Original Article

Research Synthesis Methods

Received 26 June 2014, Revised 20 May 2015, Accepted 24 June 2015 Published online 2 September 2015 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/jrsm.1164

Methods to estimate the between-study variance and its uncertainty in meta-analysis[†]

Areti Angeliki Veroniki,[a*] Dan Jackson,[b] Wolfgang Viechtbauer,[c] Ralf Bender,[d] Jack Bowden,[e] Guido Knapp,[f] Oliver Kuss,[g] Julian PT Higgins,[h,i] Dean Langan[i] and Georgia Salanti[j]

Meta-analyses are typically used to estimate the overall/mean of an outcome of interest. However, inference about between-study variability, which is typically modelled using a between-study variance parameter, is usually an additional aim. The DerSimonian and Laird method, currently widely used by default to estimate the between-study variance, has been long challenged. Our aim is to identify known methods for estimation of the between-study variance and its corresponding uncertainty, and to summarise the simulation and empirical evidence that compares them. We identified 16 estimators for the between-study variance, seven methods to calculate confidence intervals, and several comparative studies. Simulation studies suggest that for both dichotomous and continuous data the estimator proposed by Paule and Mandel and for continuous data the restricted maximum likelihood estimator are better alternatives to estimate the between-study variance. Based on the scenarios and results presented in the published studies, we recommend the Q-profile method and the alternative approach based on a 'generalised Cochran between-study variance statistic' to compute corresponding confidence intervals around the resulting estimates. Our recommendations are based on a qualitative evaluation of the existing literature and expert consensus. Evidence-based recommendations require an extensive simulation study where all methods would be compared under the same scenarios. © 2015 The Authors. *Research Synthesis Methods* published by John Wiley & Sons Ltd.

Keywords: heterogeneity; mean squared error; bias; coverage probability; confidence interval

1. Introduction

Meta-analysis combines estimates of quantities of interest, as obtained from studies addressing the same research question. A degree of variability in study estimates is inevitably present because of within-study sampling error. Additional variability might occur for many reasons such as differences in the way studies are conducted and

[a] *Li Ka Shing Knowledge Institute, St. Michael's Hospital, 209 Victoria Street, East Building, Toronto, Ontario, M5B 1T8, Canada*
[b] *MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 0SR, UK*
[c] *Department of Psychiatry and Psychology, School for Mental Health and Neuroscience, Maastricht University, The Netherlands*
[d] *Department of Medical Biometry, Institute for Quality and Efficiency in Health Care (IQWiG), Im Mediapark 8, 50670 Cologne, Germany*
[e] *MRC Biostatistics Unit Hub for Trials Methodology Research, Cambridge, UK*
[f] *Department of Statistics, TU Dortmund University, 44221 Dortmund, Germany*
[g] *Institute for Biometrics and Epidemiology, German Diabetes Center, Leibniz Institute for Diabetes Research at Heinrich Heine University, 40225 Düsseldorf, Germany*
[h] *School of Social and Community Medicine, University of Bristol, Bristol, UK*
[i] *Centre for Reviews and Dissemination, University of York, York, UK*
[j] *Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece*
**Correspondence to: Areti Angeliki Veroniki, Li Ka Shing Knowledge Institute, St. Michael's Hospital, 209 Victoria Street, East Building, Toronto, Ontario, M5B 1T8, Canada.*
E-mail: veronikia@smh.ca

[†] *The legal statement for this article was changed on 27 June 2016, after original online publication.*

how the treatment effects are measured; this additional variability is usually modelled using a between-study variance parameter. Between-study variance refers to variation across study findings beyond random sampling error, and its quantification is often of interest and aids in the interpretation of results of a meta-analysis. Several methods have been suggested to quantify the amount of between-study variance in meta-analytic data. In the most popular family of methods, the between-study variance is represented by the variance of the distribution of the true study effects, commonly denoted as τ^2 in the meta-analytic literature.

Two approaches are most commonly applied for combining the study findings in a meta-analysis: (1) the fixed-effect (FE) model, and (2) the random-effects (RE) model (for a detailed description of the models, see section 2) (Borenstein *et al*., 2009). These two models are associated with different assumptions and the selection between the two might importantly impact on meta-analytic conclusions. In the FE model, the observed treatment effects are distributed around one common true treatment effect with distribution variance informed entirely by the within-study variances. In the RE model, the observed treatment effects estimate different study-specific true treatment effects, which are related and assumed to come from the same underlying distribution. The variability in the distribution is therefore attributed to both within-study variance, because of sampling error, and between-study variance. A different approach is the fixed-effects model, in which the true treatment effect parameters are unknown and estimated by the observed effects as in the RE model, but are treated as fixed and unrelated constants (Gardiner *et al*., 2009; Laird and Mosteller, 1990). Although the names of the 'fixed-effect' and 'fixed-effects' models are so similar, the assumptions under which the models are constructed are completely different. The description of the fixed-effects model is not presented in section 2 as it is beyond the scope of this review.

Several estimators for the between-study variance component (τ^2) have been proposed that vary in popularity and complexity. The DerSimonian and Laird (1986) (DL) method is the most commonly implemented approach and is the default approach in many software routines. However, its default use has often been challenged in the sense that DL may underestimate the true between-study variance, potentially producing overly narrow confidence intervals (CIs) for the mean effect (Cornell *et al*., 2014), especially when the between-study variance is large (Novianti *et al*., 2014). The use of alternative estimation methods is in practice limited by their availability in software. In Table 1, we summarise the between-study variance estimators currently available in various software packages.

Simulation studies (Chung *et al*., 2014; Kontopantelis *et al*., 2013; Sidik and Jonkman, 2007) have shown that estimates of the between-study variance are particularly inaccurate when the number of studies included in a meta-analysis is small. Inferences about the amount of between-study variance can therefore be misleading when only a point estimate is considered. Confidence intervals for τ^2 can facilitate interpretation (Ioannidis *et al*., 2007). Again, several options exist to quantify the uncertainty in the estimated amount of the between-study variance.

While a measure of the between-study variance is arguably a core output of a meta-analysis, investigators often consider it as an intermediate step when fitting a RE model. In particular, the inverse-variance meta-analysis method estimates a summary treatment effect as the weighted average of individual study findings, with weights depending on both within-study (sampling) variance and the estimated between-study variance. Consequently, the estimated amount of the between-study variance influences the weights assigned to each study and hence the overall summary treatment effect and, importantly, its precision. Although it is typical to assume that the weights are known constants when constructing confidence intervals and hypothesis tests, they are in fact unknown random variables that need to be estimated. This issue has previously been discussed for the usual tests that, as a consequence, do not have the properties assumed for them, such as the chi-square distribution for the Q-statistic under the null hypothesis of homogeneity (Kulinskaya *et al*., 2011a, 2011b), and some attempts have been made to account for the uncertainty in the weights (Böhning *et al*., 2002; Malzahn *et al*., 2000). Hence, among other factors, the performance of the between-study variance estimators depends on how well we estimate the study weights.

Special attention should also be paid to the case of rare events in dichotomous outcome data using the inverse-variance approach, where the degree of bias in the between-study estimation is proportional to the rarity of the study events. In such cases, it has been suggested to avoid the inverse-variance method altogether and instead either use the FE model and the Mantel–Haenszel method for unbalanced group sizes or Peto's method for balanced group sizes (Bradburn *et al*., 2007; Sweeting *et al*., 2004), or to switch to models/methods based on exact distributional assumptions (Kuss, 2014; Stijnen *et al*., 2010).

In this paper, we provide a comprehensive overview of methods used for estimating the between-study variance and its uncertainty. We searched in PubMed to identify research articles that describe or compare these methods in simulation or empirical studies, and we scanned the references of the selected articles for additional relevant literature (see Supporting Information). Eligible were methods that can be applied for any type of outcome data. Several published papers present and compare various between-study variance estimators (Chung *et al*., 2014; DerSimonian and Kacker, 2007; Novianti *et al*., 2014; Pullenayegum, 2011; Sidik and Jonkman, 2007; Viechtbauer, 2005) and their CIs (Jackson, 2013; Knapp *et al*., 2006; Viechtbauer, 2007), while many published empirical and simulation studies suggest different methods (Chung *et al*., 2014; Kontopantelis *et al*., 2013; Novianti *et al*., 2014; Sidik and Jonkman, 2007; Thorlund *et al*., 2011; Viechtbauer, 2007). Therefore, there is a need to summarise the alternative estimation options and the studies' conclusions, and to indicate whether some methods are preferable to others with regard to the studies' results.

