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

Running title: A comparison of heterogeneity variance estimators

Dean Langan (d.langan@ucl.ac.uk) (corresponding author) [1 7]

Julian PT Higgins (julian.higgins@bristol.ac.uk) [2]

Dan Jackson (daniel.jackson1@astrazeneca.com) [3]

Jack Bowden (jack.bowden@bristol.ac.uk) [2]

Areti Angeliki Veroniki (veronikia@smh.ca) [4]

Evangelos Kontopantelis (e.kontopantelis@manchester.ac.uk) [5]

Wolfgang Viechtbauer (wolfgang.viechtbauer@maastrichtuniversity.nl) [6]

Mark Simmonds (mark.simmonds@york.ac.uk) [7]

[1] Great Ormond Street Institute of Child Health, UCL, London, WC1E 6BT, UK

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

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

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

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

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

[7] Centre for Reviews and Dissemination, University of York, York, YO10 5DD, UK

Abstract

Studies combined in a meta-analysis often have differences in their design and conduct that can lead to heterogeneous results. A random-effects model accounts for these differences in the underlying study effects, which includes a heterogeneity variance parameter. The DerSimonian-Laird method is often used to estimate the heterogeneity variance, but simulation studies have found the method can be biased and other methods are available. This paper compares the properties of nine different heterogeneity variance estimators using simulated meta-analysis data. Simulated scenarios include studies of equal size and of moderate and large differences in size. Results confirm that the DerSimonian-Laird estimator is negatively biased in scenarios with small studies, and in scenarios with a rare binary outcome. Results also show the Paule-Mandel method has considerable positive bias in meta-analyses with large

differences in study size. We recommend the method of restricted maximum likelihood (REML) to estimate the heterogeneity variance over other methods. However, considering that meta-analyses of health studies typically contain few studies, the heterogeneity variance estimate should not be used as a reliable gauge for the extent of heterogeneity in a meta-analysis. The estimated summary effect of the meta-analysis and its confidence interval derived from the Hartung-Knapp-Sidik-Jonkman method is more robust to changes in the heterogeneity variance estimate and shows minimal deviation from the nominal coverage of 95% under most of our simulated scenarios.

Keywords

Heterogeneity, simulation, random-effects, DerSimonian-Laird, REML

1 Introduction

Meta-analysis is the statistical technique of combining the results of multiple comparable studies. These studies often have differences in their design and conduct that lead to heterogeneity in their underlying effects. When heterogeneity is thought to be present, researchers should first attempt to find its causes, but these causes may be too numerous to isolate or may simply be unknown. Unexplained heterogeneity of study effects can be quantified in a random-effects model. This model typically assumes a normal distribution of the underlying effects across studies. A reliable estimate of the variance of this distribution can provide valuable insight into the degree of heterogeneity between studies, even if such studies are not formally synthesised in a meta-analysis.

The moment-based method proposed by DerSimonian-Laird method (DerSimonian and Laird, 1986) is most commonly used to estimate the heterogeneity variance. However, this method has been shown in previous simulation studies to be negatively biased in meta-analyses containing small studies (Malzahn et al., 2000), particularly in meta-analyses of binary outcomes (Novianti et al., 2014; Sidik and Jonkman, 2007). There are many other available methods (Veroniki et al., 2015), including those proposed by Paule and Mandel (1982), Hartung and Makambi (2003), Sidik and Jonkman (2005, 2007), and the restricted maximum likelihood method (REML) (Harville, 1977). Estimates derived from these methods in the same meta-analysis can often be notably different and in a small number of cases, these estimates can produce discordant conclusions on the summary effect and its confidence interval (Langan et al., 2015). Therefore, the choice of heterogeneity variance method is an important consideration in a meta-analysis. Research based on simulated meta-analysis data can allow a researcher to make a more informed decision.

A recent systematic review collated simulation studies that compare the properties of heterogeneity variance estimators (Langan et al., 2016). Its aim was to assess if there is consensus on which heterogeneity variance methods (if any) have better properties than DerSimonian-Laird. The review identified 12 relevant simulation studies, but there was little consensus across the various authors' recommendations (Malzahn et al., 2000; Novianti et al.,

2014; Sidik and Jonkman, 2005; Sidik and Jonkman, 2007; Panityakul et al., 2013; Viechtbauer, 2005; Rukhin et al., 2000; Bhaumik et al., 2012; Knapp and Hartung, 2003; Sanchez-Meca and Marin-Martinez, 2008; Kontopantelis et al., 2013; Chung et al., 2013). This may have been caused by a potential conflict of interest among the authors of all but four of these studies (Novianti et al., 2014; Panityakul et al., 2013; Viechtbauer, 2005; Sanchez-Meca and Marin-Martinez, 2008); the authors of these eight studies recommended their own newly proposed methods over existing methods. Three of the simulation studies (Novianti et al., 2014; Panityakul et al., 2013; Viechtbauer, 2005) compared only pre-existing methods and made an explicit recommendation for estimating the heterogeneity variance; the authors of these studies recommended the method of Paule and Mandel (1982) and/or REML (Harville, 1977), but only compared a subset of methods.

