A comparison of 20 heterogeneity variance estimators in statistical synthesis of results from studies: A simulation study

Maria Petropoulou,[1] Dimitris Mavridis [1, 2] *†

When we synthesize research findings via meta-analysis, it is common to assume that the true underlying effect differs across studies. There is a plethora of estimation methods available for the between-study variability. The widely used DerSimonian and Laird estimation method has been challenged but knowledge for the overall performance of heterogeneity estimators is incomplete. We identified 20 heterogeneity estimators in the literature and evaluated their performance in terms of bias, type error I rate and power via a simulation study. Moreover, we compared the Knapp and Hartung and the Wald-type method for estimating confidence interval for the summary estimate. Although previous simulation studies have suggested the Paule-Mandel (PM) estimator, it has not been compared with all the available estimators. For dichotomous outcomes, estimating heterogeneity through Markov Chain Monte Carlo is a good choice if the prior distribution for heterogeneity is informed by published Cochrane reviews. Non parametric bootstrap (DLb) performs well for all assessment criteria for both dichotomous and continuous outcomes. The positive DerSimonian and Laird (DLp) and the Hartung-Makambi (HM) estimators can be an alternative choice for dichotomous outcome when the heterogeneity values are close to 0.07 and for continuous outcome for all and for medium heterogeneity values ($0.01, 0.05$), respectively. Hence, they are heterogeneity estimators (DLb; DLp) which perform better than the suggested PM. Maximum likelihood (ML) provide the best performance for both types of outcome in the absence of heterogeneity.

Keywords: bias; type I error; heterogeneity variance estimators; power; simulation study

1. Introduction

When synthesizing results from different studies via a meta-analysis model, it is likely that we will encounter two sources of variation. Summary estimates are expected to differ because of sampling variability within studies but we also expect that studies would differ in methodological (study design, risk of bias etc.) and clinical (participants, dosage, duration etc.) characteristics [1]. If these characteristics modify the effect of the intervention, the true underlying intervention effect would differ across studies. Between-study variation of the true underlying effect is commonly

[1]Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece
[2]Department of Primary Education, University of Ioannina School of Education, Ioannina, Greece
* Correspondence to: Dr. Dimitris Mavridis, Department of Primary Education, University of Ioannina School of Education, Ioannina, Greece
†E-mail: dmavridi@cc.uoi.gr,

called statistical heterogeneity or heterogeneity. Two approaches are typically employed to synthesize study findings via meta-analysis. The fixed-effect and the random-effects model, which will be described in detail in Section 2. In a nutshell, the former allows only for within study variation whereas the latter allows for between study variation [1, 3]. Within study variation, although unknown, is approximated through its sample estimate. Between-study variation (heterogeneity) is estimated using a plethora of methods. Veroniki *et al*. [4] gave a thorough review of the available methods to estimate heterogeneity.

In this study, we conduct a simulation study to assess the performance of 20 different estimators for heterogeneity variance in terms of bias, type error I rate and power. The different estimators are; DerSimonian and Laird (DL) [5], general Hedges-Olkin (GHO) [6], Paule-Mandel (PM) [7], positive DerSimonian and Laird (DLp) [8], two-step DerSimonian and Laird (DL2) [9], two-step General Hedges-Olkin (GHO2) [9], Hartung-Makambi (HM) [10, 11], Hunter-Schmidt (HS) [12], maximum likelihood (ML) [13, 14], restricted maximum likelihood (REML) [5, 15], approximate restricted maximum likelihood (AREML) [14, 15, 16], Sidik Jonkman (SJ) [17], alternative Sidik-Jonkman (SJgho) (the same estimator with Sidik-Jonkman with a GHO estimator for a priori estimate) [15], Rukhin Bayes (RB) [18], positive Rukhin Bayes (RBp) [18], fully Bayesian (FB) using Markov Chain Monte Carlo in WinBUGS [19], Bayes modal (BM) [20, 21], non-parametric bootstrap DerSimonian and Laird (DLb) [8], empirical Bayes (EB) [16] and Malzahn, Böhning and Holling (MBH) [22]. Moreover, we compare the Wald-type and the Knapp and Hartung method of confidence interval for μ with the use of the most well-known estimators DL, GHO, ML and REML.

Although the DerSimonian and Laird estimator is by far the most commonly used and the only option in RevMan [23], there is much criticism to this choice and several empirical and simulation studies have been recently conducted to address this issue. Bowden *et al*. [24] suggested the use of PM estimator as one estimation method with good properties, while Knapp and Hartung [25] found that PM is less sufficient than DL and REML. Kontopantelis *et al*. [8] suggested the use of DLb after the comparisons with DL, DLp, DL2, GHO2, REML, SJ, RB and RBp. Novianti *et al*. [26] compared the DL, GHO, SJ, SJgho, PM, REML and DL2 estimators and recommended DL2 and PM as the best choice for dichotomous and continuous outcomes. In addition, both Novianti *et al*. [26] and Viechtbauer [27] suggested the REML estimator as a preferable alternative to DL for continuous outcomes.

There are some individual conclusions of empirical studies in the literature that compare a subset of the estimators considered in this work. Veroniki *et al*. [4] provided recommendation based *on a qualitative evaluation of the existing literature and expert consensus* and suggested that a thorough simulation study is needed to provide evidence-based recommendation. This motivated us to conduct a simulation study where all the available heterogeneity variance methods would be compared under the same scenarios under a representative simulation design of systematic review. In this study, we identified 20 estimators and evaluated them in terms of bias, empirical type error I and power.