Table 1. Software option (with packages or macros) for each τ^2 estimation method. To our knowledge, routines for Hartung and Makambi, two-step DerSimonian and Laird, positive DerSimonian and Laird, two-step Hedges and Olkin, Rukhin Bayes, positive Rukhin Bayes, and non-parametric bootstrap methods are not available in any of the software options listed below. The relevant references for the underlying packages and macros are presented at the end of the table.

		Estimation methods (packages/macros)									
Software	License type	DerSimonian and Laird (DL)	Paule and Mandel (PM)	Hedges and Olkin (HO)	Hunter and Schmidt (HS)	Maximum likelihood (ML)	Restricted maximum likelihood (REML)	Approximate restricted maximum likelihood (AREML)	Sidik and Jonkman (SJ)	Full Bayes (FB)	Bayes modal (BM)
Comprehensive Meta-Analysis (Borenstein *et al.*, 2005) www.meta-analysis.com/	Commercial	Yes	—	—	—	Yes	—	—	—	—	—
Excel using the MetaEasy AddIn (Kontopantelis and Reeves, 2009) http://www.jstatsoft.org/v30/i07	Freeware	Yes	—	—	—	Yes	—	—	—	—	—
HLM (Raudenbush *et al.*, 2004) http://www.ssicentral.com/hlm/	Commercial	—	—	—	—	Yes	Yes	—	—	—	—
Meta-DiSc (Zamora *et al.*, 2006) ftp://ftp.hrc.es/pub/programas/metadisc/	Freeware	Yes	—	—	—	Yes	Yes	—	—	—	—
Metawin (Rosenberg *et al.*, 2000) http://www.metawinsoft.com/	Commercial	Yes	—	—	—	Yes	—	—	—	—	—
MIX (Bax, 2011) www.mix-for-meta-analysis.info/	Commercial	Yes	—	—	—	—	—	—	—	—	—
MLwin (Rasbash *et al.*, 2014) http://www.bristol.ac.uk/cmm/software/mlwin/	Freeware	—	—	—	—	Yes	Yes	—	—	Yes	—
Open Meta Analyst (Wallace *et al.*, 2012) http://www.cebm.brown.edu/open_meta	Freeware	Yes	Yes	Yes	—	Yes	Yes	—	Yes	—	—
RevMan (The Nordic Cochrane Centre, 2014) www.cochrane.org/	Freeware	Yes	—	—	—	—	—	—	—	—	—
R (R Development Core Team, 2008) http://www.r-project.org/	Freeware	Yes (meta, metafor, netmeta, mvmeta)	Yes (meta, metafor)	Yes (meta, metafor, mvmeta)	Yes (meta, metafor)	Yes (meta, metaSEM, metafor, mvmeta)	Yes (meta, metaSEM, metafor, mvmeta)	—	Yes (meta, metafor)	Yes (R2WinBUGS, BRugs, rjugs)	Yes (blme)

(*Continues*)

Table 1. (Continued)

		Estimation methods (packages/macros)									
Software	License type	DerSimonian and Laird (DL)	Paule and Mandel (PM)	Hedges and Olkin (HO)	Hunter and Schmidt (HS)	Maximum likelihood (ML)	Restricted maximum likelihood (REML)	Approximate restricted maximum likelihood (AREML)	Sidik and Jonkman (SJ)	Full Bayes (FB)	Bayes modal (BM)
SAS (SAS Institute Inc., 2003) http://www.sas.com/technologies/analytics/statistics/stat/	Commercial	Yes (marandom.sas)	—	—	—	Yes (marandom.sas, PROC IML, PROC MIXED, PROC GLIMMIX)	Yes (PROC IML, PROC MIXED, PROC GLIMMIX)	—	—	Yes (SASBUGS, RASmacro, PROC MCMC)	—
Stata (StataCorp, 2013) www.stata.com/	Commercial	Yes (metareg, metan, metaan, mvmeta)	Yes (metareg)	—	—	Yes (metareg, metaan, mvmeta)	Yes (metareg, metaan, mvmeta)	—	—	—	Yes (gllamm)
SPSS (IBM Corp., 2013) http://www.spss.co.in/	Commercial	Yes (meanes.sps, metaf.sps, metareg.sps)	—	—	—	Yes (metaf.sps, metareg.sps)	—	Yes (metaf.sps, metareg.sps)	—	—	—
BUGS (Thomas, 1994), OpenBUGS (Thomas, 2010), or WinBUGS (Lunn *et al*., 2000) www.mrc-bsu.cam.ac.uk/bugs/	Freeware	—	—	—	—	—	—	—	—	Yes	—

R: meta (http://cran.r-project.org/web/packages/meta/meta.pdf), metafor (Viechtbauer, 2013) (http://www.metafor-project.org/doku.php), netmeta (http://cran.r-project.org/web/packages/netmeta/netmeta.pdf), mvmeta (http://cran.r-project.org/web/packages/mvmeta/mvmeta.pdf), metaSEM (http://courses.nus.edu.sg/course/psycwlm/Internet/metaSEM/), R2WinBUGS (http://cran.r-project.org/web/packages/R2WinBUGS/R2WinBUGS.pdf), BRugs (http://cran.r-project.org/web/packages/BRugs/BRugs.pdf), rjugs (http://cran.r-project.org/web/packages/rjags/rjags.pdf), blme (http://cran.r-project.org/web/packages/blme/blme.pdf)
SAS: marandom.sas (http://www.senns.demon.co.uk/SAS%20Macros/SASMacros.html), PROC IML (http://support.sas.com/documentation/cdl/en/imlug/63541/PDF/default/imlug.pdf), PROC MIXED (https://support.sas.com/documentation/cdl/en/statugmixed/61807/PDF/default/statugmixed.pdf), PROC GLIMMIX (https://support.sas.com/documentation/cdl/en/statugglmmix/61788/PDF/default/statugglmmix.pdf), SASBUGS (Zhang *et al*., 2008), RASmacro (https://github.com/rsparapa/rasmacro), PROC MCMC (http://support.sas.com/documentation/cdl/en/statugmcmc/63125/PDF/default/statugmcmc.pdf)
Stata: metareg (Harbord and Higgins, 2008), metan (Harris *et al*., 2008), metaan (Kontopantelis and Reeves, 2010), mvmeta (White, 2009), gllamm (Rabe-Hesketh *et al*., 2003) (http://www.gllamm.org/programs.html)
SPSS: meanes.sps (http://mason.gmu.edu/~dwilsonb/ma.html), metaf.sps (http://mason.gmu.edu/~dwilsonb/ma.html), metareg.sps (http://mason.gmu.edu/~dwilsonb/ma.html)

After a brief description of the usual meta-analysis models in section 2, we describe 16 methods to estimate between-study variance in section 3. We describe seven methods to quantify uncertainty in the between-study variance in section 4, and then illustrate the application of the various methods in example meta-analyses along with our recommendations in section 5. We conclude with a summary of the comparative studies that evaluate the performance of the various methods in section 6. Our recommendations are based on a qualitative evaluation of the existing literature and expert consensus. Evidence-based recommendations require an extensive simulation study where all methods would be compared under the same scenarios.

2. Models for meta-analysis

Consider the situation where the quantity of interest is the effect of an intervention compared to a control condition or the presence of a risk factor compared to its absence, measured in terms of an effect size (such as the log-odds ratio or the standardised mean difference). The main two parametric models used to combine study results are the FE model and the RE model. The FE model assumes that all studies share the same (fixed) effect, that is, there is one 'true effect' size and all differences in the observed effects are because of sampling error. In the RE model, the effects in the studies are assumed to represent a random sample from a distribution of true treatment effects, most commonly a normal distribution. The width of the distribution describes the degree of the between-study variance. The RE model encompasses within-study (v_i, index i refers to the i^{th} study, with $i = 1, \ldots, k$) and between-study (τ^2) variation, in contrast to the FE model which includes within-study variation only. Uncertainty in the location of the mean effect in a RE meta-analysis depends on the magnitude of the between-study variance, the number of studies, and the precision of the individual study estimates (Hardy and Thompson, 1996). Several approaches have been suggested to estimate a CI around the mean effect, the performance of which has been examined in simulation studies under various scenarios (Brockwell and Gordon, 2001; Hartung, 1999; Kontopantelis and Reeves, 2012).

In the presence of between-study variance ($\tau^2 > 0$), the RE model results in a wider CI compared with the FE model, reflecting greater uncertainty around the mean (Villar *et al*., 2001). In Table 2, we provide notation and define both models formally.