The tentative conclusions of that review provided motivation for a new simulation study, which we present in this paper. The limitations of previous simulation studies helped inform the design of this study. We consider the inclusion of all methods identified in recent reviews of heterogeneity variance methods (Veroniki et al., 2015; Langan et al., 2016), compare methods comprehensively in a range of simulated scenarios representative of meta-analyses of health studies, and report a wide range of performance measures. Performance measures include those that relate directly to the heterogeneity variance estimates, and those that measure the impact of heterogeneity variance estimates on the summary effect estimate and its confidence interval. Our recommendations are based on a subjective trade-off between many performance measures. To minimise any conflict of interest, we do not propose any new methods in this paper.

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

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

2 Methods

2.1 The heterogeneity variance parameter in a random-effects model

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

$$\hat{\theta}_i = \theta_i + \varepsilon_i$$

$$\theta_i = \theta + \delta_i \,, \tag{1}$$

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

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

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

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

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

2.2 Heterogeneity variance estimators

Nine estimators were identified from two systematic reviews of heterogeneity variance methods (Veroniki et al., 2015; Langan et al., 2016). Estimators proposed by Hunter and Schmidt (2004), Rukhin (2000), Malzahn et al. (2000) and the maximum likelihood method proposed by Hardy and Thompson (1996) are present in these reviews but excluded from the main results because preliminary analysis showed they are clearly inferior to other methods (as shown in appendix 1). Furthermore, Bayesian methods that rely on a subjective choice of prior distribution are excluded because of difficulty in objectively comparing them to frequentist methods. The method proposed by Morris (1983) is excluded because it is an approximation to REML. We excluded the positive DerSimonian-Laird estimator (Kontopantelis et al., 2013), which truncates heterogeneity variance estimates below 0.01, because any positive cut-off value could be applied.

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

Table 1: Nine heterogeneity variance estimators included in this simulation study

Method of moments approach (estimators 1-5)

Five estimators included in this study can be derived from the method of moments approach, which is based on the generalised Q-statistic (DerSimonian and Kacker, 2007):

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

The weight assigned to study *i* is denoted by a_i and calculated differently depending on which of the five method of moments estimators is used. $\hat{\theta}$ is given by formula (2) with study weights $w_i = a_i$. By equating Q_{MM} to its expected value, the following general formula for the heterogeneity variance can be derived (see DerSimonian and Kacker (2007)) for a detailed derivation):

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

1. The DerSimonian-Laird estimator (DL) (DerSimonian and Laird, 1986) uses the fixed-effect model weights $a_i = 1/\hat{\sigma}_i^2$, which leads to the formula:

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

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

$\hat{\tau}_{CA}^2 = max\left\{0, \frac{1}{k-1}\sum_{i=1}^{k}\left(\hat{\theta}_i - \hat{\theta}_{CA}\right)^2 - \frac{1}{k}\sum_{i=1}^{k}\hat{\sigma}_i^2\right\}$, where $\hat{\theta}_{CA}$ is calculated from formula (2) with study weights $w_i = 1/k$.

3. The Paule-Mandel estimator (PM) uses the random-effects model study weights, defined by substituting $a_i = 1/(\hat{\sigma}_i^2 + \hat{\tau}_{PM}^2)$ into formula (3). Since a_i is a function of $\hat{\tau}_{PM}^2$, there is no closed-form expression for $\hat{\tau}_{PM}^2$ and iteration is required to find the solution. Iterative algorithms including those suggested by Bowden et al. (2011) and Jackson et al. (2014) always converge. The same estimator has been derived independently of the methods of moments approach and is therefore often referred to as the empirical Bayes estimator in the literature (Rukhin, 2013).

4. The two-step Cochran's ANOVA estimator also uses Paule-Mandel random-effects weights but restricts iteration to two-steps (PM_{CA}). Cochran's ANOVA is used to initially estimate τ^2, thus, a closed form expression can be derived by substituting $a_i = 1/(\hat{\sigma}_i^2 + \hat{\tau}_{CA}^2)$ into formula (3).

5. The two-step DerSimonian-Laird estimator (PM_{DL}) has similar weights as PM_{CA} above, but uses the DerSimonian-Laird method to calculate an initial estimate of τ^2. Therefore the study weights are $a_i = 1/(\hat{\sigma}_i^2 + \hat{\tau}_{DL}^2)$.

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

Hartung-Makambi (estimator 6)

Hartung and Makambi (2003) proposed a correction to the DerSimonian-Laird estimator so that $\hat{\tau}^2$ is always positive and truncation is not required. The formula is given by:

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

Sidik-Jonkman (estimators 7 and 8)

Sidik and Jonkman (2005) proposed the following two-step estimator that only produces positive τ^2 estimates:

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

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

Sidik and Jonkman (2005) noted that an alternative formula for $\hat{\tau}_0^2$ may lead to an estimator with better properties. In a subsequent paper (2007), they proposed an alternative initial estimate $\hat{\tau}_0^2 = max\{0.01, \hat{\tau}_{CA}^2\}$, where $\hat{\tau}_{CA}^2$ is Cochran's ANOVA estimate of the heterogeneity variance (estimator 2).