The paper is organized as follows: Section 2 provides the models commonly used to summarize evidence in meta-analysis, confidence intervals for summary estimates and

the various heterogeneity estimators; Section 3 describes the design of our simulation study; Section 4 presents the simulation study results for the two types of outcomes, dichotomous and continuous; Section 5 summarizes the findings of our simulation and finally Section 6 concludes the findings of the study.

2. Methods

There are two popular statistical models for the synthesis of study findings in a meta-analysis, the fixed effect (FE) model and the random-effects (RE) model [1, 3]. Although only the latter model is of interest in this project, we will present the mathematical details of both models because that will help us explain some of the estimators presented in Section 2.2. Fixed-effect meta-analysis, assumes that all studies estimate a common true effect size μ (fixed effect) and variability between the observed effect sizes is due to sampling error. Let us consider a group of k studies from the same population where each study i population variance ${\sigma_i}^2$ and observed effect estimates $y_i, i = 1, \ldots . k$ from which we wish to estimate the overall true mean effect μ. Therefore, under the fixed effect model, the effect size will vary across studies due to error attributed to within-study variability:

$$y_i = \mu + \varepsilon_i \quad , \quad \varepsilon_i \sim N(0, {\sigma_i}^2)$$

Each study is weighted by the inverse of its variance. In the fixed effect model, variation is equal to the population variance. In practice, population variance is unknown and we approximate it with the within-study sample variance estimators s_i^2. The weights assigned to each study are

$$w_{i,FE} = \frac{1}{s_i^2}$$

, the overall mean is the weighted average of the observed effect sizes

$$\hat{\mu}_{FE} = \frac{\sum_{i=1}^{k} w_{i,FE} y_i}{\sum_{i=1}^{k} w_{i,FE}}$$

with variance equal to the inverse of the sum of the weights of the meta-analysis

$$\hat{V}_M = \frac{1}{\sum_{i=1}^{k} w_{i,FE}}$$

Under the random effects model, the effect size will vary across studies due to within $({\sigma_i}^2)$ and between (τ^2) study variability. This can be written as a hierarchical model:

$$y_i = \theta_i + \varepsilon_i \quad , \ \varepsilon_i \sim N(0, {\sigma_i}^2)$$

$$\theta_i = \mu + \xi_i \quad , \ \xi_i \sim N(0, \tau^2)$$

In practice, variability is unknown and we use estimates for the within (s_i^2) and between $(\hat{\tau}^2)$ study variation. The weights for the random effects model are now defined as

$$w_{i,RE} = \frac{1}{s_i^2 + \hat{\tau}^2}$$

and the overall mean is the weighted average of the observed effect sizes

$$\hat{\mu}_{RE} = \frac{\sum_{i=1}^{k} w_{i,RE} y_i}{w_{i,RE}}$$

with variance equal to the inverse of the sum of the weights of the meta-analysis [1]:

$$\hat{V}_M = \frac{1}{\sum_{i=1}^{k} w_{i,RE}}$$

2.1 Confidence intervals for summary estimate $\boldsymbol{\mu}$

Uncertainty around the summary estimate is expressed by $(1-\alpha)\%$ confidence intervals. There are two types of intervals commonly used:

(a) Wald-type (Wt) confidence interval

The Wald-type method is the most popular technique for establishing confidence intervals for a parameter of interest. DerSimonian and Laird [5] used Wald-type confidence intervals for expressing uncertainty for summary estimate μ in the meta-analysis. Assuming that the summary estimate is asymptotically normally distributed, a $(1-\alpha)\%$ confidence interval is given by

$$\hat{\mu}_{RE} \pm z_{a/2}\sqrt{V_M}$$

(b) Knapp and Hartung (KH) confidence interval

Hartung [28] and Knapp and Hartung [25] suggested the use of the t-distribution with $k-1$ degrees of freedom for constructing confidence interval for the summary estimate μ. Sidik and Jonkman [29] independently developed this approach. The motivation behind this approach is that the normal approximation is more liberal for meta-analysis with few studies and large heterogeneity and the t-distribution has heavier tails than the normal one [30]. Hartung [28] showed that

$$\frac{\hat{\mu}_{RE} - \mu}{\sqrt{Var(\hat{\mu}_{RE})}} \sim t_{k-1}$$

with $Var(\hat{\mu}_{RE}) = Q_{gen}\frac{1}{(k-1)\sum_{i=1}^{k} w_{i,RE}}$, where Q_{gen} is a generalised Q-statistic which quantifies the amount of between-study variance and it is given by $Q_{gen} = \sum_i w_{i,RE}(y_i - \hat{\mu}_{RE})^2$ [28]. Therefore, a $(1-\alpha)\%$ confidence interval for summary estimate μ can be described as [28]:

$$\hat{\mu}_{RE} \pm t_{k-1,a/2}\sqrt{Var(\hat{\mu}_{RE})}$$

2.2 Estimators for $\boldsymbol{\tau^2}$

Estimation of between-study variance is a necessary step to quantify uncertainty around the summary estimate in random-effects meta-analysis. A plethora of estimators involving both non-iterative (closed form expressions) and iterative (e.g. maximum likelihood) methods, have been suggested. Some of the estimators yield only positive values (positive estimators) whereas others may give negative values (non-negative estimators) that are subsequently truncated to zero.

Table 1 lists, to our knowledge, all the heterogeneity estimators that have been developed. A comprehensive review and details for heterogeneity estimators can be

found in Veroniki *et al*. [4]. In this study, we consider the 19 estimators reviewed in Veroniki *et al*. [4] plus the non-parametric Malzahn, Böhning and Holling (MBH) estimator for continuous outcomes. We provide details for the MBH estimator. For more details for the remaining estimators, you may resort to the corresponding references or to Veroniki *et al*. [4].