Given observed treatment effect y_i in study $i = 1, \ldots, k$ (e.g. the log-odds ratio), the common treatment effect μ_{FE} under the FE model is estimated as

$$\hat{\mu}_{FE} = \frac{\sum w_{i,FE}\, y_i}{\sum w_{i,FE}} \tag{1}$$

where $w_{i,FE}$ is the weight (the inverse of the variance) assigned to each study. Note that throughout the paper all summations go from $i = 1$ to $i = k$. Under the assumptions of the model, the summary treatment effect in (1) is the uniformly minimum variance unbiased estimator of μ_{FE} (Viechtbauer, 2005).

Under the RE model, we instead estimate μ_{RE}, the mean of the distribution of true treatment effects. Let δ_i denote the difference between this mean and the underlying study-specific true effect θ_i in a particular study. The estimated summary treatment effect $\hat{\mu}_{RE}$ is computed as in Equation (1) using weights appropriate to the RE model, $w_{i,RE}$, as provided in Table 2. In the following sections, we will use the notation $\hat{\mu}_{RE}(\hat{\tau}^2)$ to emphasise that the overall treatment effect depends on the estimated amount of between-study variance. Model parameters are usually estimated under the assumption that the study variances v_i are known when in fact they are estimated from the observed study data. We make this assumption throughout the paper, so the distributions of statistics presented are good approximations only when the study sizes are large.

3. Methods to estimate the between-study variance

We consider 16 between-study variance estimators, and these are summarised in Table 3, along with abbreviations which we will use in the text of the sections that follow. The methods may be divided into two main

Table 2. Models to synthesise study results in a meta-analysis.

Fixed-effect model	Random-effects model
	$y_i = \theta_i + \varepsilon_i$
$y_i = \mu_{FE} + \varepsilon_i$	$\theta_i = \mu_{RE} + \delta_i$
$\varepsilon_i \sim N(0, v_i)$	$\varepsilon_i \sim N(0, v_i)$
$Var(y_i) = v_i$	$\delta_i \sim N(0, \tau^2)$
$w_{i,FE} = 1/v_i$	$Var(y_i) = v_i + \tau^2$
	$w_{i,RE} = 1/(v_i + \tau^2)$

Table 3. Overview of the estimators for the between-study variance.

Estimator	Abbreviation	Iterative/Non-iterative	Positive/Non-negative
Method of moments estimators			
DerSimonian and Laird	DL	Non-iterative	Non-negative
Positive DerSimonian and Laird	DLp	Non-iterative	Positive
Two-step DerSimonian and Laird	DL2	Non-iterative	Non-negative
Hedges and Olkin	HO	Non-iterative	Non-negative
Two-step Hedges and Olkin	HO2	Non-iterative	Non-negative
Paule and Mandel	PM	Iterative	Non-negative
Hartung and Makambi	HM	Non-iterative	Positive
Hunter and Schmidt	HS	Non-iterative	Non-negative
Maximum likelihood estimators			
Maximum likelihood	ML	Iterative	Non-negative
Restricted maximum likelihood	REML	Iterative	Non-negative
Approximate restricted maximum likelihood	AREML	Iterative	Non-negative
Model error variance estimator			
Sidik and Jonkman	SJ	Non-iterative	Positive
Bayes estimators			
Rukhin Bayes	RB	Iterative	Non-negative
Positive Rukhin Bayes	RBp	Iterative	Positive
Full Bayes	FB	Iterative	Non-negative
Bayes Modal	BM	Iterative	Positive
Bootstrap estimator			
Non-parametric bootstrap DerSimonian and Laird	DLb	Iterative	Non-negative

groups: closed-form (or non-iterative) methods and iterative methods. Closed-form methods provide a parameter estimator in a predetermined number of steps, whereas iterative methods converge to a solution when a specific criterion is met (however, some iterative methods do not always produce a result because of failure to converge). Methods may also be distinguished by whether they yield only positive values or whether they yield non-negative estimates (i.e. $\hat{\tau}^2$ can be zero).

We describe known properties of the methods in terms of bias, mean squared error (MSE), and efficiency. Bias is the difference between the expected value of the estimator and its true value, and is given by

$$Bias\left(\hat{\tau}^2\right) = E\left(\hat{\tau}^2\right) - \tau^2 = E\left(\hat{\tau}^2 - \tau^2\right).$$

Negatively or positively biased estimators lead to an under- or over-estimation of the true between-study variance. Therefore, it is desirable that an estimator for the amount of the between-study variance satisfies $E(\hat{\tau}^2) = \tau^2$, that is, it is unbiased.

A good estimator should not only be unbiased, but also remain unaffected as much as possible by sampling fluctuation (efficiency), that is, the extent to which the estimator takes on different values with different samples. MSE is the squared distance between the estimator and its true value:

$$MSE\left(\hat{\tau}^2\right) = E\left[\left(\hat{\tau}^2 - \tau^2\right)^2\right] = Var\left(\hat{\tau}^2\right) + \left(Bias\left(\hat{\tau}^2\right)\right)^2.$$

If $MSE\left(\hat{\tau}_1^2\right) < MSE\left(\hat{\tau}_2^2\right)$, then $\hat{\tau}_1^2$ is said to be more efficient than $\hat{\tau}_2^2$. Finally, CIs produced by different estimation methods for either τ^2 or μ are often compared in terms of coverage probability (i.e. the proportion of times the interval includes the true value of the parameter being estimated) and width. Methods that provide narrower CIs with coverage probability close to the nominal level are preferable.

3.1. DerSimonian and Laird (DL) method

The DL estimator is possibly the most frequently used approach as it is a non-iterative method that is simple to implement (DerSimonian and Laird, 1986). In fact, many software routines have DL as the default method to estimate the between-study variance. The estimator is derived by equating the expected value of Cochran's *Q*-statistic with its observed value, yielding

$$E(Q) = \tau^2\left(\sum w_{i,FE} - \frac{\sum w_{i,FE}^2}{\sum w_{i,FE}}\right) + (k-1),$$

where Q is calculated based on an estimate from a FE analysis, $\hat{\mu}_{FE}$ with:

$$Q = \sum w_{i,FE}(y_i - \hat{\mu}_{FE})^2 = \sum \frac{(y_i - \hat{\mu}_{FE})^2}{v_i}.$$

The DL estimator can therefore be obtained as

$$\hat{\tau}^2_{DL} = \max\left\{0, \frac{Q-(k-1)}{\sum w_{i,FE} - \frac{\sum w^2_{i,FE}}{\sum w_{i,FE}}}\right\}.$$

The Cochran's Q-statistic belongs to the 'generalised Cochran between-study variance statistics' (DerSimonian and Kacker, 2007):

$$Q_a = \sum a_i (y_i - \hat{\mu}_a)^2,$$

with a_i representing weights assigned to each study that are equal to any positive value, and $\hat{\mu}_a = \frac{\sum a_i y_i}{\sum a_i}$. Similarly to DL, equating Q_a to its expected value

$$E(Q_a) = \tau^2\left(\sum a_i - \frac{\sum a_i^2}{\sum a_i}\right) + \left(\sum a_i v_i - \frac{\sum a_i^2 v_i}{\sum a_i}\right),$$

and solving for τ^2 we can obtain the generalised method of moments (GMM) estimator:

$$\hat{\tau}^2_{GMM} = \max\left\{0, \frac{Q_a - \left(\sum a_i v_i - \frac{\sum a_i^2 v_i}{\sum a_i}\right)}{\sum a_i - \frac{\sum a_i^2}{\sum a_i}}\right\}.$$

Therefore, the DL estimator is a special case of the general class of method of moments estimators with weights $a_i = w_{i,FE} = 1/v_i$.

Under the assumptions of the RE model assuming known within-study variances v_i and before the truncation of negative values, the generalised method of moments estimator is unbiased. However, the need to truncate negative values to zero introduces positive bias into the estimator (Rukhin, 2013; Viechtbauer, 2005) and consequently the DL estimator is positively biased, over-estimating the true amount of between-study variance on average. When k decreases and/or the v_i increase, the estimator becomes more variable and truncation is often needed, leading to positive bias (Viechtbauer, 2005).

However, before truncation, the DL estimator is unbiased if the sampling variances are known rather than estimated. Any bias in the estimator is therefore not inherent to the estimator per se, but depends on how well sampling variances are estimated. Potential bias in the DL estimator might also stem from other factors apart from estimation of the v_i values, such as bias in the treatment effect estimates and/or correlation between the treatment effect estimates and their corresponding sampling variances. It should be noted that DerSimonian and Laird (1986) suggested the truncated estimator as shown above, and from now on (unless stated otherwise) we will refer to the truncated version of their estimator as DL.