Restricted maximum likelihood (estimator 9)

To derive the restricted maximum likelihood (REML) estimator, the log-likelihood function from the random-effects model (1) derived from the maximum likelihood method (Hardy and Thompson, 2004) is transformed so that it excludes the parameter θ (Harville, 1977). In doing so, REML avoids making the assumption that θ is known and is therefore thought to be an improvement on the original maximum likelihood estimator (Viechtbauer, 2005). This results in the following modified log-likelihood function:

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

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

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

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

The heterogeneity variance estimate is calculated through a process of iteration. Fisher's scoring algorithm is used for iteration in this study, as implemented in the *metafor* package in R (Viechtbauer, 2010).

2.3 Confidence interval methods for the summary effect

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

A Wald-type confidence interval can be calculated as (DerSimonian and Laird, 1986):

$$\hat{\theta} \pm \mathrm{Z}_{(1-C)/2}\sqrt{Var(\hat{\theta})}$$

$$Var(\hat{\theta}) = 1/\left(\sum_{i=1}^{k} 1/(\hat{\sigma}_i^2 + \hat{\tau}^2)\right) \quad (4)$$

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

Alternatively, a t-distribution can be assumed for the summary effect with $k-1$ degrees of freedom (Follmann and Proschan, 1999):

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

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

The Hartung-Knapp-Sidik-Jonkman method (HKSJ) (Hartung and Knapp, 2001; Sidik and Jonkman, 2002) also relies on a t-distribution and uses an alternative weighted variance for $\hat{\theta}$:

$$\hat{\theta} \pm t_{k-1,\,(1-C)/2}\sqrt{Var_{HKSJ}(\hat{\theta})}$$

$$Var_{HKSJ}(\hat{\theta}) = \frac{\sum_{i=1}^{k} a_i\left(\hat{\theta}_i - \hat{\theta}\right)^2}{(k-1)\sum_{i=1}^{k} a_i},$$

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

This method is equivalent to the t-distribution method above, but its variance is multiplied by a scaling factor $\sum_{i=1}^{k} a_i\left(\hat{\theta}_i - \hat{\theta}\right)^2/(k-1)$ (Sidik and Jonkman, 2002; Wiksten et al., 2016). In certain cases, this scaling factor can be less than one, which leads to a narrower confidence

interval than the standard t-distribution approach and can also lead to a narrower interval compared to the Wald-type method in few cases (Higgins and Thompson, 2002). A variation of this method has been proposed to deal with this by constraining the scaling factor to be ≥ 1 (Hartung and Makambi, 2003). However, throughout this study, the HKSJ method without constraint is used.

3 Simulation study design

All simulations and analyses were carried out in R version 3.2.2. The package *metafor* (Viechtbauer, 2010) was used to run simulated meta-analyses and calculate heterogeneity variance estimates from methods coded in this package, bespoke code was used for those that are not. A study protocol was agreed by all authors before running these simulations and is available upon request from the first author.

3.1 Simulation methods

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

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

(1) Generate n_{1i} observations from $N(0, \sigma_{1i}^2)$ and n_{2i} observations from $N(\theta_i, \sigma_{2i}^2)$. We assume variances σ_{1i}^2 and σ_{2i}^2 in the two groups are equal and, without loss of generality, set them equal to 1.

(2) Calculate the sample means and standard deviations of these observations.

(3) Calculate $\hat{\theta}_i$ and $\hat{\sigma}_i^2$ for standardised mean differences by Hedges' g method, thus accounting for small sample bias of standardised mean differences (Borenstein et al., 1999, equations 2.23 and 2.24).

For studies with an odds ratio outcome measure:

(1) Generate an average event probability between the two study groups ($\bar{p}_i$) from one of the distributions as defined in section 3.2. Although this simulation approach is not common, Smith et al. (1995) has previously defined a Bayesian meta-analysis model that included the same $\bar{p}_i$ parameter.

(2) Derive underlying event probabilities for each study group (p_{1i} and p_{2i}) from the solutions to the following simultaneous equations:

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

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

(3) Simulate cell counts of the 2×2 contingency table from the distributions $Bin(n_{1i}, p_{1i})$ and $Bin(n_{2i}, p_{2i})$. Apply a continuity correction of 0.5 to studies with zero cell counts.
(4) Calculate $\hat{\theta}_i$ and $\hat{\sigma}_i^2$ for log odds ratios from the standard formulae in Borenstein et al. (1999).

3.2 Parameter values

Parameter values are chosen to represent the range of values observed in published meta-analyses in the Cochrane Database of Systematic Reviews (Langan et al., 2015) and based on parameter values from previous simulation studies (Langan et al., 2016). For all combinations of parameter values as outlined in this section, 5000 meta-analyses are simulated. Binary outcome meta-analyses are simulated with log-odds ratios of $\theta = \{0, 0.5, 1.1, 2.3\}$ (corresponding to odds ratios of 1, 1.65, 3, and 10). Standardised mean difference meta-analyses are simulated with $\theta = 0.5$ only, because previous simulation studies suggest θ has no noticeable effect on any of the results (Viechtbauer, 2005; Sanchez-Meca and Marin-Martinez, 2008). Sample sizes are generated from the following five distributions to represent meta-analyses containing small, small-to-medium, medium, large, and small and large studies: (1) $N_i = 40$, (2) $N_i \sim U(40, 400)$, (3) $N_i = 400$, (4) $N_i \sim U(2000, 4000)$, and (5) $N_i = 40$ (small) in half of studies and half selected from $N_i \sim U(2000, 4000)$ (large). If k is odd in the last scenario, one study is selected randomly (with probability 0.5) to be small or large. For odds ratio meta-analyses, the average event probability ($\bar{p}_i$) takes the values (1) 0.5, (2) 0.05, (3) 0.01, and (4) generated from the distribution $U(0.1, 0.5)$. Simulated meta-analyses contain 2, 3, 5, 10, 20, 30, 50, and 100 studies.

