OR and Generic	$REML$, PM and PM_{DL} for continuous outcomes and non-rare binary events

less affected by outliers than conventional estimators based on the Q statistics in (1). For the purpose of robustness, they are based on Q_r and Q_m, defined as the weighted sums of absolute differences between the study-specific treatment effects and the overall treatment effect, namely

$$Q_r = \sum_{k=1}^{K} \frac{1}{s_k} \left| y_k - \hat{\theta}_{Fe} \right|, \quad Q_m = \sum_{k=1}^{K} \frac{1}{s_k} \left| y_k - \hat{\theta}_m \right|.$$

Here, $\hat{\theta}_{Fe} = \sum_{k=1}^{K} \frac{y_k}{s_k^2} / \sum_{k=1}^{K} \frac{1}{s_k^2}$, the fixed-effect estimate of θ as defined in Section 2, and $\hat{\theta}_m$ is the weighted median estimator that is the solution to the equation $\sum_{k=1}^{K} w_k \left[I(\theta \geq y_i) - 0.5 \right] = 0$, where $I(\cdot)$ is the indicator function. The estimators $\hat{\tau}_r^2$ and $\hat{\tau}_m^2$, based on Q_r and Q_m, respectively, can be derived similarly as $\hat{\tau}_{DL}^2$ by equating observed Q_r and Q_m to their corresponding expected values.

Multistep DL We first introduce the generalized method of moments (GMM) estimator of τ^2 based on the Q statistic in (1). DerSimonian and Kacker [9] showed that if the weights w_k's are treated as known constants, the expected value of Q is

$$(2)\quad \mathrm{E}(Q) = \tau^2 \left(\sum_{k=1}^{K} w_k - \frac{\sum_{k=1}^{K} w_k^2}{\sum_{k=1}^{K} w_k} \right) + \left(\sum_{k=1}^{K} w_k \sigma_k^2 - \frac{\sum_{k=1}^{K} w_k^2 \sigma_k^2}{\sum_{k=1}^{K} w_k} \right).$$

By equating Q to its expected value, replacing σ_k^2 by s_k^2 in (2), solving for τ^2 and truncating any negative solution to zero:

$$(3)\quad \hat{\tau}_{GMM}^2 = \max \left\{ \frac{Q - \left(\sum_{k=1}^{K} w_k s_k^2 - \frac{\sum_{k=1}^{K} w_k^2 s_k^2}{\sum_{k=1}^{K} w_k} \right)}{\sum_{k=1}^{K} w_k - \frac{\sum_{k=1}^{K} w_k^2}{\sum_{k=1}^{K} w_k}}, 0 \right\}.$$

The DL estimator $\hat{\tau}_{DL}^2$ [10] is a special case of $\hat{\tau}_{GMM}^2$, with $w_k = 1/s_k^2$ and $Q = Q_{Fe}$.

As discussed in Section 2, the inverse-variance weighing scheme yields $w_k = 1/(s_k^2+\hat{\tau}^2)$ when calculating the (generalized) Q statistic (1) under the *Re* model. Recall that the original DL estimator $\hat{\tau}^2_{DL}$ can be obtained by specifying $w_k = 1/s_k^2$ in (3), which is equivalent to setting $\hat{\tau}^2 = 0$ in the *Re* weights. The two-step DL method [9] first obtains $\hat{\tau}^2_{DL}$ and then sets $\hat{\tau}^2 = \hat{\tau}^2_{DL}$ in the *Re* weights to obtain $\hat{\tau}^2_{DL_2}$ from (3).

van Aert and Jackson [44] proposed the multistep DL estimator as a natural extension of the two-step DL estimator. The M-step DL estimator $\hat{\tau}^2_{DL_M}$ can be obtained recursively by computing $\hat{\tau}^2_{DL}$, $\hat{\tau}^2_{DL_2}$, ..., $\hat{\tau}^2_{DL_M}$ using (3). It has been shown that the limit of the multistep DL estimator, $\hat{\tau}^2_{DL_\infty}$, when it exists, is equivalent to the PM estimator. As further suggested by the authors, divergence problems seldom happen in practice and the convergence is usually achieved quickly.

Improved Paule and Mandel (IPM) For meta-analysis of rare binary events, Bhaumik et al. [2] adopted a standard binomial-normal random-effects model (labeled BN_{BA}), which can be specified by

$$\begin{aligned}
&x_{ki} \sim \text{Binomial}(n_{ki}, p_{ki}) \text{ for } i = 0, 1;\\
&\text{logit}(p_{k0}) = \mu_k, \text{logit}(p_{k1}) = \mu_k + \theta_k;\\
&\mu_k \sim N(\mu, \sigma^2), \theta_k \sim N(\theta, \tau^2), \mu_k \perp \theta_k \text{ for } k = 1, \ldots, K.
\end{aligned}$$

They proposed a simple average estimator, $\hat{\theta}_{sa}$, for the overall treatment effect θ and then developed the IPM estimator for τ^2 based on $\hat{\theta}_{sa}$ and the iterative PM method. The treatment effect θ_k (measured by log-odds ratio) in study k is estimated with a correction factor a added to each cell count, namely, $y_{ka} = \log[(x_{k1}+a)/(n_{k1}-x_{k1}+a)] - \log[(x_{k0}+a)/(n_{k0}-x_{k0}+a)]$. The simple average estimator for θ is then given by $\hat{\theta}_{sa} = \sum_{k=1}^K y_{ka}/K$. The authors further proved that a should be $\frac{1}{2}$ in order for $\hat{\theta}_{sa}$ to be the least biased for large samples. They noticed that the PM estimator for τ^2 depends on s_k^2 and proposed to improve PM by borrowing strength from all component studies when estimating each within-study variance,

$$\begin{aligned}
s_k^2(*) = &\frac{1}{n_{k0}+1}\left[\exp(-\hat{\mu}) + 2 + \exp(\hat{\mu})\right] + \frac{1}{n_{k1}+1}\\
&\cdot\left[\exp\left(-\hat{\mu} - \hat{\theta}_{s_{\frac{1}{2}}} + \frac{\tau^2}{2}\right) + 2 + \exp\left(\hat{\mu} + \hat{\theta}_{s_{\frac{1}{2}}} + \frac{\tau^2}{2}\right)\right].
\end{aligned}$$

Denote the corresponding weights by $w_k(*) \equiv 1/[s_k^2(*)+\tau^2]$ and $\hat{\tau}^2_{IPM}$ can be obtained by solving $Q-(K-1)=0$ iteratively with weights $w_k(*)$ in the calculation of Q.

5. CONFIDENCE INTERVALS

Table 4 reports 16 existing methods for constructing CIs for τ^2 in terms of key features including whether the algorithm for computing a CI is iterative, whether truncation for non-negativity is needed, which distribution is used for construction, and whether the CI is exact under the *Re* model. All the methods are general-purpose and so can be applied to meta-analysis of binary events except for the generalized variable approach [43], which is specifically designed for the mean difference (MD) metric based on normally distributed outcomes. Some of the CIs are obtained via a test-inverting process based on different statistics for testing $\mathcal{H}_0 : \tau^2 = 0$.

In Table 5, we list existing review papers on constructing confidence intervals for τ^2. Clearly, none of these reviews is comprehensive.

5.1 Confidence intervals based on (modified) Q statistics

Q-profile and modified Q-profile CIs Knapp, Biggerstaff and Hartung [25] and Viechtbauer [48] considered the Q-profile CIs based on the generalized Q statistic in equation (1) with weights $w_k = 1/(\tau^2+s_k^2)$, denoted by $Q(\tau^2)$, which depends on τ^2 and treats s_k^2's as if they were σ_k^2's. It can be shown that $Q(\tau^2)$ follows the χ^2_{K-1} distribution under the *Re* model in which $\theta_k \sim N(\theta, \tau^2)$ and $\epsilon_k \sim N(0, \sigma_k^2)$. It follows that $P(\chi^2_{K-1,\alpha/2} < Q(\tau^2) < \chi^2_{K-1,1-\alpha/2}) = 1-\alpha$. Based on the test-inversion principle, a $100(1-\alpha)\%$ confidence interval for τ^2 can be obtained as the interval $(\tilde{\tau}_l^2, \tilde{\tau}_u^2)$ satisfying $Q(\tilde{\tau}_l^2) = \chi^2_{K-1,1-\alpha/2}$ and $Q(\tilde{\tau}_u^2) = \chi^2_{K-1,\alpha/2}$. Since τ^2 is non-negative, $\tilde{\tau}_l^2$ is truncated to 0 if $Q(0) < \chi^2_{K-1,1-\alpha/2}$ (meaning that $\tilde{\tau}_l^2$ is negative); and the CI is set to $[0,0]$ (or $\{0\}$, the set containing only zero) if $Q(0) < \chi^2_{K-1,\alpha/2}$ (meaning that $\tilde{\tau}_u^2$ is also negative). This type of CIs is referred to as the Q-profile (QP) CIs as we are profiling $Q(\tau^2)$ with different τ^2 values when solving the above equations for $\tilde{\tau}_l^2$ and $\tilde{\tau}_u^2$ iteratively.