Malzahn, Böhning and Holling [22] proposed a nonparametric estimator of the population heterogeneity variance which can be used only when effect sizes are expressed as standardized mean differences using Hedges' g. More specifically, it is given by

$$\hat{\tau}^2_{MBH} = \frac{\sum_{i=1}^{k}(1-\varphi_i)(y_i-\hat{\mu}_{FE})^2}{k-1} - \frac{1}{k}\sum_{i=1}^{k}\left(\frac{n_i}{n_{it}n_{ic}}\right) - \frac{1}{k}\sum_{i=1}^{k}\varphi_i y_i^2$$

where $n_i = n_{ic} + n_{it}$ is the total size for i study, n_{ic} the total size for control group, n_{it} the total size for treatment group and φ_i is given by $\varphi_i = 1 - \frac{n_i-4}{J^2(n_i-2)}$ with the correction J_i for unbiased estimators with type $J_i = 1 - \frac{3}{4(n_{ic}+n_{it})-9}$.

3. Simulation design

We conducted a simulation study to assess the performance of all the available heterogeneity variance estimators. We considered a range of dichotomous and continuous outcomes to represent observed values from a practical meta-analysis. The simulation design was similar to the one used in Novianti *et al*. [26]. In individual studies of randomized control trials (RCTs) the outcome is compared in two groups, the experimental group (T) and the control group (C). Assuming k number of trials included in a meta-analysis we follow the simulation of Novianti *et al*. [26] with the following steps:

- For study $i = 1, \dots k$, generate the treatment effect $\theta_i \sim N(\theta, \tau^2)$
- Generate within-study sample sizes, n_{it} for the treatment group and n_{ic} for the control group, assuming equal sample sizes, $n_i = n_{ic} = n_{it}$ generated from a discrete uniform assuming values from 20 to 200

3.1 Dichotomous outcome data

For dichotomous outcomes we need the number of events and non-events (or sample size) for each group in all studies. We considered the number of events α_i and c_i and the group sizes n_{it} and n_{ic} for experimental and control group, respectively. We simulated data with the study specific treatment effect being the logarithmic odds ratio (OR).

- Obtain the total number of events c_i for the control group from a $Binomial(n_{ic}, p_{ic})$ distribution with $p_{ic} \sim Uniform\ (0.05, 0.65)$
- Obtain the total number of events α_i for the treatment group from a $Binomial(n_{it}, p_{it})$ distribution with $p_{it} = \frac{p_{ic}e^{\theta_i}}{1-p_{ic}+p_{ic}e^{\theta_i}}$
- Calculate the total number of non-events for treatment group and control group $b_i = n_{it} - \alpha_i$, $d_i = n_{ic} - c_i$, respectively (If any number of events or non-events is zero, add the value 0.5)

- Calculate the treatment effect $y_i = log(OR)_i = log\left\{\frac{\alpha_i \times d_i}{b_i \times c_i}\right\}$ and the within-study variance $s_i^2 = \frac{1}{\alpha_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}$

3.2 Continuous outcome data

For continuous outcomes, the study specific effect size is estimated by the standardized mean difference Hedges' g which is defined as the standardized difference between the average in the treatment group and the average of control group.

- Simulate n_{ic} observations Z_{ic} for control group with $Z_{ic} \sim N(0,1)$
- Simulate n_{it} observations Z_{it} for treatment group with $Z_{it} \sim N(\theta_i, 1)$
- Calculate the pooled variance $S_{ip} = \sqrt{\frac{(n_c-1)S_{ic}^2 + (n_t-1)S_{it}^2}{n_c+n_t-2}}$ with the sample variances for control group and treatment group S_{ic}^2, S_{it}^2 , respectively
- Calculate the treatment effect $y_i = \frac{\bar{Z}_{it} - \bar{Z}_{ic}}{S_{ip}}$ and its within-study variance $S_i^2 = \frac{8+{y_i}^2}{4n_i} (n_i = n_{ic} = n_{it})$
- Use the correction $J_i = 1 - \frac{3}{8n_i - 9}$ to avoid biased estimators
- Calculate the unbiased estimate of the treatment effect $J_i \times y_i$ and the within-study variance $s_i^2 = {J_i}^2 \times S_i^2$

3.3 Scenarios

The simulation was performed in the freeware statistical software R [31], which allows estimation of most of the heterogeneity variance estimators via the metafor package [32]. PM estimator is actually equivalent method to the EB estimator [4]. Additionally, AREML estimator is an approximate of REML when the population variances are equal [27]. In practice, population variances are not equal and therefore AREML is excluded from simulation study. So, in the simulation study we included 17 heterogeneity estimators for dichotomous outcome and an additional estimator (MBH) for continuous outcome. We used the Wald-type confidence interval to obtain the summary estimate but for four of the estimators (DL, GHO, ML, REML) we additionally used the Knapp and Hartung method for constructing the confidence interval with the aim to compare the two methods for the confidence interval for μ (Wald-type and Knapp-Hartung methods). We focused on these four estimators as they are the most commonly used in practice.