Heterogeneity variance parameter values (τ^2) are defined such that the resulting meta-analyses span a wide range of levels of inconsistency between study effects. We measured inconsistency using the I^2 statistic (Higgins and Thompson, 2002), an approximate measure of the relative size of the heterogeneity variance to the total variability in effect estimates (the sum of within-study error variance and between-study heterogeneity). The chosen τ^2 values result in meta-analyses with average I^2 values of 0%, 15%, 30%, 45%, 60%, 75%, 90%, and 95% and are given in appendix 2. I^2 values are calculated using the true τ^2 parameter estimates, but still vary between simulated meta-analyses because of the simulated variation in the standard errors. Parameter values for τ^2 vary between scenarios with different distributions for N_i and $\bar{p}_i$ to maintain a consistent range of I^2. In each scenario, τ^2 is fixed and I^2 varies between meta-analyses, therefore, we also present the range of I^2 next to the graphs in the results.

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

3.3 Performance measures

Properties of heterogeneity variance estimators are measured in terms of bias and mean squared error. These two measures are plotted proportional to the heterogeneity variance parameter

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

4 Results

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

4.1 Properties of heterogeneity variance parameter estimates

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

DerSimonian-Laird (DL)

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

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

Cochran's ANOVA (CA)

CA tends to produce higher estimates of the heterogeneity variance than most other estimators for both standardised mean difference and odds ratio meta-analyses. As such, CA is roughly unbiased in scenarios with high I^2 when most other estimators are negatively biased. However, CA is one of the most positively biased estimators for low to moderate I^2. CA's positive bias is particularly prominent in scenarios with small and large studies (figures 1 and 2, third column); it is counterintuitive to assign equal study weights (as the CA estimator does) in these scenarios with large differences in study size. CA also has higher mean squared error than most other estimators when the estimator is positively biased (figures 3 and 4).

Paule-Mandel (PM)

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

Two-step Cochran's ANOVA (PM_{CA})

PM_{CA} uses CA as an initial estimate of heterogeneity. PM_{CA}'s bias and mean squared error are equal to, or somewhere between, CA and PM in all scenarios. Given than CA and PM have high positive bias and large mean squared error in scenarios with small and large studies, so too does PM_{CA} (figures 1-4, third column).

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

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

Hartung-Makambi (HM)

In meta-analyses with small or small-to-medium study sizes and zero or low I^2, HM tends to produce relatively high estimates of heterogeneity and therefore has relatively high positive bias (figures 1 and 2, top-left). This is perhaps because HM is a transformation of the DL estimator that only produces positive estimates. HM tends to produce comparatively low estimates when I^2 is moderate or high and has more negative bias than DL in these scenarios. HM has a comparatively low mean squared error in all scenarios presented (figures 3 and 4), including scenarios where HM has high positive bias. HM is one of the best performing

estimators in meta-analyses containing small and large studies (figures 1-4, third column), with properties comparable with DL, PM_{DL}, and REML.

Sidik-Jonkman (SJ)

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

Alternative Sidik-Jonkman (SJ_{CA})

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

REML

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

Figure 1: Bias of heterogeneity variance estimates in standardised mean difference outcome meta-analyses.
Scenarios containing small studies (first row), small-to-medium studies (second row), and small and large studies (third row). Effect size $\theta = 0.5$. Note: the y-axis limits differ between plots.

Figure 2: Bias of heterogeneity variance estimates in odds ratio meta-analyses with underlying summary odds ratio 1.65 and an average event probability between 0.1 and 0.5
Scenarios containing small studies (first row), small-to-medium studies (second row), and small and large studies (third row). Effect size $\theta = 0.5$. Note: the y-axis limits differ between plots.

Figure 3: Mean squared error of heterogeneity variance estimates in standardised mean difference outcome meta-analyses.

Scenarios containing small studies (first row), small-to-medium studies (second row), and small and large studies (third row). Effect size $\theta = 0.5$. *Note: the y-axis limits differ between plots.*

Figure 4: Mean squared error of heterogeneity variance estimates in odds ratio meta-analyses with underlying summary odds ratio 1.65 and an average event probability between 0.1 and 0.5

Scenarios containing small studies (first row), small-to-medium studies (second row), and small and large studies (third row). Effect size $\theta = 0.5$. *Note: the y-axis limits differ between plots.*

4.2 Summary effect estimates

Results show that summary effect estimates ($\hat{\theta}$) are almost unbiased in all scenarios of standardised mean difference meta-analyses ($\theta = 0.5$) and odds ratio meta-analyses with common events. However, summary effect estimates are biased towards the null value of zero in odds ratio meta-analyses with rare events. This is likely to be partly a consequence of the choice of continuity correction (we added 0.5 to zero cell counts) and the degree of bias was similar across all heterogeneity variance estimators. We present bias of the summary effect in the supplementary results only.