Knapp, Biggerstaff and Hartung [25] considered the fact that s_k^2's are only estimates and so have error variability, and constructed CIs using the test statistic $\tilde{Q}_r$ that replaces the weights in $Q(\tau^2)$ with regularized variants $w_{rk} = r_k/(\tau^2+s_k^2)$ to achieve a closer approximation to χ^2_{K-1}, where the regularization factor r_k is derived through a moment matching approach based on approximating the distribution of $\tau^2+s_k^2$ by a scaled χ^2 distribution [15]. The lower bound $\tilde{\tau}_l^2$ is obtained by profiling $\tilde{Q}_r(\tau^2)$ while the upper bound $\tilde{\tau}_u^2$ is still obtained by profiling $Q(\tau^2)$, satisfying $\tilde{Q}_r(\tilde{\tau}_l^2) = \chi^2_{K-1,1-\alpha/2}$ and $Q(\tilde{\tau}_u^2) = \chi^2_{K-1,\alpha/2}$. We refer to this type of CIs as the modified Q-profile (MQP) CIs.

Like the Q-profile CIs, the MQP CIs need left truncation to zero if the lower bound $\tilde{\tau}_l^2$ turns out to be negative, and they are set to $\{0\}$ if the upper bound $\tilde{\tau}_u^2$ is also negative. The same rule applies to all other types of CIs based on (modified) Q statistics in Section 5.1, as discussed below.

BT and BJ CIs based on the Q_{Fe} statistic Biggerstaff and Tweedie [4] proposed to approximate the distribution of Q_{Fe} by a gamma distribution with a shape parameter $r(\tau^2) \equiv \text{E}^2(Q_{Fe})/\text{Var}(Q_{Fe})$ and a scale parameter $\lambda(\tau^2) \equiv \text{Var}(Q_{Fe})/\text{E}(Q_{Fe})$. The mean and variance

Table 4. CI methods for τ^2 in random-effects meta-analysis

Method	Abbreviation	Iterative? (Y/N)	Truncation to 0? (Y/N)	Distribution Used	Exact Method for *Re*? (Y/N)	Reference
CIs based on (modified) Q statistics						
Q-Profile	QP	Y	Y	χ^2_{K-1}	Y	[15, 25]
Modified Q-Profile	MQP	Y	Y	χ^2_{K-1}	N	[15, 25]
Biggerstaff and Tweedie	BT	Y	Y	$Ga(r,\lambda)$	N	[4]
Biggerstaff and Jackson	BJ	Y	Y	A positive linear combination of χ^2_1	Y	[3]
Jackson	J	Y	Y	A positive linear combination of χ^2_1	Y	[21]
Approximate Jackson	AJ	N	Y	Normal	N	[22]
Unequal-tail Q-profile	UTQ	Y	Y	χ^2_{K-1}	Y	[23]
Profile likelihood CIs						
PL based on ML estimation	PL_{ML}	Y	Y	χ^2_1	N	[14]
PL based on REML estimation	PL_{REML}	Y	Y	χ^2_1	N	[48]
Wald CIs						
Wald based on ML estimation	W_{ML}	N	Y	$N(0,1)$	N	[4, 49]
Wald based on REML estimation	W_{REML}	N	Y	$N(0,1)$	N	[49]
Others						
Sidik and Jonkman	SJ	N	N	χ^2_{K-1}	N	[39]
Sidik and Jonkman with *HO* priori	SJ_{HO}	N	N	χ^2_{K-1}	N	[40]
Bayesian credible intervals	—	Y	N	—	N	[46]
Bootstrap	$\text{BS}_{\text{P}}/\text{BS}_{\text{NP}}$	Y	Y	—	N	[11, 27]
Generalized variable approach	GV	Y	Y	—	N	[43]

Table 5. Existing comparative studies on constructing CIs for τ^2 in random-effects meta-analysis

Review paper	CI methods reviewed/compared	Effect measure	Recommendations
Knapp, Biggerstaff and Hartung [25]	QP, MQP, BT, PL_{ML}, W_{ML}	MD/OR	QP and MQP
Viechtbauer [48]	QP, BT, PL, W, SJ, BS	OR	QP
Veroniki et al. [46]	PL, W, BT, BJ, J, QP, SJ, BS, BC	Generic	—
van Aert, van Assen and Viechtbauer [45]	QP, BJ, J	OR	None recommended when $p_{ki} < 0.1$ in combination with either $K \geq 80$ or ($K \geq 40$ and $n_{ki} < 30$)

of Q_{Fe} under the *Re* model are given by $\text{E}(Q_{Fe}) = (K-1) + (S_1 - S_2/S_1)\,\tau^2$ and $\text{Var}(Q_{Fe}) = 2(K-1) + 4\,(S_1 - S_2/S_1)\,\tau^2 + 2\,(S_2 + S_2^2/S_1^2 - 2S_3/S_1)\,\tau^4$, where $S_r \equiv \sum_{k=1}^{K}[1/s_k^2]^r$. CIs for τ^2 can be obtained similarly based on this gamma approximation instead of χ^2_{K-1} using the above profiling approach, which we refer to as the BT intervals.

Biggerstaff and Jackson [3] derived the exact CDF of Q_{Fe} under the *Re* model, denoted by $F_Q(q;\tau^2)$, as a positive linear combination of χ^2_1 random variables, whose cumulative distribution function can be obtained using Farebrother's algorithm [12]. They then obtained $(\tilde{\tau}_l^2, \tilde{\tau}_u^2)$ by solving the following two equations numerically, $F_Q(c\hat{\tau}^2_{uDL} + K - 1; \tilde{\tau}_l^2) = 1 - \alpha/2$ and $F_Q(c\hat{\tau}^2_{uDL} + K - 1; \tilde{\tau}_u^2) = \alpha/2$, where $c = S_1 - S_2/S_1$ and $\hat{\tau}^2_{uDL} = [Q_{Fe} - (K-1)]/c$ is the untruncated version of the *DL* estimator of τ^2. This type of CIs is referred to as the BJ intervals.

Jackson and approximate Jackson CIs Following the numerical approach in [3], Jackson [21] proposed CIs by test inversion based on the generalized Q in equation (1), which is also distributed as a positive linear combination of χ^2_1 random variables under the *Re* model. Jackson, Bowden and Baker [22] further proposed to apply the arcsinh transformation to the untruncated version of $\hat{\tau}^2_{GMM}$ for variance stabilization and then constructed CIs for τ^2 based on a normal approximation. These types of CIs are referred to as the Jackson (J) and approximate Jackson (AJ) CIs, respectively. Based on simulation, Jackson [21] further commented that weighting component studies by the reciprocal of their within-study standard errors (i.e. $1/s_k$), rather than by their variances (i.e. $1/s_k^2$) as the convention dictates, appears to provide a sensible and viable option when there is little a priori knowledge about the extent of heterogeneity.

Unequal-tail Q profile CIs Jackson and Bowden [23] advocated to use unequal tail probabilities to obtain shorter intervals whenever such methods are justifiable. For example, when constructing a $100(1-\alpha)\%$ unequal-tail Q-profile (UTQ) confidence interval, the lower and upper bounds, $\tilde{\tau}_l^2$ and $\tilde{\tau}_u^2$, are obtained by solving $Q(\tilde{\tau}_l^2) = \chi^2_{K-1,1-\alpha_1}$ and $Q(\tilde{\tau}_u^2) = \chi^2_{K-1,\alpha_2}$, respectively, where $\alpha_2 > \alpha_1$ and $\alpha_1 + \alpha_2 = \alpha$. They further suggested to use a pre-specified α-split with $\alpha_1 = 0.01$ and $\alpha_2 = 0.04$ for a 95% CI, which was shown to be able to retain the nominal coverage and

reduce the width under the *Re* model. Obviously, the idea of unequal tails can be applied to all kinds of confidence intervals. In our numerical evaluation, we examine the performance of the Q-profile CIs with $\alpha_1 = 0.01$ and $\alpha_2 = 0.04$ as a representative case.