Simulation studies have suggested that the DL between-study estimate is acceptable when true levels of the between-study variance are small or close to zero and k is large, whereas when τ^2 is large, the DL estimator can produce estimates with significant negative bias (Bowden *et al*., 2011; Novianti *et al*., 2014; Sidik and Jonkman, 2007, 2005a, 2005b). The negative bias that has been reported with respect to the DL estimator seems to be something especially related to using effect size measures based on 2×2 table data (e.g. odds ratios, risk ratios), where problems arise when using very large τ^2 values in simulation studies. In particular, very large τ^2 can lead to extreme values of the effect size measure, at which point many tables will include zero cells and the accuracy and applicability of the inverse-variance method becomes questionable.

The DL estimator is associated with lower MSE than the HO, SJ, and PM estimators (which we describe below) when the true between-study variance is not too large (Sidik and Jonkman, 2007). Jackson *et al*. (2010) evaluated the efficiency of the DL estimator asymptotically, that is, they assessed whether the variance of this estimator attains the Cramér–Rao bound for $k \to \infty$.They showed that DL is inefficient when the studies included in the meta-analysis are of different sizes and particularly when τ^2 is large. However, they suggested that the DL estimator can be efficient for inference on μ when the number of studies included in the meta-analysis is large.

3.2. Positive DerSimonian and Laird (DLp) method

Kontopantelis *et al*. (2013) carried out an empirical and a simulation study for dichotomous outcome data and concluded that $\hat{\tau}^2_{DL}$ is often estimated to be zero when its true value is positive, especially for small k. They claimed that positive estimation methods are preferable to non-negative methods and proposed an alternative approach to the DL method (which we denote DLp) that ensures a positive value:

$$\hat{\tau}^2_{DLp} = \begin{cases} \hat{\tau}^2_{DL} \ , & \hat{\tau}^2_{DL} > 0 \\ c \ , & \hat{\tau}^2_{DL} \leq 0 \end{cases}$$

with c denoting an arbitrary positive constant. In a simulation study reflecting the meta-analysis of log-odds ratios, the authors selected $c = 0.01$ and showed that DLp has lower bias than other positive estimators (SJ, HO, RBp), irrespective of the distribution of the study-specific true effects θ_i.

3.3. Two-step estimator with DerSimonian and Laird (DL2) method

DerSimonian and Kacker (2007) proposed a non-iterative, two-step estimator (DL2). It is based on the generalised method of moments estimator and the generalised Cochran between-study variance statistic, with $a_i = w_{i,RE} = 1/\left(v_i + \hat{\tau}^2_{DL}\right)$, and $\hat{\mu}_{RE}\left(\hat{\tau}^2_{DL}\right)$ computed as in (1) using these RE weights. The estimator can be obtained by

$$\hat{\tau}^2_{DL2} = \max\left\{0, \frac{Q_{w_{RE}} - \left(\sum w_{i,RE} v_i - \frac{\sum w^2_{i,RE} v_i}{\sum w_{i,RE}}\right)}{\sum w_{i,RE} - \frac{\sum w^2_{i,RE}}{\sum w_{i,RE}}}\right\}. \qquad (2)$$

When all of the sampling variances are equal to each other, the estimator reduces to the HO method. DerSimonian and Kacker (2007) compared the DL, HO, DL2, HO2, and PM methods (which we describe below) and showed that the DL2 estimator approximates the PM estimator, which may have some desirable statistical properties. However, Bhaumik *et al*. (2012) found that for rare events the DL2 estimator is downwardly biased.

3.4. Non-parametric bootstrap DerSimonian and Laird method (DLb)

Kontopantelis *et al*. (2013) suggested a non-parametric bootstrap version of the DL method (DLb) by randomly sampling B sets of studies with replacement. In each set, they estimate τ^2 using the DL method and then estimate $\hat{\tau}^2_{DLb}$ as the mean of these B estimates. The authors carried out a simulation study and suggested that DLb was associated with lower bias than SJ or RBp (which we describe below) for $k \geq 5$. The same study suggested that DLb performed better compared to DL in terms of identifying the presence of between-study variance, especially when the number of studies was small. However, non-parametric bootstrap methods perform well only when a large number of studies is included in the meta-analysis and the observed benefit may be artificial. The study also showed that DLb revealed greater bias compared with DL, which was more profound in small meta-analyses. Although described for the DL method, this method can in fact be employed for every between-study variance estimator.

3.5. Hedges and Olkin (HO) method

The HO (Hedges and Olkin, 1985) estimator (also known as Cochran estimator or variance component type estimator) was first introduced in a RE analysis of variance context by Cochran (1954). Hedges (1983) discussed the estimation method for the between-study variance component in the meta-analytic context. The estimator is obtained by setting the sample variance

$$S^2_y = \frac{1}{k-1}\sum (y_i - \bar{y})^2$$

equal to its expected value and solving for τ^2, which yields

$$\hat{\tau}^2_{HO} = \max\left\{0, \frac{1}{k-1}\sum (y_i - \bar{y})^2 - \frac{1}{k}\sum v_i\right\},$$

where $\bar{y}$ is the unweighted average of y_i. DerSimonian and Laird (1986) noted that the difference between the DL and HO method of moments estimators is that HO is based on the unweighted variance of the treatment effect estimates, whereas DL is based on their weighted variance. The method is a special case of the generalised method of moments estimator with $a_i = 1/k$ as DerSimonian and Kacker (2007) suggested, or with a_i equal to any other positive constant independent of i. Although the HO estimator is simple to compute and does not require an iterative numerical solution, it is not widely used. However, it is worth noting that the HO estimator is exactly unbiased (before being truncated) when the sampling variances can be estimated unbiasedly (Viechtbauer, 2005). Unbiased estimates of the sampling variances can in fact be obtained for some outcome measures used in meta-analyses (e.g. risk differences, raw mean differences, and standardised mean differences).

DerSimonian and Laird (1986) compared the HO method with the DL, ML, and REML methods. They concluded that, on average, the DL and REML methods yield slightly larger values than ML, but all three gave lower values than the HO estimator. This was corroborated by an empirical study (Thorlund *et al*., 2011) showing that the HO estimator produces on average larger estimates than the DL method. The HO method performs well in the presence of substantial between-study variance, especially when the number of studies is large (i.e. $k \geq 30$), but produces large MSE (Chung *et al*., 2014; Panityakul *et al*., 2013; Sidik and Jonkman, 2007; Viechtbauer, 2005). Sidik and Jonkman (2007) showed that for large τ^2 and moderate to large k, the HO estimator has the largest MSE in comparison with the DL, ML, REML, PM, and SJ estimators, but for large k, it has smaller bias than the DL, ML, and REML estimators. Friedman (2000) derived the variances of the DL and HO estimators and found that the DL estimator is more efficient than the HO estimator when τ^2 is zero, while the opposite occurs when the amount of between-study variance is large. However, Sidik and Jonkman (2007) in their simulations concluded the opposite, namely that the DL estimator is more efficient than the HO estimator when τ^2 is large. They attribute the different conclusions to the fact that Friedman derives variances for the estimators that do not take into account the truncation of negative values.

3.6. Two-step estimator for the Hedges and Olkin (HO2) method

The method of DerSimonian and Kacker (2007) is a two-step estimator and belongs to the family of the generalised method of moments estimators. In the first step, we start with the HO estimator and use the weights $a_i = w_{i,RE} = 1/\left(v_i + \hat{\tau}^2_{HO}\right)$ and the overall treatment effect $\hat{\mu}_{RE}\left(\hat{\tau}^2_{HO}\right)$ as in (1). In the second step, we obtain the HO2 estimator as in (2). When all sampling variances are equal, the estimator reduces to the HO method. DerSimonian and Kacker (2007) showed that the two-step estimators HO2 and DL2 approximate the PM estimator better than the one-step HO and DL estimators, and suggest the use of either DL2 or HO2 when a non-iterative method is desired.

3.7. Paule and Mandel (PM) method

Paule and Mandel (1982) proposed to profile a special form of Q_a, with $a_i = w_{i,RE} = 1/(v_i + \tau^2)$, the generalised Q-statistic:

$$Q_{gen} = \sum w_{i,RE}\left(y_i - \hat{\mu}_{RE}\left(\tau^2\right)\right)^2 \sim \chi^2_{k-1},$$

until Q_{gen} equals its expected value (i.e. $E(Q_{gen}) = k - 1$). Q_{gen} is a pivotal quantity to test the null hypothesis that the true between-study variance is equal to a certain amount $\tau^2_0 (\geq 0)$, and depends on the unknown τ^2. PM is an iterative estimator that belongs to the family of the GMM estimators, where the RE weights and overall effect are simultaneously calculated using the true value of τ^2 that is part of the pivotal quantity. Similarly to the distribution of Cochran's Q-statistic, the chi-square distribution of Q_{gen} depends on how well the study-specific weights, variances, and treatment effects are estimated.