We computed ten estimators (DL, GHO, HS, SJ, ML, REML, PM, DL2, GHO2, SJgho) with the metafor package [32]. In addition, metafor package was applied for the computation of the Wald-type and the Knapp and Hartung confidence intervals of μ. FB estimator was computed using Markov Chain Monte Carlo (MCMC) within the Bayesian software WinBUGS [19] via the R package of R2WinBUGS [33]. MCMC was conducted with a vague normal prior distribution for the summary estimate $\theta \sim N(0, 10^6)$, the log-normal $logN(-2.56, 1.74^2)$ prior distribution on the untransformed τ^2 scale for dichotomous outcome and the $log(\tau^2) \sim t(-3.44, 2.59^2, 5)$ for continuous outcomes. These distributions for τ^2 are the predictive distributions for a future meta-analysis in a general setting suggested by

Turner *et al*. [34] and Rhodes *et al*. [35], respectively. Moreover, we compared the above informative prior distributions with the $Uniform(0,100)$ prior distribution (vague prior distribution). We symbolized the FB estimator as FB_{vague} using a vague prior and as $FB_{informative}$ using the informative prior. We created R functions to compute seven estimators (DLp, DLb, HM, RBp, RBo, MBH and BM) which, to our knowledge, are not offered in any R package. The R and WinBUGS codes used in this simulation study are given in the supplementary material.

For a detailed comparison of the heterogeneity variance estimators, we included various values for the number of studies $k = 10, 20, 30, 50$ reflecting small, medium and large number of studies and different values for the true overall effect $\theta = 0, 0.3, 0.5, 0.8$ that represent a range of intervention effects that can be found in practice. We also assumed between-study heterogeneity values taking into account the empirical heterogeneity variance distributions [34, 35], $\tau^2 = 0, 0.025, 0.07, 0.3$ values for dichotomous outcomes and $\tau^2 = 0, 0.01, 0.05, 0.5$ values for continuous outcomes. Each scenario was repeated $B = 1000$ times with a significance level of $\alpha = 0.05$.

3.4 Assessment criteria

The performance of the heterogeneity variance estimators was assessed using three measures: (a) average absolute bias, (b) type error I, (c) power

(a) Average absolute bias

Bias is the difference between the expected value of the estimator and its true value, and is given by $bias(\hat{\tau}^2) = E(\hat{\tau}^2) - \tau^2 = E(\hat{\tau}^2 - \tau^2)$. Ideally, a good estimator must be unbiased which means that the expected value must be equal to its true value ($E(\hat{\tau}^2) = \tau^2$). In the simulation study, we used the average absolute bias, which is an empirical measure of bias and is given by

$$bias = \frac{1}{B}\sum_{i=1}^{B}\left|\hat{\tau}_i^2 - \tau^2\right|$$

where $\hat{\tau}_i^2$ is the observed heterogeneity variance estimator for the study i and τ^2 the true value of heterogeneity. The close the average absolute bias is to zero, the better the estimator is. The performance of simulation in terms of bias is conducted for all values of k, θ and τ^2.

(b) Empirical type error I

Given the null hypothesis $H_0: \theta = \theta_0$, we produced B simulated datasets and for each dataset we estimated the $(1-\alpha)\%$ confidence interval for θ_0, $\left[\hat{\theta}_{i,L}, \hat{\theta}_{i,U}\right], i = 1, \ldots, B$. Empirical type error I is defined as the number of times that the estimated interval for θ_0, did not include its true value θ_0.

$$empirical\ type\ error\ I = \frac{\#(\hat{\theta}_{i,L} >= \theta_0\ ||\ \hat{\theta}_{i,U} <= \theta_0\ )}{B}$$

A good estimator must have empirical type error I close to its nominal level. With level of significance a, a good estimator must minimize the empirical type error I and

its values must range in the confidence interval of $\left(\alpha - 1.96\sqrt{\frac{\alpha(1-\alpha)}{B}}, \alpha + 1.96\sqrt{\frac{\alpha(1-\alpha)}{B}}\right)$. With a level of significance of $\alpha = 0.05$ and $B = 1000$, an estimator has good performance if its empirical type error I is located within the interval $(L, U) = (0.0365, 0.0635)$. Empirical type error I has a direct relationship with the coverage probability for μ. This means that high values of empirical type error I are associated with more rejections of the null hypothesis than what we should have observed. This fact produces lower values in coverage probability for μ and therefore heterogeneity estimators have similar behavior in these two measures. We calculated the empirical type error I for all heterogeneity estimators in all scenarios.

(c) Power

Suppose that the null hypothesis is $H_0: \theta = \theta_0$ and the alternative is $H_\alpha: \theta = \theta_1$. We produced B simulated datasets under the alternative hypothesis and estimated the $(1-\alpha)\%$ confidence interval for θ_0, denoted as $\left[\hat{\theta}_{i,L}, \hat{\theta}_{i,U}\right], i = 1, \dots, B$ in each simulated sample. Power is defined as the number of times out of the B simulated datasets that θ_0 does not belong to the closed confidence interval for θ_0.

$$power = \frac{\#(\hat{\theta}_{i,L} >= \theta_0 \ || \ \hat{\theta}_{i,U} <= \theta_0)}{B}$$

Therefore, a good estimator provides high power. The performance of simulation in terms of power is conducted for all the values of k and τ^2 with the null hypothesis for overall mean $H_0: \theta = \theta_0 = 0$ and the alternatives $\theta_1 = 0.3, 0.5$.