5.2 Profile likelihood confidence intervals

Under the *Re* model, Hardy and Thompson [14] proposed the profile likelihood CIs based on maximum likelihood (ML) estimation, referred to as PL_{ML}. The joint log-likelihood function of (θ, τ^2) is

$$l(\theta, \tau^2) = -\frac{K}{2}\ln 2\pi - \frac{1}{2}\sum_{k=1}^{K}\ln(\tau^2 + \sigma_k^2) - \frac{1}{2}\sum_{k=1}^{K}\frac{(y_k - \theta)^2}{\tau^2 + \sigma_k^2}.$$

Given the value of τ^2, the ML estimator of θ is

$$\hat{\theta}_{ML}(\tau^2) = \sum_{k=1}^{K}\frac{1}{\tau^2 + \sigma_k^2}y_k \Bigg/ \sum_{k=1}^{K}\frac{1}{\tau^2 + \sigma_k^2}.$$

The profile log-likelihood for τ^2, written as $l(\hat{\theta}_{ML}(\tau^2), \tau^2)$, takes into account the fact that θ is also unknown and must be estimated. Then a $100(1-\alpha)\%$ CI for τ^2 is given by the set of τ^2 values satisfying $l(\hat{\theta}_{ML}(\tau^2), \tau^2) > l(\hat{\theta}_{ML}(\hat{\tau}^2_{ML}), \hat{\tau}^2_{ML}) - \chi^2_{1,1-\alpha}/2$.

Viechtbauer [48] proposed to construct profile likelihood CIs based on restricted maximum likelihood (REML) estimation, referred to as PL_{REML}. The $100(1-\alpha)\%$ CI for τ^2 is given by the set of τ^2 values satisfying $l_R(\tau^2) > l_R(\hat{\tau}^2_{REML}) - \chi^2_{1,1-\alpha}/2$, where the restricted log-likelihood function of τ^2 is

$$l_R(\tau^2) = -\frac{K}{2}\ln 2\pi - \frac{1}{2}\sum_{k=1}^{K}\ln(\tau^2 + \sigma_k^2) - \frac{1}{2}\ln\sum_{k=1}^{K}\frac{1}{\tau^2 + \sigma_k^2} - \frac{1}{2}\sum_{k=1}^{K}\frac{\left(y_k - \hat{\theta}_{ML}(\tau^2)\right)^2}{\tau^2 + \sigma_k^2},$$

and $\hat{\tau}^2_{REML}$ is the REML estimate of τ^2 (by maximizing l_R). Viechtbauer [48] found that the REML-based CIs were slightly more accurate than the ML-based CIs in terms of coverage probability, especially for small K.

Because ML and REML estimates of τ^2 require non-negativity, the lower bounds of profile likelihood (PL) intervals are always non-negative and the upper bounds are strictly positive after applying the same truncation for Q-profile CIs.

5.3 Wald confidence intervals

The Wald test statistics for testing $\mathcal{H}_0 : \tau^2 = 0$ under the *Re* model have the form $W = \hat{\tau}^2/SE(\hat{\tau}^2)$, where $\hat{\tau}^2$ can be $\hat{\tau}^2_{ML}$ or $\hat{\tau}^2_{REML}$, and the standard error is estimated by

$$\widehat{SE}(\hat{\tau}^2_{ML}) = \sqrt{2\left[\sum_{k=1}^{K} w^2_{ML.k}\right]^{-1}},$$

$$\widehat{SE}(\hat{\tau}^2_{REML}) = \sqrt{2\left[\sum_{k=1}^{K} w^2_{REML.k} - 2\frac{\sum_{k=1}^{K} w^3_{REML.k}}{\sum_{k=1}^{K} w_{REML.k}} + \left(\frac{\sum_{k=1}^{K} w^2_{REML.k}}{\sum_{k=1}^{K} w_{REML.k}}\right)^2\right]^{-1}}$$

with $w_{ML.k} = \left(\hat{\tau}^2_{ML} + s_k^2\right)^{-1}$ and $w_{REML.k} = \left(\hat{\tau}^2_{REML} + s_k^2\right)^{-1}$. We label the Wald statistics based on ML and REML estimation by W_{ML} and W_{REML}, respectively. The corresponding $100(1-\alpha)\%$ Wald (W) CIs for τ^2 are $\hat{\tau}^2_{ML} \pm z_{1-\alpha/2}\widehat{SE}(\hat{\tau}^2_{ML})$ or $\hat{\tau}^2_{REML} \pm z_{1-\alpha/2}\widehat{SE}(\hat{\tau}^2_{REML})$ [4, 48], where z_α is the 100α-th percentile of the standard normal distribution. Negative lower bounds of the Wald CIs should be truncated to 0 since both ML and REML estimates of τ^2 are constrained to be non-negative.

5.4 Other confidence intervals

Sidik and Jonkman (SJ) CIs Sidik and Jonkman [39] proposed confidence intervals based on the SJ estimator of τ^2, which is derived from the weighted residual sum of squares in the framework of a linear regression model. Let the crude estimate $\hat{\tau}_0 = \sum_{k=1}^{K}(y_k - \bar{y})^2/K$ be an a priori value for τ^2. Then the SJ estimator is given by $\hat{\tau}^2_{SJ} = \frac{\hat{\tau}^2_0}{K-1}\sum_{k=1}^{K}\hat{w}_k(y_k - \hat{\theta}_0)^2$, where $\hat{w}_k = 1/(s_k^2 + \hat{\tau}^2_0)$, and $\hat{\theta}_0 = \sum_{k=1}^{K}\hat{w}_k y_k / \sum_{k=1}^{K}\hat{w}_k$. It follows that $(K-1)\hat{\tau}^2_{SJ}/\tau^2$ has an asymptotic distribution of χ^2_{K-1}. Thus an approximate $100(1-\alpha)\%$ confidence interval can be calculated as

$$\frac{(K-1)\hat{\tau}^2_{SJ}}{\chi^2_{K-1,1-\alpha/2}} \le \tau^2 \le \frac{(K-1)\hat{\tau}^2_{SJ}}{\chi^2_{K-1,\alpha/2}}.$$

Since $\hat{\tau}^2_{SJ}$ is always positive, the SJ confidence intervals have positive lower and upper bounds. Sidik and Jonkman [40] later proposed an improved estimator $\hat{\tau}^2_{SJ_{HO}}$ by using $\hat{\tau}^2_{HO}$ as the a priori value. Then improved confidence intervals can be constructed correspondingly.

Bayesian credible intervals Bayesian credible (BC) intervals can be obtained when a Bayesian approach is employed and posterior samples are drawn from the (joint) posterior distribution of all parameters involved using an MCMC algorithm. The lower and upper points of a $100(1-\alpha)\%$ CI can be the $100(\alpha/2)$th and $100(1-\alpha/2)$th percentiles, respectively, of the posterior sample of τ^2's, or can be determined by the region that gives the highest posterior density. Such intervals may be heavily affected by the prior selection when K, the number of studies is small.

Bootstrap CIs Bootstrap techniques can be used to obtain confidence intervals for nearly all τ^2 estimators. For nonparametric bootstrap (denoted by BS_{NP}), we sample K studies with replacement from the observed set of studies B

times to get B bootstrap samples. For parametric bootstrap (denoted by BS_P), we first obtain the parameter estimates and then generate B samples from the assumed distributions with these estimates. For each (parametric or nonparametric) sample, we calculate the corresponding estimate $\hat{\tau}^2$. Then the $100(\alpha/2)$th and $100(1-\alpha/2)$th percentiles of the B estimates of τ^2 are, respectively, the lower and upper bounds of a $100(1-\alpha)\%$ bootstrap confidence interval. In our numerical experiment, we only perform the nonparametric bootstrap procedure for the DL estimator for illustration.

The generalized variable (GV) approach For meta-analysis of normally distributed outcomes, Tian [43] proposed inference procedures based on the generalized pivotal quantity for τ^2. A pivotal quantity is a function of observations and parameters such that the distribution of the function does not depend on the parameters including nuisance parameters. Let σ_{k0}^2 (σ_{k1}^2) be the population variance of the control (treatment) group in study k; let s_{k0}^2 (s_{k1}^2) be the corresponding sample variance. For normally distributed outcomes, it is well known that $V_{ki} \equiv (n_{ki}-1)s_{ki}^2/\sigma_{ki}^2 \sim \chi_{n_{ki}-1}^2$ for $k = 1, \ldots, K$ and $i = 0, 1$. Denote Q in (1) with weight $w_k = 1/(\sigma_{k0}^2/n_{k0} + \sigma_{k1}^2/n_{k1} + \tau^2)$ by $Q(\tau^2)$, which follows χ_{K-1}^2 and is a monotonic decreasing function of τ^2. Thus, given a real number $\eta \geq 0$, there exists a unique $\tau_\eta^2 \geq 0$ such that $Q(\tau_\eta^2) = \eta$. Based on this, Tian [43] defined the generalized pivotal quantity R_{τ^2} for τ^2 as $R_{\tau^2} = \tau_\eta^2$ if $\eta \leq Q(0)$ and $R_{\tau^2} = 0$ otherwise. Given the observed treatment effects y_k's and sample variances s_{ki}^2's, the distribution of R_{τ^2} does not depend on any nuisance parameters. A series of R_{τ^2} values can be obtained by first simulating $V_{ki} \sim \chi_{n_{ki}-1}^2$ and $\eta \sim \chi_{K-1}^2$ and setting $\sigma_{ki}^2 = (n_{ki}-1)s_{ki}^2/V_{ki}$ in $Q(\tau^2)$ for $k = 1, \ldots, K$ and $i = 0, 1$, and then solving for τ_η^2. A $100(1-\alpha)\%$ confidence interval is given by $(R_{\tau^2,\alpha/2}, R_{\tau^2,1-\alpha/2})$, where the lower and upper bounds are the $100(\alpha/2)$th and $100(1-\alpha/2)$th percentiles, respectively,x of the generated R_{τ^2}'s.