The method is actually equivalent to the empirical Bayes estimator that was discussed by Morris (1983) and introduced into the meta-analytic context by Berkey *et al*. (1995). Provided that the sampling variances and τ^2 are fixed and known and that the first two moments exist, the expectation of Q_{gen} is equal to $k - 1$ even when the underlying distributions are not normal (Rukhin, 2013). Because Q_{gen} is a monotonically decreasing function of τ^2, $\hat{\tau}^2_{MP}$ is set equal to zero when $Q_{gen} < k - 1$ for $\tau^2 = 0$ (DerSimonian and Kacker, 2007).

Rukhin *et al*. (2000) showed that when assumptions underlying the method do not hold, the method is still more robust compared to the DL estimator, which depends on large sample sizes. Panityakul *et al*. (2013) showed that the PM estimator is approximately unbiased for large sample sizes and also provided R code for computing the PM estimator. It has been shown that the PM method has upward bias for small k and τ^2, whereas for large k and τ^2 it is downwardly biased (Sidik and Jonkman, 2007); but generally the method is less biased than its alternatives. One simulation study suggested that PM outperforms the DL and REML estimators in terms of bias (Panityakul *et al*., 2013). Novianti *et al*. (2014) compared the DL, DL2, PM, HO, REML, and SJ estimators, and showed that the PM method performed best in terms of bias for both dichotomous and continuous outcome data. Sidik and Jonkman (2007) highlighted the methodological similarity between the SJ and PM estimators, stating that differences are because of the fact that SJ is simplified to two steps and avoids zero between-study variance estimates, and they showed that the two estimators have similar MSE. In fact, if the SJ estimator would be iterated, then it would yield the same exact value as the PM method. Although the PM estimator seems to perform well in terms of bias, Knapp and Hartung (2003) found that it is less efficient than the DL and REML estimators.

Bowden *et al*. (2011) carried out an empirical study comparing the DL and PM estimators and showed that as the between-study variance increases, $\hat{\tau}^2_{PM}$ becomes greater than $\hat{\tau}^2_{DL}$. The authors recommend the use of $\hat{\tau}^2_{PM}$ and provide R code to obtain the estimator, based on the general algorithm of DerSimonian and Kacker (2007), who also recommended the PM estimator for its good properties.

For the meta-analysis of log-odds ratios, Bhaumik *et al*. (2012) proposed an improved PM estimator for rare adverse events by borrowing strength from all studies when estimating each sampling variance. They suggest instead of the conventional v_i to use

$$v_i^* = \frac{1}{n_{it}+1}\left[e^{-CGR_i-\bar{y}_{cor}+\frac{\tau^2}{2}} + 2 + e^{CGR_i+\bar{y}_{cor}+\frac{\tau^2}{2}}\right] + \frac{1}{n_{ic}+1}\left[e^{-CGR_i} + 2 + e^{CGR_i}\right],$$

where n_{it} and n_{ic} are the number of subjects assigned to the treatment and control group, respectively, in study i, CGR is the control group risk, $\bar{y}_{cor}$ is the simple average of the y_i values, and *cor* is referred to as the continuity correction. In particular, the authors added a positive constant *cor* to the observed frequencies, estimated the relative treatment effect, and then determined the optimal value of *cor* so as to retain an unbiased estimate. To obtain the improved PM estimator, the authors suggested applying the same process as in PM using weights $w_{i,RE}^* = 1/\left(\tau^2 + v_i^*\right)$. They concluded that this improved method reduces bias compared with the DL, DL2, and PM estimators. This approach could in principle be implemented for every between-study variance estimator.

3.8. Hartung and Makambi (HM) method

The HM method is a modification of the DL estimation method which does not require truncation to zero (Hartung and Makambi, 2003, 2002). Taking into account the quadratic form of the random variables y_i,

$$\frac{\sum w_{i,FE}\left(y_i - \hat{\mu}_{FE}\right)^2}{\sum w_{i,FE} - \frac{\sum w_{i,FE}^2}{\sum w_{i,FE}}} = \frac{Q}{\sum w_{i,FE} - \frac{\sum w_{i,FE}^2}{\sum w_{i,FE}}},$$

the HM method involves multiplying the quadratic form above by the factor $Q/(2(k-1)+Q)$ to ensure positivity. The estimator is therefore given by:

$$\hat{\tau}_{HM}^2 = \frac{Q^2}{(2(k-1)+Q)\left(\sum w_{i,FE} - \frac{\sum w_{i,FE}^2}{\sum w_{i,FE}}\right)}.$$

This is a non-iterative method that always produces positive values. Thorlund *et al*. (2011) carried out an empirical study comparing the DL method with the HM, REML, HO, and SJ estimators, and concluded that for small to moderate true between-study variance values, HM and SJ produce large estimates of τ^2.

3.9. Hunter and Schmidt (HS) method

The Hunter and Schmidt (2004) estimator is given by

$$\hat{\tau}_{HS}^2 = \max\left\{0, \frac{Q-k}{\sum w_{i,FE}}\right\}.$$

The HS estimation method has been shown to be negatively biased (Viechtbauer, 2005). The HS and ML estimators have similar MSEs, which in turn are lower than the MSEs of the DL, REML, and HO estimators. Nevertheless, if unbiasedness is considered to be of importance, then the HS estimator should be avoided (Viechtbauer, 2005).

3.10. Maximum likelihood (ML) method

The ML method is asymptotically efficient but requires an iterative solution (Hardy and Thompson, 1996; Thompson and Sharp, 1999). Based on the marginal distribution $y_i \sim N(\mu, v_i + \tau^2)$ the estimate $\hat{\tau}_{ML}^2$ is obtained by maximising the log–likelihood function

$$lnL\left(\mu, \tau^2\right) = -\frac{k}{2}\ln(2\pi) - \frac{1}{2}\sum \ln\left(v_i + \tau^2\right) - \frac{1}{2}\sum \frac{(y_i - \mu)^2}{\left(v_i + \tau^2\right)}.$$

Setting partial derivatives with respect to μ and τ^2 equal to zero and solving the likelihood equations for the two parameters to be estimated, the ML estimators for μ and τ^2 can be obtained by

$$\hat{\mu}_{RE}\left(\hat{\tau}_{ML}^2\right) = \frac{\sum w_{i,RE} y_i}{\sum w_{i,RE}}, \qquad (3)$$

$$\hat{\tau}^2_{ML} = \max\left\{0, \frac{\sum w^2_{i,RE}\left(\left(y_i - \hat{\mu}_{RE}\left(\hat{\tau}^2_{ML}\right)\right)^2 - v_i\right)}{\sum w^2_{i,RE}}\right\}, \tag{4}$$

where $w_{i,RE} = 1/\left(v_i + \hat{\tau}^2_{ML}\right)$. One way to perform the maximisation is to start with an initial estimate for $\hat{\tau}^2_{ML}$, which can be decided *a priori* as a plausible value of the between-study variance, or it can be estimated with any other non-iterative estimation method. Then the ML estimates are obtained by iterating over $\hat{\tau}^2_{ML}$ and $\hat{\mu}_{RE}\left(\hat{\tau}^2_{ML}\right)$ until they converge and do not change from one iteration to the next. In each iteration step, a negative between-study variance estimate is set equal to zero. The maximisation of the likelihood can also be performed using several techniques, such as the Newton–Raphson method, the method of scoring, the simplex method, or the expectation–maximisation (EM) algorithm.

A disadvantage of iterative estimators is that they depend on the choice of maximisation method, which might fail to converge to a solution, and hence the estimator does not provide an estimated τ^2 value. This is mainly because of the maximisation method selected and a potentially flat likelihood that is hard to maximise, which is more likely to happen when k is small. In such cases, one could apply one of the closed-form estimators or incorporate an informative prior on the between-study variance within a Bayesian framework (Pullenayegum, 2011; Rhodes *et al*., 2015; Turner *et al*., 2012). Likelihood-based methods are asymptotically unbiased, with variance approaching the Cramér–Rao lower bound. Hence, the ML and REML methods are asymptotically equivalent, but not in finite samples.

Simulation studies have suggested that although the ML estimator has a small MSE, it exhibits large negative bias for large τ^2 when k is small to moderate and small studies are included in the meta-analysis (Chung *et al*., 2014; Kontopantelis *et al*., 2013; Panityakul *et al*., 2013; Sidik and Jonkman, 2007; Viechtbauer, 2005). The method described above, also assumes effect estimates are normally distributed and there is currently little evidence to suggest how the ML estimator performs under non-normal conditions. Alternative forms of the likelihood can be used to relax the normality assumption, which result in different maximum likelihood estimators, but are beyond the scope of this paper.