4.3 Coverage of 95% summary effect confidence intervals

Coverage is presented in figure 5 for all combinations of heterogeneity variance estimators and (95%) Wald-type, t-distribution, and HKSJ confidence interval methods for the summary effect. Results are presented for standardised mean difference meta-analyses only, but results are consistent with the equivalent scenarios of odds ratio meta-analyses with common events (event probabilities 0.1 to 0.5, see appendix 3 in the supplementary results).

Wald-type method

Coverage of the 95% Wald-type confidence interval can differ by up to 5% between heterogeneity variance estimators, up to 30% between numbers of studies, and up to 20% between heterogeneity values. Coverage varies between 96-100% when studies are homogeneous and can be as low as 65% when the typical I^2 is 90% ($\tau^2 = 0.187$) and meta-analyses have two or three studies. When heterogeneity is present, the confidence interval's coverage tends towards the nominal value of 95% as the number of studies increases.

Standard t-distribution method

Coverage of the t-distribution 95% confidence interval is generally more robust to changes in the mean I^2, as shown in figure 5. In these scenarios, however, coverage can differ by up to 5% depending on the heterogeneity variance estimator used and the number of studies. When

there are 20 studies or more, 95% t-distribution confidence intervals have coverage 94-97%, but perform conservatively with coverages close to 100% when there are fewer than 20 studies. The heterogeneity variance estimator that works best with this confidence interval method varies considerably between scenarios, so it is difficult to select one overall.

Hartung-Knapp-Sidik-Jonkman (HKSJ) method

The HKSJ confidence interval for the summary effect has better coverage than the other two methods in all scenarios. This method has coverage 94-96% in standardised mean difference meta-analyses presented in figure 5 and is insensitive to the choice of heterogeneity variance estimator. The method can produce confidence intervals with sub-optimal coverage in odds ratio meta-analyses with rare events, where all meta-analysis methods perform poorly (as demonstrated in the supplementary results, appendix 4).

Figure 5: Coverage of 95% confidence intervals of the summary effect in standardised mean difference meta-analyses with small-to-medium studies ($N_i = U(40, 400)$)
Coverage of Wald-type (first row), t-distribution (second row), and HKSJ (third row) confidence intervals presented.

4.4 Generalisability of presented results

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

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

Second, results are not presented in the main paper from scenarios where all heterogeneity variance methods failed with considerable negative bias. This occurred in all scenarios of odds ratio meta-analyses with rare events (event probability = 0.05 and 0.01) except where study sizes were large (sample size >4000 per study). In these scenarios, summary effects were considerably biased and confidence interval methods also failed to produce reasonable coverage. For example, simulation results show that the HKSJ method can have coverage as low as 85% in odds ratio meta-analyses with small-to-medium sized studies with an underlying event probability of 0.05 (see appendix 4). Poor properties were perhaps observed in these scenarios because many studies contained zero events and a continuity correction was applied (0.5 was added to all 2x2 cell counts in these simulations). An alternative continuity correction may have produced different results.

Finally, results were presented thus far are from meta-analyses with typical I^2 values of 0%, 30%, 60%, and 90% (corresponding to four heterogeneity variance parameter values). Meta-

analyses with other typical I^2 values were simulated, but the four presented gave an adequate description of the properties of methods across all levels of inconsistency.

5 Discussion

The DerSimonian-Laird heterogeneity variance estimator is not recommended for widespread use in two-stage random-effects meta-analysis and therefore, should not be the default method for meta-analysis in statistical software packages; it produces estimates with more negative bias than most other methods in odds ratio meta-analyses with small studies or rare events and to a lesser extent in standardised mean difference meta-analyses with small studies. This finding can perhaps be explained by DerSimonian-Laird's fixed-effect study weights that are based solely on estimated within-study variances; these variances are imprecise and likely to be biased under such conditions. This observation is in agreement with previous simulation studies (Sidik and Jonkman, 2007; Panityakul et al., 2013), as identified in a systematic review (Langan et al., 2016). Viechtbauer (2005) and Böhning et al. (2002) stated that DerSimonian-Laird is unbiased when within-study variances are known. However, DerSimonian-Laird is one of the better performing estimators in meta-analyses with large differences in study size.

This simulation study identified three heterogeneity variance estimators with more reasonable properties; REML (Harville, 1977), Paule-Mandel (1982), and the two-step Paule-Mandel that uses a DerSimonian-Laird initial estimate (DerSimonian and Kacker, 2007). Paule-Mandel is often approximately unbiased when DerSimonian-Laird is negatively biased. However, results also show Paule-Mandel has high positive bias when there are large differences in study size. This can perhaps be attributed to the random-effects study weights used in this method, which can lead to small studies being given a relatively large weight under heterogeneous conditions. A similar issue regarding the use of random-effects study weights for summary effect estimation has been noted elsewhere (Higgins and Spiegelhalter, 2002). The two-step DerSimonian-Laird estimator (PM_{DL}) inherits most of the best properties of DerSimonian-Laird and Paule-Mandel methods and is simple to compute. REML has very similar properties to this two-step estimator and is already widely known, recommended in two previous simulation studies for meta-analyses with continuous (Novianti et al., 2014; Viechtbauer, 2005) and binary (Viechtbauer, 2005) outcomes. Furthermore, REML is already available in many statistical software packages (Viechtbauer, 2010; Kontopantelis and Reeves, 2010). Of those with reasonable properties, REML is the only estimator that assumes normality of effect sizes, but a previous simulation study (Kontopantelis and Reeves, 2012a; Kontopantelis and Reeves, 2012b) showed all these methods are reasonably robust under non-normal conditions.