6. SIMULATION FOCUSING ON RARE BINARY EVENTS

For meta-analysis of rare binary events, Li and Wang [30] conducted a comprehensive simulation study to compare the performance of various estimators of the overall treatment effect θ measured by log-odds ratio, where a flexible binomial-normal model was used to accommodate treatment groups with unequal variability. This model, labeled BN_{LW}, specifies the event probabilities by

$$\text{logit}(p_{k0}) = \mu_k - \omega\theta_k, \quad \text{logit}(p_{k1}) = \mu_k + (1-\omega)\theta_k,$$

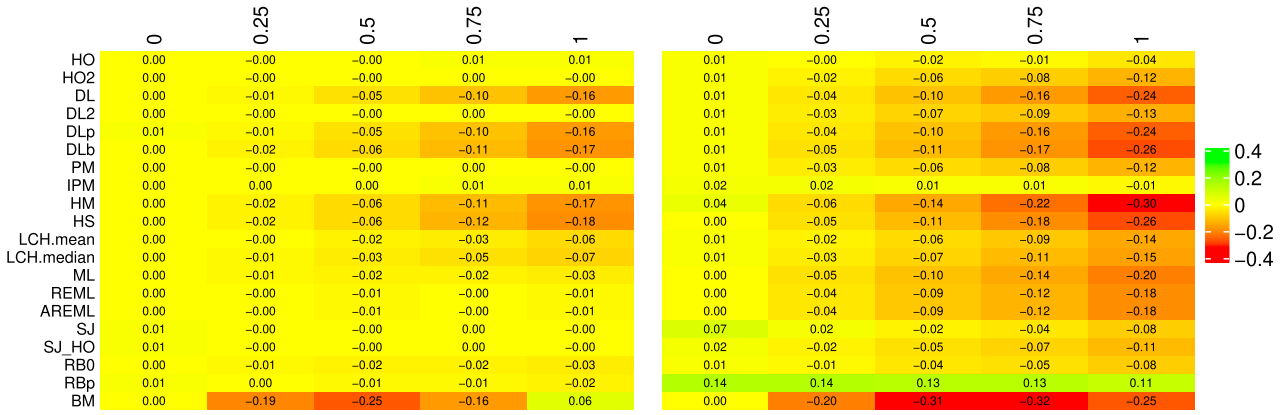
where $\mu_k \sim N(\mu, \sigma^2)$, $\theta_k \sim N(\theta, \tau^2)$, $\mu_k \perp \theta_k$, and ω is a constant in [0, 1]. The random-effects model BN_{BA} in [2] is a special case of BN_{LW} with $\omega = 0$. Further, when $\omega = 1/2$, it reduces to the model in [41], which assumes the equality of the variances of $\text{logit}(p_{k0})$ and $\text{logit}(p_{k1})$.

In this section, we adopt the same model and simulation setup from [30], to examine the performance of various methods. Results are summarized in Sections 6.1 and 6.2 for estimating the between-study variance τ^2 of log-odds ratios θ_k's. Here, bias and MSE are reported for point estimation, and the actual coverage probability and width of confidence intervals are reported for interval estimation. To be specific, we set the number of studies K to 10, 20 and 50 to reflect different sizes of meta-analysis. We generate the number of events x_{ki} from $\text{Binomial}(n_{ki}, p_{ki})$ for $k = 1, \ldots, K$ and $i = 0, 1$. The number of subjects n_{k0}'s in the control groups are generated from $\text{Uniform}[2000, 3000]$ to examine large-sample performance and from $\text{Uniform}[20, 1000]$ to examine small-sample performance, and then rounded to the nearest integers. To allow varying allocation ratios across studies, the within-study sample sizes are set to follow the relationship $n_{k1} = R_k n_{k0}$, where $\log_2 R_k \sim N(\log_2 R, \sigma_R^2)$, $R \in \{1, 2, 4\}$ and $\sigma_R^2 = 0.5$. For small sample sizes, as noted in [30], the range $[20, 1000]$ is chosen so that the empirical means of $\min(n_{k0}p_{k0}, n_{k1}p_{k1})_{k=1}^K$ in all the settings are below one while it still allows for cases where most component studies have small sample sizes but a few can have sample sizes close to 1000. To generate p_{ki}'s, we fix σ^2 at 0.5, and set $\tau^2 \in \{0, 0.25, 0.5, 0.75, 1\}$ for evaluating different estimators and $\tau^2 \in \{0, 0.1, 0.2, \cdots, 0.9, 1\}$ for evaluating different types of CIs. We further set $\theta \in \{-1, 0, 1\}$ to reflect different directions of the overall treatment effect, set $\mu \in \{-2.5, -5\}$ to represent low and very low incidence rates of the binary event (i.e., 0.076 and 0.0067 in the probability scale), and set $\omega \in \{0, 0.5, 1\}$ to represent smaller/equal/larger variability in the control group, compared to the treatment group. For each setting, 1000 datasets are simulated to compute empirical values of the performance measures by taking the average.

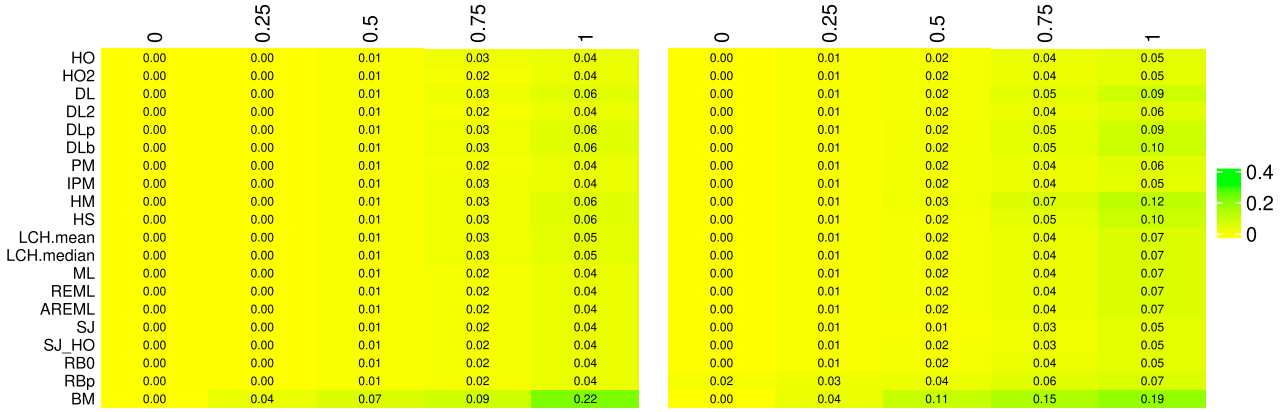
6.1 Comparison of different heterogeneity estimators

We compare all the methods listed in Table 2 except for FB and MBH. Since the full Bayesian method can be greatly affected by the prior choice and other factors (such as convergence), we exclude FB from our simulation. The MBH method is designed specifically for standard mean difference, thus not suitable for binary events. In addition, the empirical Bayes method EB is equivalent to PM, and the multistep DL method has the property that DL_∞ converges to PM. Therefore, we include PM in the comparison and leave EB and DL_M out. We use heat maps to visualize the bias and MSE results where the rows of each map represent different methods and columns represents different τ^2 values in [0,1].

Large-sample results Figure 1 presents the bias and MSE results of different estimators for $\mu = -2.5$ and $\mu = -5$ based on large-sample settings with $R = 1$, $K = 50$, $\theta = 0$, and $w = 0$. As shown in Figure 1(a), when the event of



(a) Comparison of estimation bias. Left panel: $\mu = -2.5$; Right panel: $\mu = -5$.