It has been shown that the ML method has the smallest MSE in comparison to the REML, SJ, HO, and PM methods, but exhibits the largest amount of bias among them (Chung *et al*., 2014; Sidik and Jonkman, 2007; Thompson and Sharp, 1999; Swallow and Monahan, 1984). Another simulation study (Viechtbauer, 2005) showed that the ML and HS methods have approximately the same MSE across all values of k and τ^2 simulated, which in turn was lower than the MSE of the DL and REML methods. However, because of its downward bias, both Panityakul *et al*. (2013) and Viechtbauer (2005) recommended avoiding the ML estimator.

3.11. *Restricted maximum likelihood (REML) method*

The REML method can be used to correct for the negative bias associated with the ML method. The estimate $\hat{\tau}^2_{REML}$ is produced by setting the derivative of the restricted log-likelihood function (Raudenbush, 2009)

$$lnL\left(\tau^2\right) = -\frac{k}{2}\ln(2\pi) - \frac{1}{2}\sum\ln\left(v_i + \tau^2\right) - \frac{1}{2}\sum\frac{\left(y_i - \hat{\mu}_{RE}\left(\hat{\tau}^2_{ML}\right)\right)^2}{\left(v_i + \tau^2\right)} - \frac{1}{2}ln\left(\sum\frac{1}{\left(v_i + \tau^2\right)}\right),$$

with respect to τ^2 equal to zero and solving the resulting equation for τ^2. This yields

$$\hat{\tau}^2_{REML} = \max\left\{0, \frac{\sum w^2_{i,RE}\left(\left(y_i - \hat{\mu}_{RE}\left(\hat{\tau}^2_{ML}\right)\right)^2 - v_i\right)}{\sum w^2_{i,RE}} + \frac{1}{\sum w_{i,RE}}\right\},$$

where $w_{i,RE} = 1/\left(v_i + \hat{\tau}^2_{REML}\right)$ (DerSimonian and Laird, 1986; Sidik and Jonkman, 2007). Again, $\hat{\tau}^2_{REML}$ is calculated by a process of iteration with an initial estimate of $\hat{\tau}^2_{REML}\geq 0$. Each iteration step requires non-negativity.

Simulation studies suggested that the REML method underestimates τ^2 especially when the data are sparse (Goldstein and Rasbash, 1996; Novianti *et al*., 2014; Sidik and Jonkman, 2007, 2005b). For dichotomous outcome data, it has been shown that the REML estimator is less downwardly biased than the DL estimator but has greater MSE (Chung *et al*., 2014; Sidik and Jonkman, 2007). Viechtbauer (2005) used continuous simulated data to compare the DL, ML, REML, HS, and HO methods and calculated bias and MSE with the non-truncated estimates of τ^2. He showed that the REML estimator has smaller MSE than the HO estimator, larger MSE than the ML and HS estimators, and comparable MSE to the DL estimator. The same study showed that REML is the preferable approach when large studies are included in the meta-analysis. For continuous outcomes, Novianti *et al*. (2014) also suggest that REML may be a valid alternative than the DL method. Knapp and Hartung (2003) found that the REML estimator has lower variance than the DL and PM estimators. Jackson *et al*. (2010) investigated the asymptotic efficiency of the DL, ML, and REML methods and showed that ML estimation performs better for small amounts of between-study variance, whereas for large τ^2, the DL and REML estimators are more efficient. Chung *et al*. (2014) showed that the DL and REML methods produce similar proportions of zero estimates and lower than the ML estimator when k is small, but for large k, the DL, ML, and REML estimators are similar and lower in magnitude than the HO method. An empirical study (Thorlund *et al*., 2011) of 920 Cochrane reviews with dichotomous outcome data and meta-analyses including at least three studies showed that the REML estimator can be smaller or larger in magnitude than the DL method. This agrees with a simulation study comparing DL with REML estimates under several non-normal distributions for the effect measures, and suggests that REML is a computationally intensive iterative method and does not perform better than DL (Kontopantelis and Reeves, 2012).

3.12. Approximate restricted maximum likelihood (AREML) method

An approximate REML (AREML) estimate is also available and it is an iterative solution to (Morris, 1983; Sidik and Jonkman, 2007; Thompson and Sharp, 1999)

$$\hat{\tau}^2_{AREML} = \max\left\{0, \frac{\sum w^2_{i,RE}\left(\left(\frac{k}{k-1}\right)\left(y_i - \hat{\mu}_{RE}\left(\hat{\tau}^2_{AREML}\right)\right)^2 - v_i\right)}{\sum w^2_{i,RE}}\right\},$$

where $w_{i.RE} = 1/\left(v_i + \hat{\tau}^2_{AREML}\right)$. Although Thompson and Sharp (1999) describe the method as REML, this is an approximation of REML using a direct adjustment for the loss of degrees of freedom. The method yields almost identical estimates to REML (Sidik and Jonkman, 2007). In the scenario where the sampling variances are equal ($v_i = v$), AREML and REML estimates are identical.

3.13. Sidik and Jonkman (SJ) method

Sidik and Jonkman (2005b) introduced a non-iterative estimation method based on weighted least squares. To obtain the SJ estimator (also known as the model error variance estimator) we first calculate the values $\hat{q}_i = \hat{r}_i + 1$ with $\hat{r}_i = v_i/\hat{\tau}^2_0$ (assuming $\hat{\tau}^2_0 \neq 0$) where $\hat{\tau}^2_0 = \sum (y_i - \bar{y})^2/k$ is an initial estimate of the between-study variance. Then the SJ estimator is obtained by setting the quantity $\sum \hat{q}_i^{-1}\left(y_i - \hat{\mu}_{\hat{q},RE}\right)^2$ equal to its expected value

$$\hat{\tau}^2_{SJ} = \frac{1}{k-1}\sum \hat{q}_i^{-1}\left(y_i - \hat{\mu}_{\hat{q},RE}\right)^2,$$

where $\hat{\mu}_{\hat{q},RE} = \sum \hat{q}_i^{-1} y_i / \sum \hat{q}_i^{-1}$ is the weighted RE pooled estimate. An improvement on this estimation method has been recommended (Sidik and Jonkman, 2007), using $\hat{r}_i = v_i/\hat{\tau}^2_{HO}$ (if $\hat{\tau}^2_{HO} = 0$, we set $\hat{\tau}^2_{HO} = 0.01$). The method always yields a positive estimate of the between-study variance.

The SJ estimator has methodological similarities with the PM estimator. As in the PM method, the weights assigned to each study when estimating $\hat{\tau}^2_{SJ}$ can be re-expressed as $\hat{q}_i = \hat{r}_i + 1 = \left(v_i + \hat{\tau}^2_0\right)\left(\hat{\tau}^2_0\right)^{-1}$, that is, RE weights multiplied by the constant term $\hat{\tau}^2_0$. In practice, the SJ method differs from the PM estimator in being always positive and non-iterative.

Simulation studies suggested that the SJ estimation method has smaller MSE and substantially smaller bias than the DL estimator for large values of k and τ^2, whereas the opposite occurs when k and τ^2 are small (Sidik and Jonkman, 2007, 2005b). It was shown that the SJ method has smaller MSE compared with the HO method irrespective of the magnitude of k and τ^2, but that the latter performs better in terms of bias when τ^2 is small (Sidik and Jonkman, 2007). Additionally, the SJ estimator has been shown to have the largest bias among the DL, ML, REML, HO, and PM estimators for relatively small values of τ^2, with the bias decreasing as τ^2 increases (Novianti *et al*., 2014; Panityakul *et al*., 2013; Sidik and Jonkman, 2007). For large τ^2, the SJ and PM methods are the best estimators in terms of bias according to Sidik and Jonkman (2007). In agreement with these findings, an empirical study (Thorlund *et al*., 2011) showed that the SJ estimator produces larger estimates than the DL method.