One of the aims of this simulation study was to investigate when it is appropriate to rely on one estimate of the heterogeneity variance. Results show all estimators are imprecise and often fail to detect high levels of heterogeneity in meta-analyses containing fewer than ten studies. Furthermore, only 14% of meta-analyses in the Cochrane Database of Systematic Reviews contain ten studies or more (Langan et al., 2015), so it is rarely appropriate to rely on one estimate of heterogeneity in this setting. All estimators have poor properties even in meta-

analyses containing high numbers of studies when study sizes are small or the event of interest is rare.

Estimates of the summary effect and its HKSJ confidence interval are of less cause for concern, and perform well even in meta-analyses with only two studies. In particular, the HKSJ confidence interval offers a large improvement in coverage over the commonly used Wald-type confidence interval. However, caution must still be applied when dealing with meta-analysis datasets with rare events, where summary effects are biased and the HKSJ confidence interval method can have coverage as low as 85%. Summary effect estimates in this study were calculated using the inverse-variance approach, though the use of the Mantel-Haenszel method has been recommended for rare events (Kontopantelis et al., 2013; Bradburn et al., 2007) and may have improved properties in these scenarios. These findings agree with a previous simulation study (IntHout et al., 2014), in which the HKSJ method was compared with other confidence interval methods for both continuous and binary outcome measures. The results presented in this paper show the HKSJ method is robust to changes in the heterogeneity variance estimate.

Our findings do not concur with some previous simulation studies. In all cases, this can be attributed to differences in parameter values and other differences in simulation study design. The original estimator proposed by Sidik and Jonkman (2005) performed well in the author's own simulations, yet simulations in this study shows they have considerable positive bias in meta-analyses of up to moderate I^2. This was not observed by Sidik and Jonkman (2005) because simulated meta-analyses were only presented with high I^2 (Langan et al., 2016). The method of Paule-Mandel has been recommended based on the results of three previous simulation studies (Novianti et al., 2014; Panityakul et al., 2013; Bhaumik et al., 2012), but these studies did not simulate meta-analyses with moderate-to-large differences in study size, where Paule-Mandel has considerable positive bias. Novianti et al. (2014) only recommended REML for continuous outcome meta-analyses and observed a small negative bias when the outcome is binary and high I^2; this bias was less pronounced in our simulations with low to moderate I^2 that Novianti et al (2014) did not include in their simulations (Langan et al., 2016).

The limitations of this simulation study are as follows. First, only a subset of all confidence interval methods for the summary effect are included. Results show the HKSJ method is more robust than the Wald method to the choice of heterogeneity variance estimator, but no confidence interval method can be recommended solely from the results of this study. Other methods include the profile likelihood method (Hardy and Thompson, 1996), which has also been shown as a better alternative to the Wald method in simulated meta-analysis data (Henmi and Copas, 2010) and recommended elsewhere (Cornell, 2014). Second, a continuity correction of 0.5 was applied whenever simulated studies with a binary outcome contained zero events, but other methods with a better performance are available (Sweeting et al., 2004). This choice may have affected the results in scenarios where the event is rare (i.e. 0.05), but alternative continuity corrections are unlikely to have led to meaningful improvements where the event rate is extremely rare (i.e. 0.01) and all random-effects methods fail in terms of all performance measures. We assumed effects to be normally distributed and although this is a limitation, it has been shown that most of the investigated methods are robust even in extreme non-normal

distributions (Kontopantelis and Reeves. 2012a). Third, our analyses assume that all studies provide unbiased estimates of the true effects underlying them. In practice, results of studies may be biased if the studies are performed sub-optimally, and meta-analyses may be biased if studies are missing for reasons related to their results (e.g. due to publication bias). These biases can affect estimation of heterogeneity (both upwardly or downwardly) and lead to inappropriate conclusions. Finally, although the study aimed to simulate a comprehensive range of scenarios, this range could never be complete given how diverse meta-analyses are in practice; not all outcome measures were included (e.g. hazard ratios) and the distributions from which sample sizes were drawn in this study cannot be considered representative of all observed distributions because study sample sizes are unlikely to conform to a defined distribution.

We compared methods in the context of a classical two-stage meta-analysis where study effect estimates and their standard errors are calculated first, then combined at the second final stage. Alternatively, one-stage meta-analyses can be undertaken using individual participant data (IPD) using mixed modelling techniques; these raw data can be derived trivially from study-level 2x2 contingency tables for binary outcome meta-analyses (Stijnen et al., 2010; Simmonds and Higgins, 2016). Stijnen et al. (2010) explains that this approach makes random-effects meta-analyses more feasible with sparse data and does not require a continuity correction in case of zero events. Jackson et al. (2018) reviewed modelling approaches for this type of meta-analysis data and suggest these models can offer improved statistical inference on the summary effect. However, these models can present additional numerical issues given their complexity. Future work comparing the properties of heterogeneity variance methods between one-stage and two-stage binary outcome meta-analyses would be informative.