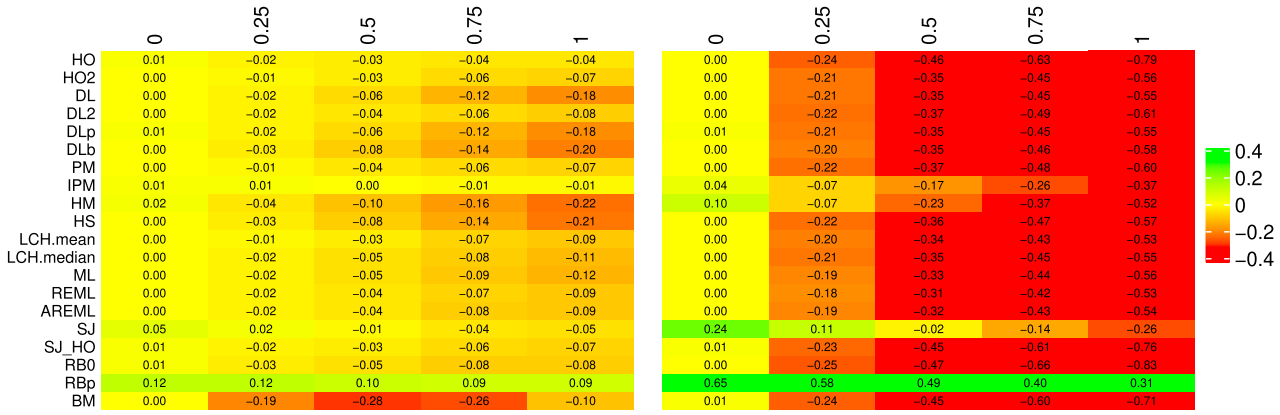


(b) Comparison of MSE. Left panel: $\mu = -2.5$; Right panel: $\mu = -5$.

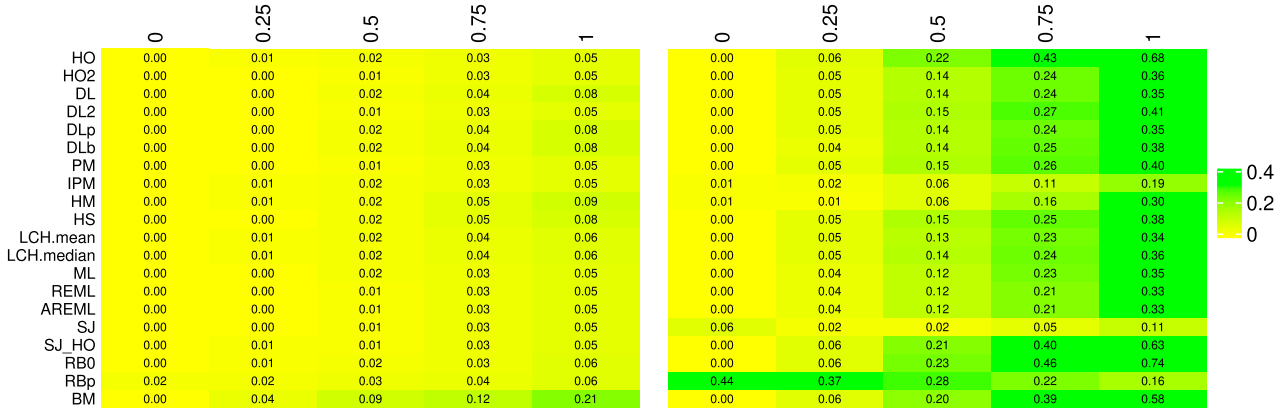
Figure 1. Large-sample performance of different τ^2 estimators based on settings with $R = 1$, $K = 50$, $\theta = 0$, and $w = 0$.

interest becomes rarer, all methods seem to produce more bias in estimating τ^2. Almost all methods underestimate the between-study heterogeneity when $\tau^2 > 0$. The RBp estimator, however, consistently overestimates τ^2 when the event is very rare ($\mu = -5$). As τ^2 increases, most estimators produce more bias except for BM and RB_p; the bias from BM first increases then decreases, and the bias from RB_p decreases for very rare events ($\mu = -5$). When the events are not that rare ($\mu = -2.5$), most estimators have similarly low bias except for the one-step DL estimators (DL, DLp, DL_b), HM, HS, and BM. However, IPM stands out with the lowest bias when the incidence rate becomes very low, especially when $\tau^2 \geq 0.5$. The HS, HM, BM and one-step DL family methods remain the worst and should be avoided in terms of bias. All three likelihood-based methods, ML, $REML$ and $AREML$, produce similar results with a moderate level of bias. In terms of MSE, most methods have similar performance except for HM and BM, which are the most inefficient according to Figure 1(b). Those with relatively large magnitude of bias tend to have relatively large MSE.

We next discuss the potential impacts of R, K, θ, and w on the estimation performance for the large-sample case. Figures S1 and S2 in the Supplementary Material (SM) (http://intlpress.com/site/pub/files/_supp/sii/2020/0013/0004/SII-2020-0013-0004-s003.pdf) show the bias and MSE results for different R and K values, respectively, based on settings with $\mu = -2.5$, $\theta = 0$ and $w = 0$. We can see that when $\tau^2 < 0.5$, regardless of R and K, all the methods perform somewhat similarly and have both bias and MSE close to zero except for BM which has much larger bias. As K increases, MSE decreases significantly for every estimator when $\tau^2 \geq 0.5$ but bias for a few estimators seems not to get closer to zero (e.g., DL for $\tau^2 = 1$, BM for $\tau^2 = 0.5$, and 0.75). However, the heat maps show very similar color patterns both vertically and horizontally, indicating that the impact of R and K on the *relative* performance of these methods is merely marginal. Figures S3 and S4 in the SM show the bias and MSE results for different θ and w values, respectively, based on settings with $R = 1$, $K = 50$ and $\mu = -5$. When $\theta = -1$, bias decreases as w increases while this trend reverses when $\theta = 1$. This effect of w is minimal



(a) Comparison of estimation bias. Left panel: $\mu = -2.5$; Right panel: $\mu = -5$.



(b) Comparison of MSE. Left panel: $\mu = -2.5$; Right panel: $\mu = -5$.

Figure 2. Small-sample performance of different τ^2 estimators based on settings with $R = 1$, $K = 50$, $\theta = 0$, and $w = 0$.

when there is no treatment effect ($\theta = 0$). Similar trends are observed but less obvious for MSE. Also, we find that IPM maintains the best performance in terms of both bias and MSE while DL, DLp, DL_b, HS, HM, and BM are among the worst in nearly all the settings considered.

Small-sample results Figure 2 presents the bias and MSE results of different estimators for $\mu = -2.5$ and $\mu = -5$ based on small-sample settings with $R = 1$, $K = 50$, $\theta = 0$, and $w = 0$. From Figure 2(a), we can see that when $\tau^2 > 0$, the underestimation observed in the large-sample results for all the estimators but RB_p is much more severe for small samples, where the magnitude of bias increases substantially for very rare events ($\mu = -5$). Note that RB_p consistently overestimates τ^2 for both $\mu = -2.5$ and $\mu = -5$, and unlike most other estimators, the bias decreases as τ^2 increases. When events are not that rare ($\mu = -2.5$), IPM is still the least biased. However, for very rare events ($\mu = -5$), SJ becomes the least biased estimator for $\tau^2 \geq 0.5$. The problem of SJ is that it significantly overestimates τ^2 when there is no or little heterogeneity, due to its positive nature. From Figure 2(b) we can see that MSE does not change much when $\mu = -2.5$ but dramatically increases when $\mu = -5$ compared to results from large samples. For very rare events ($\mu = -5$), SJ is the most efficient method except for $\tau^2 = 0$ and IPM seems to be the second best in terms of MSE. Note that when $\tau^2 = 1$, RB_p has smaller MSE than IPM for very rare events, but it does not perform as well as IPM for smaller τ^2 values.

The impacts of R, K, θ, and w on the estimation bias and MSE for the small-sample case are shown in Figures S5-S8 of the SM. Since several methods (e.g., the likelihood-based methods) failed in some small-sample settings for very rare events ($\mu = -5$), we show results for $\mu = -2.5$ in these figures. Although the effect of K on MSE becomes more significant for small samples (i.e., MSE decreases more as K increases), it is still the case that both R and K have little impact on the *relative* performance of different methods. Also, similar trends for both bias and MSE occur when w and θ change as in the large-sample case. For these $\mu = -2.5$ settings, IPM seems to be the best estimator due to its consistent top-level performance across various settings. This also agrees with the results in the left panels of Figure 2. On