4. Simulation Results

4.1 Dichotomous outcome data

(a) Average absolute bias

Figure 1 shows the performance of heterogeneity variance estimators in terms of bias for all values of k, θ and τ^2. In all cases, results show that bias increased with increasing heterogeneity τ^2 and decreased with increasing number of studies k. RBp estimator has the largest bias compared to other heterogeneity estimators in most cases. The SJ estimator has the second largest bias for heterogeneity values up to 0.07 and similar performance with the majority of estimators for larger heterogeneity values τ^2. In addition, HE, RBo, SJgho and BM present large values of bias. It is noteworthy that the HM has the lowest bias from all heterogeneity estimators for heterogeneity values close to 0.07. The $FB_{informative}$ estimator has comparative bias with DLb and lower values than the other estimators. In most cases, $FB_{informative}$ presents lower bias than DLb with the exception of a zero or high heterogeneity (e.g. $\tau^2 = 0.3$). Moreover, DLp might be an alternative choice to $FB_{informative}$ and DLb. The ML estimator has the lowest bias in the absence of heterogeneity followed by the HS estimator.

(b) Empirical type error I

As the number of trials increases, the closer the values of the empirical type error I are to its nominal level (Figure 2), which is assumed to be 0.05 with 95% confidence

interval $(L = 0.0365, U = 0.0635)$ denoted by dotted line. We found that RBp estimator has a much lower empirical type error I than the lower limit of the interval L for all scenarios. With the increase of heterogeneity, type error I of RBp estimator increases but still remains far lower than L. Whereas, only in case of $k = 10$ and $\tau^2 = 0.3$, RBp is located into the nominal interval. HM has empirical type error I into the nominal interval only for medium values of heterogeneity. SJ estimator has a much lower empirical type error I from L, for all cases when heterogeneity is less than or equal to 0.07. Though it ranges closer to permissible interval by increasing heterogeneity. The empirical type error I of RBo estimator is within the permissible interval in the absence of heterogeneity or when heterogeneity is 0.025 for all scenarios. For larger heterogeneity values, RBo estimator has empirical type error I larger than the allowed interval. Apart from the RBo, RBp and SJ listed above, all other estimators have similar behavior for the empirical type error I with values close to the interval (L, U). Moreover, it should be noted that the estimators from the Knapp and Hartung (DLknha, GHOknha, MLknha, REMLknha) perform better from Wald-type method for all heterogeneity values.

(c) Power

When heterogeneity increased, power reduced for all heterogeneity estimators (Figure 3). It is obvious that power increases with the number of studies and with the true intervention effect. RBp has the lowest power followed by the SJ. In addition, estimators from the Knapp and Hartung (DLknha, GHOknha, MLknha and REMLknha) present slightly lower power from Wald-type method. All other estimators have similar behavior and they present high power. Finally, in the absence of heterogeneity, all the estimators present similar performance for big number of studies and for smaller number of studies $(k = 10, 20)$ HS and ML perform better than the other estimators.

4.2 Continuous outcome data

(a) Average absolute bias

In all cases, results show that bias increased with increasing heterogeneity and decreased with increasing number of studies (Supporting Information Figure 1). RBp has large bias for heterogeneity value until 0.05, while it has the smallest bias for larger heterogeneity value for all scenarios. Moreover, SJ has large bias for all scenarios when there is no heterogeneity or it is equal to 0.01, while for larger heterogeneity value (> 0.01) it ranges with the majority heterogeneity variance estimators. We also found that REML and DL2 have small bias in all cases and perform better than DL and PM. In general, REML, as a recommended choice from Novianti *et al*. [26], has good performance in all scenarios for all heterogeneity values. Moreover, DLb and secondly DLp present small bias for all cases. HM has good performance either for 0.01 (the best performance compared to all other heterogeneity estimators) either for 0.05 or for both cases. $FB_{informative}$ with the suggested prior by Rhodes *et al*. [35] for continuous outcome has not the best performance while its bias ranges with the majority of heterogeneity. Finally, in cases with absence of heterogeneity, RBo and ML have the smallest bias.

(b) Empirical type error I

As the number of trials increases, the closer the values of the empirical type error I are to its nominal level (Supporting Information Figure 2). RBp and SJ have empirical type error I smaller from the nominal interval for heterogeneity until 0.05 and 0.01 value respectively, while for larger heterogeneity they range in the nominal interval. In addition, RBo has larger type error I up than the nominal interval for all random effects meta-analysis scenarios. Knapp and Hartung method performs better to Wald-type method in terms of type error I. All other estimators have similar behavior for the empirical type error I with values close to the interval (L, U).

(c) Power

When heterogeneity increased, power reduced for all heterogeneity estimators (Supporting Information Figure 3). Power increases when the number of studies or the true intervention effect increased for all heterogeneity estimators. For the absence or for small heterogeneity, we have similar performance (large power) for all heterogeneity estimators. In all cases, RBo and DLb have large power in comparison with the other estimators. On the opposite side, estimators from Knapp and Hartung method, $FB_{informative}$ and BM present low power in all cases.

5. General results

5.1 Dichotomous outcomes

The behavior of heterogeneity estimators is getting worse for all assessment criteria as the value of heterogeneity increases or the number of studies decreases (Figures 1-3). Table 2 shows a general visualization for the heterogeneity estimators' behavior when the type of outcome is dichotomous. The majority of heterogeneity estimators has similar performance for all heterogeneity values (Table 2). More analytically, Table 2 provides that RBp has the worst behavior compared to the others for all the assessment criteria. HM has the best behavior than other estimators for heterogeneity values close to 0.07 and SJ presents good behavior only for high heterogeneity. In general, $FB_{informative}$ and DLp have small bias and good behavior in terms of type error I and power.