The HKSJ method is generally preferred over the Wald-type method. However, Wiksten et al. (Wilksten et al., 2016) showed it can occasionally lead to less conservative results, even when the Wald method uses a fixed-effect variance structure. Sidik and Jonkman (2007) proposed a modification to the HKSJ method to ensure the resulting confidence interval is at least as wide as the Wald-type fixed-effect confidence interval. We did not apply this modification in our study. A simulation study by Rover et al. (2015) found the modified method provides coverage closer to the nominal level when differences in study size were large.

Summarising the properties of a comprehensive list of heterogeneity variance estimators, compared over many combinations of parameter values was the biggest challenge of this study. By simulating meta-analyses from a wide range parameter values, inevitably there are scenarios that reflect meta-analyses rarely observed in practice. For example, most meta-analyses contain very few studies (Langan et al., 2015; Davey et al., 2011), but meta-analyses with up to 100 studies were simulated in order to show results over the full range of possible meta-analysis sizes. An attempt was made to focus more on the scenarios representative of real meta-analyses when interpreting results, but this was inevitably subjective.

6 Conclusion

A summary of our recommendations are given in table 2. The two-step DerSimonian-Laird estimator (PM_{DL}) and REML can often be biased, but overall have the most reasonable properties in standardised mean difference and odds ratio meta-analyses. Of these two estimators, REML is recommended on the basis of these results because it is already widely known, available in most statistical software packages, and consistent with the method commonly used for one-stage meta-analyses using individual participant data (Simmonds et al., 2015). The two-step DerSimonian-Laird estimator is recommended as an alternative if a simpler, non-iterative method is required.

The Hartung-Knapp-Sidik-Jonkman confidence interval for the summary effect is generally recommended over the standard t-distribution and Wald-type methods, particularly in binary outcome meta-analyses with rare events and the number of studies included is less than 20. To be consistent, we recommend the same REML estimate of the heterogeneity variance to calculate this confidence interval. However, this is inconsequential given how robust this confidence interval is to changes in the heterogeneity variance method in most scenarios.

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

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

7 References

Bhaumik DK, Amatya A, Normand SLT, et al 2012. Meta-analysis of rare binary adverse event data. *Journal of American Statistical Association*; 107(498) 555-567.

Böhning D, Malzahn U, Dietz E, et al 2002. Some general points in estimating heterogeneity variance with the DerSimonian-Laird estimator. *Biostatistics*; 3: 445-457. DOI: 10.1093/biostatistics/3.4.445

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

Bowden J, Tierney J, Copas A, et al 2011. Quantifying, displaying and accounting for heterogeneity in the meta-analysis of RCTs using standard and generalised Q statistics. *BMC Medical Research Methodology*; 11(1): 41.

Bradburn MJ, Deeks JJ, Berlin JA, et al 2007. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Statistics in medicine*; 26(1): 53-77.

Chung Y, Rabe-Hesketh S and Choi IH 2013. Avoiding zero between-study variance estimates in random-effects meta-analysis. *Statistics in Medicine*; 32(23): 4071-4089.

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

Cornell JE 2014. Random-effects meta-analysis of inconsistent effects: a time for change. *Annals of Internal Medicine*; 160(4): 267-270. DOI:10.7326/M13-2886

Davey J, Turner RM, Clarke MJ, et al 2011. Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: a cross-sectional, descriptive analysis. *BMC Medical Research Methodology*; 11(1). DOI: 10.1186/1471-2288-11-160

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

DerSimonian R, Kacker R 2007. Random-effects model for meta-analysis of clinical trials: An update. *Contemporary Clinical Trials*; 28(2): 105-114. DOI: 10.1016/j.cct.2006.04.004

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

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

Hartung, J 1999. An alternative method for meta-analysis. *Biometrical Journal*; 41, 901–916.

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

Hartung J and Makambi KH 2003. Reducing the number of unjustified significant results in meta-analysis. *Communications in Statistics - Simulation and Computation*; 32(4): 1179-1190. DOI: 10.1081/SAC-120023884

Harville DA 1977. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*; 72(358): 320-338. DOI: 10.2307/2286796

Henmi M and Copas JB 2010. Confidence intervals for random effects meta-analysis and robustness to publication bias. *Statistics in Medicine*; 29(29) 2969-2983. DOI: 10.1002/sim.4029

Higgins JP and Spiegelhalter DJ 2002. Being sceptical about meta-analyses: a Bayesian perspective on magnesium trials in myocardial infarction. *International Journal of Epidemiology*; 31: 96-104. DOI: 10.1093/ije/31.1.96

Higgins JPT and Thompson SG 2002. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*; 21(11): 1539-1558. DOI: 10.1002/sim.1186

Hunter J and Schmidt F 2004. Methods of Meta-Analysis: Correcting Error and Bias in Research Findings. SAGE Publications.

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

Jackson D, Turner R, Rhodes K, et al 2014. Methods for calculating confidence and credible intervals for the residual between-study variance in random effects meta-regression models. *BMC medical research methodology*; 14(1): 103.

Jackson D, Law M, Stijnen T, Viechtbauer W and White IR 2018 (in press). A comparison of 7 random-effects models for meta-analyses that estimate the summary odds ratio. *Statistics in Medicine.* DOI: 10.1002/sim.7588

Knapp G and Hartung J 2003. Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*; 22(17): 2693-2710.

Kontopantelis E and Reeves D 2010. metaan: Random-effects meta-analysis. *Stata Journal*; 10(3): 395.