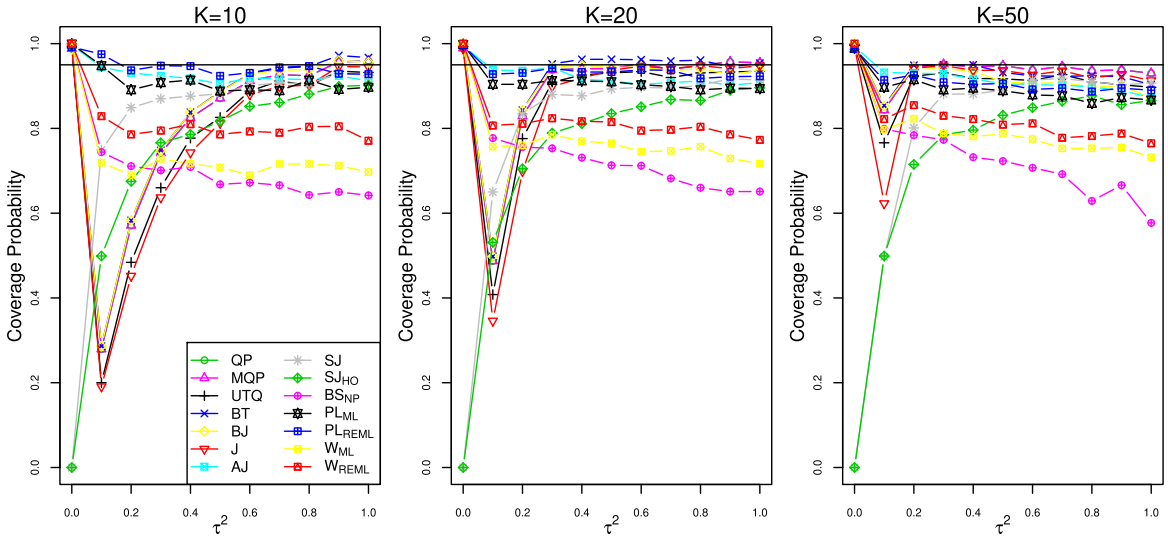


Figure 3. Actual coverage probabilities of different types of 95% CIs for different K values based on large-sample settings with $R = 1$, $\mu = -5$, $\theta = 0$, and $w = 0$.

the other hand, DL, DL_p, DL_b, HM, HS, and BM should be used with caution due to their generally large bias.

6.2 Comparison of different types of CIs

Among those summarized in Table 4, we compared 14 different types of 95% CIs for the heterogeneity parameter τ^2 in Figures 3 and 4, excluding Bayesian credible intervals and the GV method as before. As mentioned in Section 5, BS_{NP} represents the nonparametric bootstrap procedure combined with the DL estimator and UTQ represents the unequal-tail Q-profile CI with $\alpha_1 = 0.01$ and $\alpha_2 = 0.04$. Again, from our (unreported) simulation results, we find that the influences of R, θ, and w on the empirical coverage probability are marginal.

Figure 3 shows actual coverage probabilities of different types of CIs for different K values based on large-sample settings with $R = 1$, $\mu = -5$, $\theta = 0$, and $w = 0$. When there is no between-study heterogeneity ($\tau^2 = 0$), all the methods provide 100% coverage except for SJ and SJ_{HO}, which produce strictly positive intervals and so have zero coverage. When τ^2 is small, as K increases, the methods based on (modified) Q statistics gain some improvement in coverage except for AJ, which achieves relatively high coverage for all K and τ^2 values. As τ^2 gets larger, most methods do not improve their coverage by increasing K.

Figure 4 presents actual coverage probabilities of different types of CIs for both large- and small-sample cases and different μ values based on settings with $R = 1$, $K = 20$, $\theta = 0$, and $w = 0$. When $\mu = -2.5$, most methods have actual coverage close to the nominal level 0.95. Among all, the nonparametric bootstrap CI has the lowest coverage, followed by the two Wald CIs when $\tau^2 > 0$. The influence of sample sizes is not obvious except for J, SJ and SJ_{HO} that improve their coverage for large sample sizes when τ^2 is small. For very rare events ($\mu = -5$), the impact of sample sizes is much more severe and some of the CIs (e.g., SJ_{HO}, J, UTQ) do not even achieve 50% coverage in most small-sample settings. In the large-sample settings, PL_{ML}, PL_{REML}, and AJ maintain the nominal 95% coverage quite well at all positive levels of τ^2. As the sample sizes become small, all methods fail to do so for very rare events when $\tau^2 \geq 0.3$. Still, PL_{ML} and PL_{REML}, and AJ are among those with the highest coverage. We also find that when $\tau^2 \geq 0.4$, SJ joins the top-performing group with the following order SJ $\approx$ PL_{REML} > PL_{ML} > AJ. This matches with the estimation results reported in Section 6.1 that for very rare events coupled with small samples, the SJ estimator is the least biased and has the smallest MSE when $\tau^2 \geq 0.5$. In such situations, the Q statistic-based CIs have generally low coverage and thus should be avoided; meanwhile the Wald and nonparametric bootstrap CIs have moderate coverage instead of being the worst in the other three cases.

Figure 5 shows width curves of different types of CIs under the same settings of Figure 4, where for all CIs, the width shows an increasing pattern as τ^2 increases. The influence of sample sizes on the CI width is only obvious when $\mu = -5$, where all the CIs become narrower when sample sizes decrease. Though anti-intuitive, a closer examination reveals that when events are very rare and sample sizes are small, many simulation iterations produce confidence intervals of a point $\{0\}$, which makes the average width become smaller. In the first three situations (either $\mu = -2.5$ or large samples), BT and BJ produce the widest intervals, and PL and AJ intervals, which offer higher coverage than most other methods, have moderate widths among all. Unsurprisingly, the nonparametric bootstrap procedure produces the narrowest CIs. In the last situation (very rare events coupled

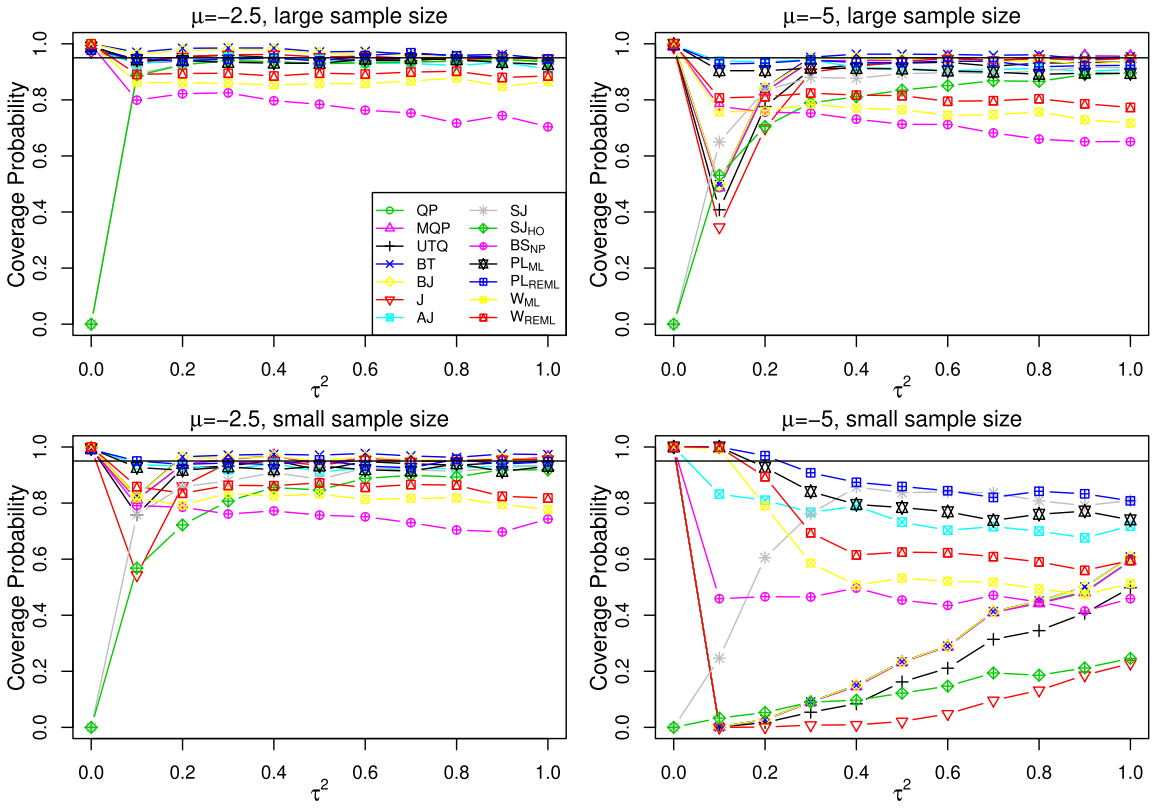


Figure 4. Actual coverage probabilities of different types of 95% CIs for both large- and small-sample cases and different μ values based on settings with $R = 1$, $K = 20$, $\theta = 0$, and $w = 0$.

with small samples), PL and AJ intervals are among the widest. Here, CIs with shorter widths are not necessarily desirable as they may reflect more {0} intervals due to sparsity. SJ produces intervals with moderate widths though it also provides higher coverage when τ^2 is large. Overall, we recommend PL and AJ intervals in meta-analysis of rare binary events for their high coverage. For very rare events with small samples, we recommend SJ intervals if we know there exists at least moderate-level heterogeneity. Besides, AJ and SJ intervals are much easier to obtain than PL intervals.