General results from Figures 1-3 and Table 2, show that the behavior of the PM estimator is similar to the majority of estimators in all of the assessment criteria. Results provide that positive estimators BM, RBp, SJ and HM perform better in the presence of heterogeneity, as it is expected. SJgho presents moderate behavior in terms of bias and power, as it has medium values considering the other estimators. We found that $FB_{informative}$ performs better than FB_{vague} in all cases (except the case of $\theta = 0, k = 10$ when the type of outcome is dichotomous) while $FB_{informative}$ presents very smaller bias (Table 3). For example, as shows Table 3, bias of $FB_{informative}$ is 0.043 while bias for FB_{vague} is 0.094 in case with $\tau^2 = 0.07, \theta = 0.5, k = 10$ (dichotomous outcome). Moreover, $FB_{informative}$ is the most appropriate choice while it has the smallest bias in the most cases and it presents good performance in terms of type error I and power. Finally, DLb estimator has generally low bias in the presence of heterogeneity and provides very high power. Also, it is located within the nominal interval of empirical type error I except for some cases with high or zero heterogeneity.

5.2 Continuous outcomes

The behavior of heterogeneity estimators is getting worse for all assessment criteria as the value of heterogeneity increases or the number of studies decreases (Supporting Information Figures 1-3). Table 2 shows a general visualization for the overall heterogeneity estimators' behavior when the type of outcome is continuous. The majority of heterogeneity estimators have similar performance for all heterogeneity values (Table 2). In general, SJ, RBo, BM and RBp present bad performance. The $FB_{informative}$ estimator performs better than FB_{vague} (Table 3) but its bias ranges with the majority of heterogeneity estimators. General results from Supporting Information Figures 1-3 and Table 2, show that the behavior of REML estimator which was suggested by Novianti *et al*. [26] presents a good performance for all heterogeneity values but not the ideally. DLb and secondly DLp present the best behavior in all cases. Moreover, HM has the best performance for medium values of heterogeneity ($\tau^2 = 0.01, 0.05$). Finally, for the fixed effect meta-analysis scenarios, RBo and ML are the most preferable.

6. Conclusion

Conclusions from this simulation study pertain to the scenarios considered. There is no guarantee that estimators would have a similar behavior with small event rates and small sample sizes. Our simulation study shows that Knapp and Hartung method presents better behavior than Wald type method in terms of empirical type error I (or the similar measure of coverage probability) which comes in agreement with IntHout *et al*. [36]. In addition, the usage of Knapp and Hartung method needs caution while it performs worse than Wald-type method for zero heterogeneity in same cases [37]. Several simulations suggest using the well-known PM estimator because it is less biased compared to DL and REML estimators [15, 24, 26, 38]. Although PM seems to be the best estimator so far, it has not been compared with the recently recommended heterogeneity estimators suggested by Kontopantelis *et al*. [8] and Rukhin *et al*. [18]. Our simulation study (after a comparison of PM with more heterogeneity estimators) presents that PM has a good performance but not markedly better than the majority of heterogeneity estimators.

In general, we found that the majority of heterogeneity estimators have similar performance for all heterogeneity values for the two types of outcome (Table 2). The FB with the suggested prior from Turner *et al*. [34] ($FB_{informative}$) performs well for dichotomous outcome in the presence of heterogeneity, while it has the smallest bias compared to all other estimators. We found that the FB with the suggested prior from Turner *et al*. [34] provides the best performance for dichotomous outcome but FB with the suggested prior from Rhodes *et al*. [35] for continuous outcome is not markedly better than the majority of the heterogeneity estimators. We selected these two above recommended priors and the Uniform prior as an indicative application of FB estimator. The values of heterogeneity considered in this study are not unlikely under the empirical distributions used for dichotomous and continuous outcomes. If true heterogeneity is much larger and its value has a small likelihood under the suggested empirical distributions, the $FB_{informative}$ will underestimate it. However, empirical work of thousands of published meta-analyses shows that these distributions cover the plausible range of values for heterogeneity. A Bayesian framework is more flexible for more complex meta-analyses and it also allows the estimation of the

uncertainty in heterogeneity. Further simulation studies, using several priors, are needed to properly evaluate the FB estimator.

We found that the DLb estimator performs well in all assessment criteria for both dichotomous and continuous outcomes and it can be an alternative choice for dichotomous outcome. A limitation of the DLb estimator is that it is time consuming, while it is an iterative estimator and according to Kontopantelis *et al*. [8] DLb performs best despite its higher bias that he found for small number of studies ($k < 5$).

Moreover, DLp can be an alternative choice for the two types of outcome and HM for dichotomous outcome, when the heterogeneity values are close to 0.07, and for continuous outcome, when the heterogeneity values are close to $(0.01, 0.05)$ interval. In addition, this recommendation of DLp and HM can be based on the fact that these two estimators present better performance from the suggested by Novianti *et al*. [26] REML estimator. Finally, ML estimator provides the best performance for both types of outcome for the application of fixed effect meta-analysis' model.

Acknowledgements

The work of MP and DM is supported by the OPERAM project from the European Union's Horizon 2020 research and innovation programme under the grant agreement No 634238.