3.14. Rukhin Bayes (RB) method

Under the assumption that v_i and τ^2 are random independent parameters and that the prior distribution of τ^2 is non-informative with large variance and mean $\hat{\tau}^2_{prior}$, Rukhin (2013) derived the general form of Bayes estimators:

$$\hat{\tau}^2_{RB} = \max\left\{0, \frac{\sum (y_i - \bar{y})^2}{k+1} + \frac{\left(\sum n_i - k\right)\left(2k\hat{\tau}^2_{prior} - (k-1)\sum v_i\right)}{\sum (n_i - k + 2)\, k(k+1)}\right\}, \tag{5}$$

where n_i is the number of subjects in study i and $\hat{\tau}^2_{prior}$ is the mean of the prior distribution of τ^2. He showed that estimators of class (5) have an inherent positive bias. Rukhin (2013) recommended this estimator for small to moderate k and more specifically to estimate $\hat{\tau}^2_{RB}$ with $\hat{\tau}^2_{prior} = 0$ (RB0) as an alternative to the DL method. Note that setting $\hat{\tau}^2_{prior} = 0.5(k-1)\sum (v_i/k)$ results in a positive estimator (RBp). A simulation study

(Kontopantelis *et al.*, 2013) for $k<5$ showed that RB0 had less bias than the DL, DLp, DLb, DL2, HO, HO2, REML, and SJ estimators.

3.15. Bayes Modal (BM) method

Chung *et al.* (2014, 2013) proposed the use of Bayes modal (BM) or maximum penalised likelihood estimators with a gamma prior $G(2, 10^{-4})$. To derive the BM estimator, they use the profile log-likelihood:

$$InL_p(\tau) = -\frac{k}{2}\ln(2\pi) - \frac{1}{2}\sum \ln\left(v_i+\tau^2\right) - \frac{1}{2}\sum \frac{\left(y_i - \frac{\sum\left(v_i+\tau^2\right)^{-1}y_i}{\sum\left(v_i+\tau^2\right)^{-1}}\right)^2}{\left(v_i+\tau^2\right)}.$$

Approximating $InL_p(\tau)$ using the ML estimator and a Taylor expansion, the BM estimator is obtained as

$$\hat{\tau}^2_{BM} = \begin{cases} Var(\hat{\tau}_{ML}) & ,\hat{\tau}_{ML}=0 \\ \left(\frac{\hat{\tau}_{ML}}{2} + \frac{\hat{\tau}_{ML}}{2}\sqrt{1+\frac{4Var(\hat{\tau}_{ML})}{\hat{\tau}^2_{ML}}}\right)^2 & ,\hat{\tau}_{ML}>0 \end{cases}$$

where $Var(\hat{\tau}_{ML})$ represents the estimated asymptotic variance of τ based on the Fisher information (see also 4.2). The method always yields positive estimates and larger values than $\hat{\tau}_{ML}$ (Chung *et al.*, 2014, 2013). It has also been shown that the REML and BM estimators provide similar results for large τ^2, but for small values of between-study variance, REML estimation underestimates τ^2 in contrast to the BM estimator. When τ^2 is zero, Chung *et al.* (2014) showed that the BM estimator overestimates the between-study variance and has larger bias than the DL, ML, REML, and HO estimators especially when the study sizes and k are small. When τ^2 is positive, the BM estimator has the lowest MSE, whereas for $\tau^2=0$, the BM estimator performed worse in terms of efficiency than the DL, ML, and REML estimators, but still better than the HO method.

3.16. Full Bayesian (FB) method

Estimates of the amount of between-study variance can also be obtained with fully Bayesian (FB) approaches, using Markov chain Monte Carlo (MCMC) methods (e.g. a Gibbs sampler) in specialised software such as WinBUGS (see Table 1) (Smith *et al.*, 1995). The FB approach is often preferable as it allows incorporation of uncertainty in all parameters (including τ^2), for which credible intervals can be derived from the posterior distribution not relying on asymptotic standard errors. However, several investigators claim that in practice the differences between frequentist and Bayesian approaches appear to be small (Morris, 1983; Thompson and Sharp, 1999). A simple hierarchical Bayesian model for meta-analysis is

$$\begin{aligned} y_i|\theta_i &\sim N(\theta_i, v_i), \\ \theta_i|\mu &\sim N(\mu, \tau^2), \\ \mu &\sim \pi_1(.), \\ \tau &\sim \pi_2(.), \end{aligned} \tag{6}$$

where $\pi_1(.)$ and $\pi_2(.)$ are prior distributions. The FB method uses non-informative priors to approximate a likelihood-based analysis. With many studies (large k) the choice of the prior distribution does not have a major influence on the results because the data dominate the analysis. However, the choice of prior distribution is important when the number of studies is small because it may impact on the estimated between-study variance and consequently estimation of the mean treatment effect (Lambert *et al.*, 2005; Senn, 2007).

A simulation study compared 13 different prior distributions for the between-study variance and suggested that the results might vary substantially when the number of studies is small (Lambert *et al.*, 2005). The study showed that, in terms of bias, none of the distributions considered performs best for all scenarios. More specifically, inverse-gamma, uniform, and Wishart distributions for the between-study variance perform poorly when k is small and produce estimates with substantial bias. The same study suggested that a uniform prior on τ performs better than other priors (e.g. uniform on log variance, inverse-gamma on variance, DuMouchel prior) in terms of bias and convergence problems. An inverse-gamma prior with small hyper-parameters is often considered to be an approximately non-informative prior, but it was shown that inferences can be sensitive to the choice of hyper-parameters (Chung *et al.*, 2013; Gelman, 2006). Chung *et al.* (2014) compared the BM estimator with a FB approach using the inverse-gamma and uniform prior distributions for τ, and found that for small τ^2 the inverse-gamma produces estimates with less bias and lower MSE than BM, but the opposite was observed for larger τ^2. The uniform prior had the largest bias and MSE among the three approaches. Informative priors were recently proposed for the between-study variance when meta-analysing (log) odds ratios (Pullenayegum, 2011; Turner *et al.*, 2012) and standardised mean differences (Rhodes *et al.*, 2015), and these might considerably improve estimation when few studies are included in the meta-analysis.

4. Confidence intervals for the between-study variance

4.1. *Profile likelihood confidence intervals (PL)*

Hardy and Thompson (1996) established the use of profile likelihood methods in meta-analysis. The profile likelihood (PL) method is based on the log-likelihood function and is an iterative process that provides CIs for the between-study variance parameter taking into account the fact that μ needs to be estimated as well. The log likelihood ratio statistic under the null hypothesis $H_0: \tau^2 = 0$ is (Hardy and Thompson, 1996):

$$U = -2ln\left[\frac{L(\hat{\mu}_{RE}(0), 0)}{L\left(\hat{\mu}_{RE}\left(\hat{\tau}^2_{ML}\right), \hat{\tau}^2_{ML}\right)}\right] \sim \chi^2_1.$$

We denote $\hat{\mu}_{RE}\left(\widetilde{\tau}^2\right)$ as the value estimated by formula (3) where $\widetilde{\tau}^2$ is used to calculate the RE model's study weights. A 95% CI for τ^2 can then be obtained by the set of τ^2 values satisfying (Jackson *et al.*, 2010; Viechtbauer, 2007):

$$lnL\left(\hat{\mu}_{RE}\left(\widetilde{\tau}^2\right), \widetilde{\tau}^2\right) > lnL\left(\hat{\mu}_{ML}, \hat{\tau}^2_{ML}\right) - \frac{3.84}{2}.$$

The method requires non-negativity at each iteration step. Viechtbauer (2007) found that profile likelihood CIs based on the restricted log-likelihood improves the coverage probability.

When $\tau^2 = 0$, the method produces large CIs with very high coverage probabilities, whereas as τ^2 increases, the coverage probabilities approach the nominal level (Viechtbauer, 2007). This is probably because the asymptotic distribution of the likelihood ratio statistic when $\tau^2 = 0$ is not χ^2_1, but a mixture of χ^2_1 and a probability mass of a random variable centred at zero (Viechtbauer, 2007). The method is implemented in Stata using the *metaan* (Kontopantelis and Reeves, 2010) command.

4.2. *Wald-type confidence intervals (Wt)*

Assuming asymptotic normality for the ML estimate, a 95% Wald-type (Wt) CI for τ^2 can be obtained as (Biggerstaff and Tweedie, 1997):

$$\hat{\tau}^2_{ML} \pm 1.96\sqrt{Var\left(\hat{\tau}^2_{ML}\right)}.$$

The normal asymptotic distribution for the ML method with mean τ^2 and variance equal to the inverse of the Fisher information suggests that

$$Var\left(\hat{\tau}^2_{ML}\right) = 2\left(\sum w^2_{i,RE}\right)^{-1}$$

and

$$\hat{\tau}^2_{REML} \pm 1.96\sqrt{Var\left(\hat{\tau}^2_{REML}\right)},$$

with

$$Var\left(\hat{\tau}^2_{REML}\right) = 2\left(\sum w^2_{i,RE} - 2\frac{\sum w^3_{i,RE}}{\sum w_{i,RE}} + \frac{\left(\sum w^2_{i,RE}\right)^2}{\left(\sum w_{i,RE}\right)^2}\right)^{-1}.$$

Because ML and REML estimates require non-negativity, the Wt CIs should always be non-negative. Therefore, a negative lower bound is truncated to zero. It is worth noting that the Wt CIs require a large k to perform well, as between-study variances are skewed and do not follow a normal distribution.