7. EXAMPLE: TYPE 2 DIABETES MELLITUS AFTER GESTATIONAL DIABETES

Women with gestational diabetes are believed to have a higher chance to develop type 2 diabetes. Bellamy et al. [1] performed a comprehensive systematic review and meta-analysis to assess the strength of this association. They selected 20 cohort studies that included 675,455 women with/without gestational diabetes and 10,859 type 2 diabetic events from 205 reports between Jan 1, 1960 and Jan 31, 2009 from Embase and Medline (see Table S1 of the SM). We reanalyzed the data focusing on inference about the heterogeneity parameter τ^2. Note that the overall event rate is ∼1.61% and many studies have very small sample sizes with zero event counts. So this data example fits in the scenario of very rare events coupled with small sample sizes. Recall that in this scenario, SJ gives the least bias and most efficient estimator when there exists a moderate or large level of heterogeneity and IPM is the second best which tends to underestimate τ^2.

Point estimates for the heterogeneity parameter τ^2 and the corresponding inverse-variance weighted estimates for the overall treatment effect θ (measured by log-odds ratio) are summarized in Table 6. Here, most methods give an estimate between 0.4 and 0.7 for τ^2, where the estimate from IPM is 0.563 and that from SJ is 0.679. This seems to suggest a moderate to high level of heterogeneity, especially after accounting for the underestimation from IPM. The RB_p method, which has been shown to severely overestimate τ^2 for very rare events, not surprisingly gives the largest estimate of 1.162. On the other hand, the HS estimate is much smaller than the others. The resulting estimated odds ratios do not vary as much except for the one from RB_p. Table 7 shows the confidence intervals from all the compared methods. BT gives a very large upper bound, which seems to be odd. All CIs except for those from BT, BJ, and Wald methods exclude zero, among which SJ yields the shortest interval with the largest lower bound and the upper bound in line

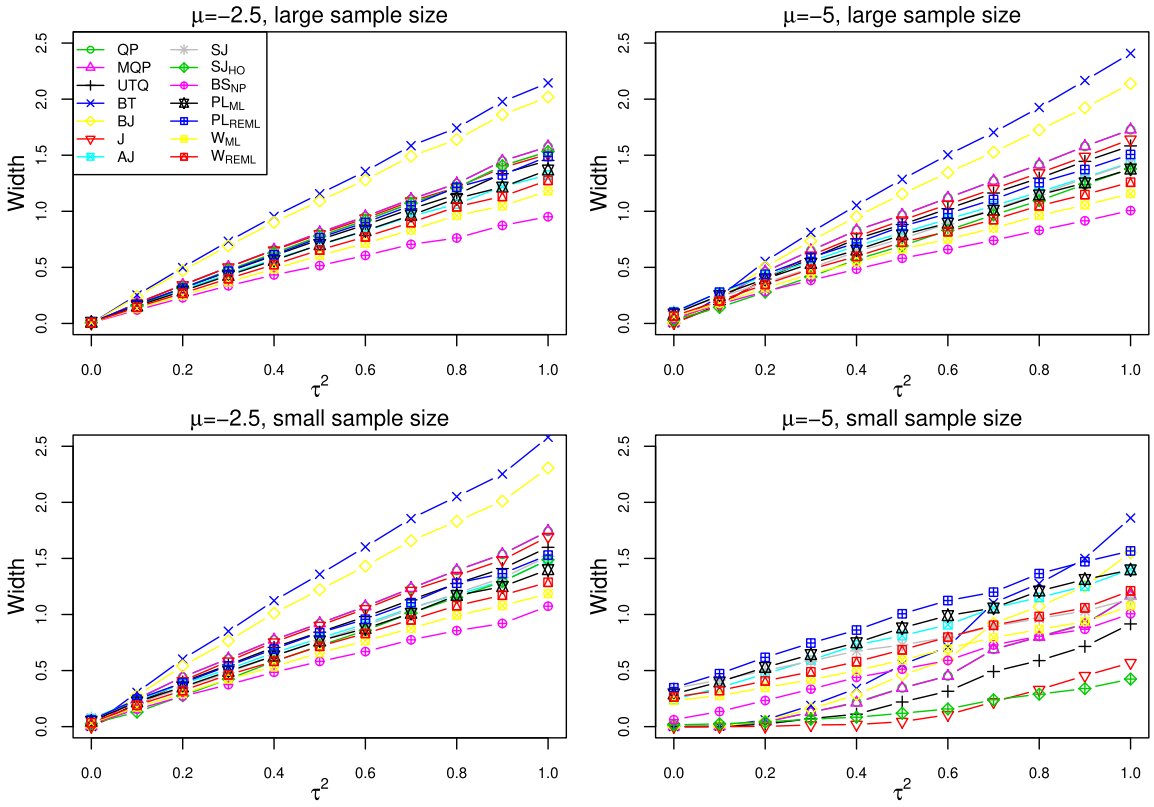


Figure 5. Width curves of different types of 95% CIs for both large- and small-sample cases and different μ values based on settings with $R = 1$, $K = 20$, $\theta = 0$, and $w = 0$.

with that from PL and AJ methods. Recall that SJ tends to produce the best interval with higher coverage and relatively shorter width when there exists at least moderate-level heterogeneity, as reported in Section 6.2. In this example, we lean toward reporting the SJ interval, among the top performing methods PL, AJ and SJ. Based on the estimation and inference results above, we believe that these studies are heterogeneous.

8. DISCUSSION AND RECOMMENDATIONS

Based on our comprehensive simulation studies for large-sample meta-analysis of rare binary events, we recommend the IPM method for estimating the heterogeneity parameter τ^2 if reducing estimation bias is of high priority, especially when the events are extremely rare. Most of the methods do not differ much in terms of MSE. We suggest to avoid using HM, HS and BM since they have relatively large bias and MSE compared with other estimators. The most widely used DL estimator and its one-step variants DL_p and DL_b do not perform satisfactorily and hence should be avoided. For small-sample meta-analysis of rare events, IPM is still recommended and SJ also performs much better than the other estimators in terms of both bias and MSE when $\tau^2 \geq 0.5$ and the events are extremely rare. In terms of interval estimation, we recommend the profile likelihood methods (PL_{ML} and PL_{REML}) and the approximate Jackson method AJ in general situations. Among the three, PL_{REML} usually produces higher coverage but with wider intervals. The SJ method is a good candidate when events are extremely rare, sample sizes are small, and $\tau^2 \geq 0.4$. We did not examine the performance of Bayesian methods because of the computation burden, convergence detection issue, and potential sensitivity to prior choices. However, Bayesian hierarchical modeling can be a good alternative especially when meaningful prior information is available.

We notice that most estimators for τ^2 are negatively biased in our simulation, an interesting phenomenon observed in other simulation studies with binary outcomes [2, 26, 39, 40] as well. In simulation studies with continuous outcomes [27], most of the estimators show positive bias when τ^2 is small (<0.1) and the magnitude of bias of RB_p is much larger than the other estimators; for larger τ^2 values, the HS and ML estimators are negatively biased and the magnitude increases as τ^2 increases [47]. Viechtbauer [47] provides some analytical results for the bias of estimators HO, DL, HS, ML, and $REML$. Most of these results were derived based on the homogeneous within-study variance assumption ($\sigma_k^2 = \sigma^2$). Under this assumption, the

Table 6. Data example of gestational diabetes meta-analysis: estimates for τ^2 and θ from different methods

Estimator	HO	HO_2	DL	DL_2	DL_p	DL_b	PM	IPM	HM	HS
$\hat{\tau}^2$	0.220	0.418	0.466	0.411	0.466	0.265	0.413	0.563	0.419	**0.046**
$\hat{\theta}$	2.093	2.136	2.146	2.135	2.146	2.104	2.135	2.162	2.137	2.092
OR	8.112	8.469	8.547	8.457	8.547	8.197	8.461	8.691	8.470	8.099
Estimator	LCH_{mean}	LCH_{median}	ML	$REML$	$AREML$	SJ	SJ_{HO}	RB_0	RB_p	BM
$\hat{\tau}^2$	0.519	0.298	0.396	0.449	0.433	0.679	0.290	0.198	**1.162**	0.195
$\hat{\theta}$	2.155	2.111	2.132	2.142	2.139	2.180	2.110	2.088	2.235	2.088
OR	8.626	8.260	8.432	8.520	8.493	8.846	8.245	8.072	9.345	8.067

Table 7. Data example of gestational diabetes meta-analysis: confidence intervals for τ^2 from different methods

Method	QP	MQP	UTQ	BT	BJ	J	AJ
CI	(0.109, 1.603)	(0.106, 1.603)	(0.083, 1.403)	[0, 8.610)	[0, 2.660)	(0.048, 1.540)	(0.004, 1.396)
Method	SJ	SJ_{HO}	BS_{NP}	PL_{ML}	PL_{REML}	W_{ML}	W_{REML}
CI	(0.393,1.449)	(0.168, 0.620)	(0.012, 0.670)	(0.113, 1.285)	(0.129, 1.458)	[0, 0.841)	[0, 0.966)

bias due to truncation is always positive for DL, HO and $REML$ with all levels of heterogeneity and is negative for HS and ML when $\tau^2 \geq 0.5$. However, we believe that in the rare events context, it is the sparsity (caused by zero counts) and lack of resolution in estimating the within-study variances that cause the large magnitude of underestimation for many methods. This underestimation is much reduced by the IPM estimator where the within-study variance estimates are improved by pooling information from all the studies.