References

1. Higgins JPT, Green S (editors). *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from www.cochrane-handbook.org.
2. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. *Front Matter, in: Introduction to Meta-Analysis*. Wiley: New York, 2009.
3. Nikolakopoulou A, Mavridis D, Salanti G. Demystifying fixed and random effects meta-analysis. *Evidence-Based Mental Health* 2014; **17**(2), 53–57.
4. Veroniki AA, Jackson D, Viechtbauer W, Bender R, Bowden J, Knapp G, Kuss O, Higgins JPT, Langan D, Salanti G. Methods to estimate heterogeneity variance and its uncertainty in meta-analysis. *Research Synthesis Methods* 2015; **7**(1), 55–79.
5. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**(3): 177–188.
6. Hedges LV, Olkin I. *Statistical Methods for Meta-Analysis*. Orlando: Academic Press, 1985.
7. Paule RC, Mandel J. Consensus values and weighting factors. *Journal of Research of the National Bureau of Standards* 1982; **87**, 377 – 385.
8. Kontopantelis E, Springate DA, Reeves D. A re-analysis of the Cochrane Library data: the dangers of unobserved heterogeneity in meta-analyses. *PLoS One* 2013; **8**(7), e69930.
9. DerSimonian R, Kacker R. Random-effects model for meta-analysis of clinical trials: An update. *Contemporary Clinical Trials* 2007; **28**(2), 105–114.

10. Hartung J, Makambi KH. Reducing the number of unjustified significant results in meta-analysis. *Communications in Statistics - Simulation and Computation* 2003; **32**(4), 1179–1190.
11. Hartung J, Makambi KH. Positive estimation of the between-study variance in meta-analysis. *South African Statistical Journal* 2002; **36**(1), 55–76.
12. Hunter JE, Schmidt FL. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. Thousand Oaks, CA: SAGE Publications, 2004.
13. Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. *Statistics in Medicine* 1996; **15**(6), 619–629.
14. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in Medicine* 1999; **18**(20), 2693–2708.
15. Sidik K, Jonkman JN. A comparison of heterogeneity variance estimators in combining results of studies. *Statistics in Medicine* 2007; **26**(9), 1964–1981.
16. Morris CN. Parametric Empirical Bayes Inference: theory and applications. *Journal of the American Statistical Association* 1983; **78**(381), 47–55.
17. Sidik K, Jonkman JN. Simple heterogeneity variance estimation for meta-analysis. *Journal of the Royal Statistical Society: Series C: Applied Statistics* 2005; **54**(2), 367–384.
18. Rukhin AL. Estimating heterogeneity variance in meta-analysis. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 2013; **75**(3), 451–469.
19. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS -- a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* 2000; **10**, 325-337.
20. Chung Y, Rabe-Hesketh S, Choi I-H. Avoiding zero between-study variance estimates in random-effects meta-analysis. *Statistics in Medicine* 2014; **32**(23), 4071–4089.
21. Chung Y, Rabe-Hesketh S, Dorie V, Gelman A, Liu J. A Nondegenerate Penalized Likelihood Estimator for Variance Parameters in Multilevel Models. *Psychometrika* 2013; **78**(4), 685–709.
22. Malzahn U, Böhning D, Holling H. Nonparametric estimation of heterogeneity variance for the standardised difference used in meta-analysis. *Biometrika* 2000; **87**(3), 619–632.
23. Review Manager (RevMan) [Computer program]. Version 5.3. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration, 2014.
24. Bowden J, Tierney JF, Copas AJ, Burdett S. Quantifying, displaying and accounting for heterogeneity in the meta-analysis of RCTs using standard and generalised Q statistics. *BMC Medical Research Methodology* 2011; **11**, 41.
25. Knapp G, Hartung J. Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine* 2003; **22**(17), 2693–2710.
26. Novianti PW, Roes KCB, van der Tweel I. Estimation of between-trial variance in sequential meta-analyses: a simulation study. *Contemporary Clinical Trials* 2014; **37**(1), 129–138.
27. Viechtbauer W. Bias and Efficiency of Meta-Analytic Variance Estimators in the Random-Effects Model. *Journal of Educational and Behavioral Statistics* 2005; **30**(3), 261–293.
28. Hartung J. An Alternative Method for Meta-Analysis. *Biometrical Journal* 1999; **41**(8), 901–916.

29. Sidik K, Jonkman JN. A simple confidence interval for meta-analysis. *Statistics in Medicine* 2002; **21**(21), 3153–3159.
30. Röver C, Knapp G, Friede T. Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. *BMC Medical Research Methodology* 2015; **15**, 99.
31. R Development Core Team. R: A language and environment for statistical computing. Version 3.2.3, R Foundation for Statistical Computing, Vienna, Austria, 2015. http://www.R-project.org.
32. Viechtbauer W. Metafor: meta-analysis package for R, 2013.
33. Sturtz S, Ligges U, Gelman A. R2WinBUGS: A Package for Running WinBUGS from R. *Journal of Statistical Software* 2005; **12**(3), 1-16.
34. Turner RM, Davey J, Clarke MJ, Thompson SG, Higgins JPT. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane Database of Systematic Reviews. *International Journal of Epidemiology* 2012; **41**(3), 818–827.
35. Rhodes KM, Turner RM, Higgins JPT. Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continuous outcome data. *Journal of Clinical Epidemiology* 2015; **68**(1), 52–60.
36. IntHout J, Ioannidis JPA, Borm GF. 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* 2014; **14**(1), 25.
37. Sánchez-Meca J, Marín-Martínez F. Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychological Methods* 2008; **13**(1), 31–48.
38. Panityakul T, Bumrungsup C, Knapp G. On estimating residual heterogeneity in random-effects meta-regression: a comparative study. *Journal of Statistical Theory and Applications* 2013; **12**(3), 253.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.

Table 1. A presentation of all the investigated heterogeneity estimators.