Kontopantelis E and Reeves D 2012a. Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: a simulation study. *Statistical methods in medical research*; 21(4): 409-26.

Kontopantelis E, Reeves D 2012b. Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: A comparison between DerSimonian–Laird and restricted maximum likelihood. *Statistical methods in medical research*; 21(6): 657-9.

Kontopantelis E, Springate DA and Reeves D 2013. A re-analysis of the Cochrane library data: the dangers of unobserved heterogeneity in meta-analyses. *PloS One*; 8(7): e69930.

Langan D, Higgins JPT and Simmonds M 2015. An empirical comparison of heterogeneity variance estimators in 12,894 meta-analyses. *Research Synthesis Methods*; 6(2): 195-205. DOI: 10.1002/jrsm.1140

Langan D, Higgins JPT and Simmonds M 2016. Comparative performance of heterogeneity variance estimators in meta-analysis: a review of simulation studies. *Research Synthesis Methods;* 8(2): 181-198. DOI: 10.1002/jrsm.1198

Malzahn U, Böhning D and Holling H 2000. Nonparametric estimation of heterogeneity variance for the standardised difference used in meta-analysis. *Biometrika*; 87(3): 619-632. DOI: 10.1093/biomet/87.3.619

Morris CN 1983. Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistics Association*; 78(381): 47–55.

Novianti PW, Roes KCB and van der Tweel I 2014. Estimation of between-trial variance in sequential meta-analyses: A simulation study. *Contemporary Clinical Trials*; 37(1): 129-138. DOI: 10.1016/j.cct.2013.11.012.

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(3): 253-265.

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

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

Rukhin AL, Biggerstaff BJ and Vangel MG 2000. Restricted maximum likelihood estimation of a common mean and the Mandel-Paule algorithm. *Journal of Statistical Planning and Inference*; 83(2): 319-330. DOI:10.1016/S0378-3758(99)00098-1

Rukhin, A.L. 2013. Estimating heterogeneity variance in meta-analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75(3) 451-469.

Sanchez-Meca J and Marín-Martínez F 2008. Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychol Methods*; 13(1): 31.

Sidik K and Jonkman JN 2002. A simple confidence interval for meta-analysis. *Statistics in Medicine*; 21(21): 3153-3159. DOI: 10.1002/sim.1262

Sidik K and Jonkman JN 2005. Simple heterogeneity variance estimation for meta-analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*; 54(2): 367-384. DOI: 10.1111/j.1467-9876.2005.00489.x

Sidik K and Jonkman JN 2007. A comparison of heterogeneity variance estimators in combining results of studies. *Statistics in Medicine*; 26(9): 1964-1981. DOI: 10.1002/sim.2688

Simmonds M, Stewart G and Stewart L 2015. A decade of individual participant data meta-analyses: a review of current practice. *Contemporary clinical trials*; 45: 76-83.

736 Simmonds M and Higgins JPT 2016. A general framework for the use of logistic regression
737 models in meta-analysis. *Statistical methods in medical research*; 25(6): 2858-2877.

738 Smith TC, Spiegelhalter DJ and Thomas A 1995. Bayesian approaches to random-effects
739 meta-analysis: A comparative study. *Statistics in Medicine*; 14(24): 2685-2699

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

743 Sweeting MJ, Sutton AJ and Lambert PC 2004. What to add to nothing? use and avoidance of
744 continuity corrections in meta-analysis of sparse data. *Statistics in Medicine*; 23(9): 1351-
745 1375. DOI: 10.1002/sim.1761

746 Veroniki AA, Jackson D, Viechtbauer W, et al 2015. Methods to estimate heterogeneity
747 variance and its uncertainty in meta-analysis. *Research Synthesis Methods*; 7: 55-79. DOI:
748 10.1002/jrsm.1164

749 Viechtbauer W 2010. Bias and efficiency of meta-analytic variance estimators in the random-
750 effects model. *Journal of Educational and Behavioral Statistics*; 30(3): 261-293. DOI:
751 10.3102/10769986030003261

752 Viechtbauer W 2010. Conducting meta-analyses in R with the metaphor package. *Journal of*
753 *Statistical Software*; 36(3): 1-48.

754 Wiksten A, Rücker G, Schwarzer G 2016. Hartung–Knapp method is not always conservative
755 compared with fixed-effect meta-analysis. *Statistics in medicine*; 35: 2503-2515. DOI:
756 10.1002/sim.6879

757 ***Table 1: Nine heterogeneity variance estimators included in this study***

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

758

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

		OR outcome with average event probability:		
		0.05	***0.1 to 0.5***	***SMD outcome***
Study sizes:	***Small***	All estimators have substantial negative bias in the presence of heterogeneity. HKSJ confidence interval can have coverage too high/low for >20 studies (appendix 4).	REML/PM/PM_{DL} recommended, but all estimators biased/imprecise for <10 studies. HKSJ confidence interval yields the nominal coverage.	
	Small-to-medium			
	Small and large		REML/PM_{DL} and HKSJ confidence interval recommended (as above), but all heterogeneity variance estimators biased/imprecise for <10 studies. PM positively biased.	

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

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

See separate file for figure.

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

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

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

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

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

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

See separate file for figure.

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

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

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

See separate file for figure.