Finally, we should mention that, when synthesizing information from multiple studies to obtain more reliable conclusions, one should not simply rely on one point estimate or one p-value (especially those from the default methods in software packages) without considering the rich selection of statistical tools offered in the literature. Each of the above reviewed models or methods has its own limitations. In practice, all kinds of evidence should be combined and evaluated together with the specific characteristics of component studies included in the meta-analysis.

Received 11 August 2019

REFERENCES

[1] Bellamy, L., Casas, J.-P., Hingorani, A. D. and Williams, D. (2009). Type 2 diabetes mellitus after gestational diabetes: a systematic review and meta-analysis. *The Lancet* **373** 1773–1779.

[2] Bhaumik, D. K., Amatya, A., Normand, S.-L. T., Greenhouse, J., Kaizar, E., Neelon, B. and Gibbons, R. D. (2012). Meta-analysis of rare binary adverse event data. *Journal of the American Statistical Association* **107** 555–567. MR2980067

[3] Biggerstaff, B. J. and Jackson, D. (2008). The exact distribution of Cochran's heterogeneity statistic in one-way random effects meta-analysis. *Statistics in Medicine* **27** 6093–6110. MR2522312

[4] Biggerstaff, B. and Tweedie, R. (1997). Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in medicine* **16** 753–768.

[5] Chung, Y., Rabe-Hesketh, S. and Choi, I.-H. (2013). Avoiding zero between-study variance estimates in random-effects meta-analysis. *Statistics in Medicine* **32** 4071–4089. MR3102435

[6] Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A. and Liu, J. (2013). A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika* **78** 685–709. MR3110938

[7] Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics* **10** 101–129. MR0067428

[8] Crippa, A., Khudyakov, P., Wang, M., Orsini, N. and Spiegelman, D. (2016). A new measure of between-studies heterogeneity in meta-analysis. *Statistics in Medicine* **35** 3661–3675. MR3538039

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

[10] DerSimonian, R. and Laird, N. (1986). Meta-analysis in linical trials. *Controlled Clinical Trials* **7** 177–188.

[11] Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* 54–75. MR0833275

[12] Farebrother, R. (1984). Algorithm AS 204: the distribution of a positive linear combination of χ^2 random variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **33** 332–339.

[13] Gart, J. J. (1966). Alternative analyses of contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)* 164–179. MR0202241

[14] Hardy, R. J. and Thompson, S. G. (1996). A likelihood approach to meta-analysis with random effects. *Statistics in Medicine* **15** 619–629.

[15] Hartung, J. and Knapp, G. (2005). On confidence intervals for the among-group variance in the one-way random effects model with unequal error variances. *Journal of Statistical Planning and Inference* **127** 157–177. MR2103031

[16] Hartung, J. and Makambi, K. (2002). Positive estimation of the between-study variance in meta-analysis: theory and methods. *South African Statistical Journal* **36** 55–76. MR1961402

[17] Hedges, L. V. and Olkin, I. (2014). *Statistical methods for meta-analysis*. Academic press. MR0798597

[18] Higgins, J. P. and Green, S. (2011). *Cochrane handbook for systematic reviews of interventions* **4**. John Wiley & Sons.

[19] Higgins, J. and Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* **21** 1539–1558.

[20] Hunter, J. E. and Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage.
[21] Jackson, D. (2013). Confidence intervals for the between-study variance in random effects meta-analysis using generalised Cochran heterogeneity statistics. *Research Synthesis Methods* **4** 220–229.
[22] Jackson, D., Bowden, J. and Baker, R. (2015). Approximate confidence intervals for moment-based estimators of the between-study variance in random effects meta-analysis. *Research Synthesis Methods* **6** 372–382.
[23] Jackson, D. and Bowden, J. (2016). 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* **16** 118.
[24] Jackson, D., White, I. R. and Riley, R. D. (2012). Quantifying the impact of between-study heterogeneity in multivariate meta-analyses. *Statistics in Medicine* **31** 3805–3820. MR3041775
[25] Knapp, G., Biggerstaff, B. J. and Hartung, J. (2006). Assessing the amount of heterogeneity in random-effects meta-analysis. *Biometrical Journal* **48** 271–285. MR2224258
[26] Knapp, G. and Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine* **22** 2693–2710.
[27] Kontopantelis, E., Springate, D. A. and Reeves, D. (2013). A re-analysis of the Cochrane Library data: the dangers of unobserved heterogeneity in meta-analyses. *PloS One* **8** e69930.
[28] Langan, D., Higgins, J. and Simmonds, M. (2017). Comparative performance of heterogeneity variance estimators in meta-analysis: a review of simulation studies. *Research Synthesis Methods* **8** 181–198.
[29] Langan, D., Higgins, J. P., Jackson, D., Bowden, J., Veroniki, A. A., Kontopantelis, E., Viechtbauer, W. and Simmonds, M. (2019). A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Research Synthesis Methods* **10** 83–98.
[30] Li, L. and Wang, X. (2019). Meta-analysis of rare binary events in treatment groups with unequal variability. *Statistical Methods in Medical Research* **28** 263–274. MR3894527
[31] Lin, L., Chu, H. and Hodges, J. S. (2017). Alternative measures of between-study heterogeneity in meta-analysis: Reducing the impact of outlying studies. *Biometrics* **73** 156–166. MR3632361
[32] Malzahn, U., Böhning, D. and Holling, H. (2000). Nonparametric estimation of heterogeneity variance for the standardised difference used in meta-analysis. *Biometrika* **87** 619–632. MR1789813
[33] Morris, C. N. (1983). Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association* **78** 47–55. MR0696849
[34] Novianti, P. W., Roes, K. C. and van der Tweel, I. (2014). Estimation of between-trial variance in sequential meta-analyses: a simulation study. *Contemporary Clinical Trials* **37** 129–138.
[35] Panityakul, T., Bumrungsup, C. and Knapp, G. (2013). On estimating residual heterogeneity in random-effects meta-regression: a comparative study. *Journal of Statistical Theory and Applications* **12** 253. MR3190282
[36] Paule, R. C. and Mandel, J. (1982). Consensus values and weighting factors. *Journal of Research of the National Bureau of Standards* **87** 377–385.
[37] Petropoulou, M. and Mavridis, D. (2017). A comparison of 20 heterogeneity variance estimators in statistical synthesis of results from studies: a simulation study. *Statistics in Medicine* **36** 4266–4280. MR3721131
[38] Rukhin, A. L. (2013). Estimating heterogeneity variance in meta-analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75** 451–469. MR3065475
[39] Sidik, K. and Jonkman, J. N. (2005). Simple heterogeneity variance estimation for meta-analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **54** 367–384. MR2135880
[40] Sidik, K. and Jonkman, J. N. (2007). A comparison of heterogeneity variance estimators in combining results of studies. *Statistics in Medicine* **26** 1964–1981. MR2364286
[41] Smith, T. C., Spiegelhalter, D. J. and Thomas, A. (1995). Bayesian approaches to random-effects meta-analysis: A comparative study. *Statistics in Medicine* **14** 2685–2699.
[42] Takkouche, B., Cadarso-Suarez, C. and Spiegelman, D. (1999). Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis. *American Journal of Epidemiology* **150** 206–215.
[43] Tian, L. (2008). Inferences about the between-study variance in meta-analysis with normally distributed outcomes. *Biometrical Journal* **50** 248–256. MR2420267
[44] van Aert, R. C. and Jackson, D. (2018). Multistep estimators of the between-study variance: The relationship with the Paule-Mandel estimator. *Statistics in Medicine* **37** 2616–2629. MR3824520
[45] van Aert, R. C., van Assen, M. A. and Viechtbauer, W. (2019). 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* **10** 225–239.
[46] Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J., Langan, D. and Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods* **7** 55–79.
[47] Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics* **30** 261–293. MR2717217
[48] Viechtbauer, W. (2007a). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine* **26** 37–52. MR2312698
[49] Viechtbauer, W. (2007b). Hypothesis tests for population heterogeneity in meta-analysis. *British Journal of Mathematical and Statistical Psychology* **60** 29–60. MR2318199

Chiyu Zhang
Department of Statistical Science
Southern Methodist University
Dallas, TX 75205
USA
E-mail address: zhchyvictor@gmail.com

Min Chen
Department of Mathematical Sciences
University of Texas at Dallas
Dallas, TX 75080
USA
E-mail address: mchen@utdallas.edu

Xinlei Wang
Department of Statistical Science
Southern Methodist University
Dallas, TX 75205
USA
E-mail address: swang@smu.edu