Estimator	Abbreviation	Iterative / Non-iterative	Positive/ Non-negative
Generalized method of moments			
DerSimonian and Laird	DL	Non-iterative	Non-negative
General Hedges-Olkin	GHO	Non-iterative	Non-negative
Paule-Mandel (or empirical Bayes)	PM (or EB)	Iterative	Non-negative
Positive DerSimonian and Laird	DLp	Non-iterative	Positive
Two-step DerSimonian and Laird	DL2	Non-iterative	Non-negative
Two-step General Hedges-Olkin	GHO2	Non-iterative	Non-negative
Hartung-Makambi	HM	Non-iterative	Positive
Hunter-Schmidt	HS	Non-iterative	Non-negative
Maximum likelihood methods			
Maximum likelihood	ML	Iterative	Non-negative
Restricted maximum likelihood	REML	Iterative	Non-negative
Approximate restricted maximum likelihood	AREML	Iterative	Non-negative
Least squared methods			
Sidik- Jonkman	SJ	Non-iterative	Positive
Alternative Sidik-Jonkman (SJgho) (the same estimator with Sidik-Jonkman with a GHO estimator for a priori estimate)	SJgho	Non-iterative	Positive
Bayes methods			
Rukhin Bayes	RB	Non-iterative	Non-negative
Positive Rukhin Bayes	RBp	Non-iterative	Positive
Fully Bayesian	FB	Iterative	Non-negative
Bayes Modal	BM	Iterative	Positive
Non-Parametric methods			
Non-Parametric bootstrap DerSimonian and Laird	DLb	Non-iterative	Non-negative
Malzahn, B**ö**hning and Holling	MBH	Non-iterative	Non-negative

Table 2. Overall behavior of the bellow heterogeneity estimators in terms of the assessment criteria.

	Type of outcome					
	Dichotomous Outcome			Continuous Outcome		
Criteria / Estimators	Absolute bias	Empirical type error I	Power	Absolute bias	Empirical type error I	Power
DL	⊕	⊕	⊕	⊕	⊕	⊕
PM	⊕	⊕	⊕	⊕	⊕	⊕
DLp	⊕	⊕	⊕	⊕	⊕	⊕
GHO	×	⊕	⊕	⊕	⊕	⊕
DL2	⊕	⊕	⊕	⊕	⊕	⊕
GHO2	⊕	⊕	⊕	⊕	⊕	⊕
HM	⊕	⊕	⊕	⊕	⊕	⊕
HS	⊕	⊕	⊕	⊕	⊕	⊕
ML	⊕	⊕	⊕	⊕	⊕	⊕
REML	⊕	⊕	⊕	⊕	⊕	⊕
SJ	×	×	×	×	×	⊕
SJgho	×	⊕	⊕	⊕	⊕	⊕
DLb	⊕	⊕	⊕	⊕	⊕	⊕
RBo	⊕	×	⊕	×	×	⊕
RBp	×	×	×	×	×	⊕
$\boldsymbol{FB_{informative}}$	⊕	⊕	⊕	⊕	⊕	×
BM	×	⊕	⊕	~	⊕	×
DLknha	⊕	⊕	×	⊕	⊕	×
GHOknha	⊕	⊕	×	⊕	⊕	×
MLknha	⊕	⊕	×	⊕	⊕	×

REMLknha	⊕	⊕	×	⊕	⊕	×
MBH	-	-	-	⊕	⊕	⊕

Legend

⊕ good performance for all heterogeneity values	× bad performance for the heterogeneity values until the value of 0.07
⊕ good performance only for medium heterogeneity close to the value of 0.07	× bad performance for heterogeneity values until 0.01 (similar performance for larger τ^2)
⊕ good performance only for medium heterogeneity close to the value of 0.01	× bad performance for heterogeneity values until 0.07 (good performance for larger τ^2)
⊕ good performance while number of studies increased	× bad performance for all heterogeneity values
~ bad performance for heterogeneity ($\tau^2 > 0$)	- estimator exist only for continuous outcome

Table 3. Comparison the bias values of FB estimator between the uniform prior $\boldsymbol{Uniform(0,100)}$ and the informative priors suggested by Turner *et al* [34] and Rhodes *et al*. [35].

		Type of outcome							
		Dichotomous Outcome				Continuous Outcome			
Scenarios	τ^2 / FB	0	0.025	0.07	0.3	0	0.01	0.05	0.5
$\boldsymbol{\theta = 0, k = 10}$	FB_{vague}	0.013	0.025	0.035	0.042	0.013	0.015	0.044	0.336
	$FB_{informative}$	0.043	0.032	0.041	0.142	0.009	0.009	0.031	0.236
$\boldsymbol{\theta = 0, k = 30}$	FB_{vague}	0.039	0.043	0.044	0.055	0.003	0.006	0.016	0.127
	$FB_{informative}$	0.020	0.014	0.029	0.085	0.003	0.005	0.015	0.112
$\boldsymbol{\theta = 0.5, k = 10}$	FB_{vague}	0.066	0.076	0.094	0.251	0.013	0.015	0.047	0.353
	$FB_{informative}$	0.042	0.034	0.043	0.141	0.010	0.010	0.033	0.244
$\boldsymbol{\theta = 0.5, k = 30}$	FB_{vague}	0.018	0.019	0.039	0.092	0.004	0.006	0.017	0.126
	$FB_{informative}$	0.019	0.014	0.031	0.083	0.003	0.005	0.015	0.111

$\boldsymbol{FB_{vague}}$: FB with $\boldsymbol{Uniform(0,100)}$ prior; $\boldsymbol{FB_{informative}}$: FB with informative prior $\boldsymbol{logN(-2.56, 1.74^2)}$ on the untransformed $\boldsymbol{\tau^2}$ scale for dichotomous outcome [34] and $\boldsymbol{log(\tau^2) \sim t(-3.44, 2.59^2, 5)}$ for continuous outcome [35].