The method produces inaccurate CIs when the distributions of $\hat{\tau}^2_{ML}$ and $\hat{\tau}^2_{REML}$ are not adequately approximated by normal distributions (Stern and Welsh, 2000), which will be the case unless k is large (Goldstein, 1995). Therefore, for $\tau^2 > 0$, the Wt CIs are not expected to yield adequate coverage probabilities, whereas for $\tau^2 = 0$ the coverage probabilities are well above the nominal 95% level (Viechtbauer, 2007). The Wt CI is implemented in Stata using the *xtreg* (Gutierrez *et al.*, 2001) and in R using *metaSEM* package (Cheung, 2014).

4.3. Biggerstaff, Tweedie, and Jackson confidence intervals (BT, BJ, and Jackson)

4.3.1. Biggerstaff and Tweedie (BT) CI. Biggerstaff and Tweedie (1997) suggested to approximate the distribution of the Cochran's Q-statistic, $Q = (S_1 - (S_2/S_1))\hat{\tau}^2_{DL} + (k-1)$, where $S_r = \sum w^r_{i,FE}$ (see also section 3.1), using a gamma distribution with shape and scale parameters r and λ respectively. Using the RE model, they obtain

$$E(Q) = \left(S_1 - \frac{S_2}{S_1}\right)\tau^2 + (k-1),$$

$$Var(Q) = 4\left(S_1 - \frac{S_2}{S_1}\right)\tau^2 + 2\left(S_2 - 2\frac{S_3}{S_1} + \frac{S_2^2}{S_1^2}\right)\tau^4 + 2(k-1).$$

It follows that $E(Q) = r\lambda$ and $Var(Q) = r\lambda^2$, $r(\tau^2) = (E(Q))^2/Var(Q)$, and $\lambda(\tau^2) = Var(Q)/E(Q)$. Denoting with $f(x|r(\tau^2))$ the density function of the gamma distribution with shape parameter $r(\tau^2)$ and scale parameter 1, the 95% Biggerstaff and Tweedie (BT) CI can be obtained as the solutions of the equations:

$$\left(\int_{\frac{Q}{\lambda(\tilde{\tau}^2)}}^{\infty} f\left(x|r\left(\tilde{\tau}^2\right)\right)dx = 0.025, \quad \int_{0}^{\frac{Q}{\lambda(\tilde{\tau}^2)}} f\left(x|r\left(\tilde{\tau}^2\right)\right)dx = 0.025\right),$$

To solve these, an iterative procedure is followed as τ^2 varies, and non-negativity is required at each iteration step. In the case that $\int_0^{Q/\lambda\left(\tilde{\tau}^2\right)} f\left(x|r\left(\tilde{\tau}^2\right)\right)dx < 0.025$, the interval is set equal to the null set. Obviously, this method depends on how well the gamma distribution approximates the true distribution of Q. SAS code to obtain the CI was provided by the authors (Biggerstaff and Tweedie, 1997).

4.3.2. Biggerstaff and Jackson (BJ) CI. An extension of the BT CI was later proposed by Biggerstaff and Jackson (2008) (BJ), using the cumulative distribution function of Q, $F_Q(x;\tau^2)$. The authors noted that the non-truncated version of the DL estimator is a linear function of Q, and that the HM estimator is a simple function of Q. Thus, they express the cumulative distribution function $F_Q(x;\tau^2)$ as:

$$F_{\hat{\tau}^2_{DL}}(x;\tau^2) = F_Q\left(s\hat{\tau}^2_{DL} + k - 1;\tau^2\right)$$

for the untruncated DL estimator and

$$F_{\hat{\tau}^2_{HM}}(x;\tau^2) = F_Q\left(s\frac{x}{2} + \sqrt{2(k-1)sx + \left(\frac{sx}{2}\right)^2}\right) - F_Q\left(s\frac{x}{2} - \sqrt{2(k-1)sx + \left(\frac{sx}{2}\right)^2}\right)$$

for the HM estimator of τ^2 with $s = \sum w_{i,FE} - \left(\sum w^2_{i,FE}/\sum w_{i,FE}\right)$. A 95% CI for the between-study variance can be obtained as the solutions of the equations (Biggerstaff and Jackson, 2008):

$$\left(1 - F_Q\left(x;\tau^2\right) = 0.025, \;\; F_Q\left(x;\tau^2\right) = 0.025\right).$$

When $1 - F_Q(x;\tau^2 = 0) > 0.025$ the lower bound of CI is set equal to zero.

The cumulative distribution function $F_Q(x;\tau^2)$ may be computed using Farebrother's algorithm (Farebrother, 1984) for positive linear combination of chi-squared random variables. The method is implemented in R using the *CompQuadForm* (Jackson, 2013) and *metaxa* packages (Preuß and Ziegler, 2014). Preuß and Ziegler (2014) presented a simplified version of the method that does not require the calculation of the density of Cochran's Q, but only the calculation of the cumulative distribution function, which reduces computation time.

4.3.3. Jackson CI. A generalisation of the BJ CI was recently suggested by Jackson (2013) and uses the Q_a statistic (see section 3.1). One option is to use $a_i = w_{i,FE}$ but other weighting schemes are also possible. Jackson (2013) showed that Q_a, like Q, is distributed as a linear combination of χ^2 random variables so that methods similar to Biggerstaff and Jackson can be used. The cumulative distribution function of $Q_a(F_{Q_a}(x;\tau^2))$ is a continuous and strictly decreasing function of τ^2 and Jackson suggested obtaining the 95% CI as:

$$\left(1 - F_{Q_a}\left(x;\tau^2\right) = 0.025, \; F_{Q_a}\left(x;\tau^2\right) = 0.025\right)$$

In the case that $F_{Q_a}(x;\tau^2 = 0) < 0.025$, the interval is equal to the null or empty set. In such cases, Jackson (2013) suggests presenting the interval [0, 0] and recommends interpreting the finding as 'the data appear to be highly homogeneous' or 'the interval estimation fails'. If $1 - F_{Q_a}(x;\tau^2 = 0) > 0.025$, the lower bound of CI is set equal to zero.

Jackson (2013) also investigated weights of the form $a_i = 1/(v_i + x)^p$ and carried out a simulation study where x and p are constants. He showed that the *Q*-profile (QP) method provides narrower CIs than BJ method for large τ^2, and vice versa for small τ^2. For moderate τ^2, he recommends using the Jackson method with weights equal to the reciprocal of the within-study standard errors $\left(a_i = 1/\sqrt{v_i}\right)$, i.e. $x = 0$ and $p = 1/2$. The method provides an exact CI under the RE model's assumptions. The CI approaches based on the generalised Cochran between-study variance statistic are competitors to the QP method (see section 4.4) and may provide shorter CIs. The method can be implemented in R software via the *CompQuadForm* (Duchesne and Lafaye De Micheaux, 2010) and *metafor* packages. R code is also provided by the author (Jackson, 2013).

4.4. *Q-profile confidence intervals (QP)*

The QP method is based on the generalised *Q*-statistic ($Q_{gen}(\tau^2)$, see section 3.7), which follows a χ^2_{k-1} distribution, so that (Viechtbauer, 2007)

$$P\left(\chi^2_{k-1,0.025} \le Q_{gen}\left(\tau^2\right) \le \chi^2_{k-1,0.975}\right) = 0.95.$$

Employing the inversion principle, we can derive a 95% CI for τ^2 as

$$\left(Q_{gen}\left(\tilde{\tau}^2\right) = \chi^2_{k-1,0.975}, Q_{gen}\left(\tilde{\tau}^2\right) = \chi^2_{k-1,0.025}\right),$$

where the two $\tilde{\tau}^2$ values are iteratively computed from $Q_{gen}\left(\tilde{\tau}^2\right)$, so that the lower and upper critical values of the distribution are reached. Note that non-negativity is required at each iteration step. If $Q_{gen}\left(\tilde{\tau}^2 = 0\right) < \chi^2_{k-1,0.025}$ then the upper bound of the CI falls below zero and no CI can be provided for τ^2 resulting in the null set ([0, 0]). It is also possible that the CI does not actually contain the estimate of the between-study variance. However, the QP method in conjunction with the PM estimator prevents such strange cases (Viechtbauer, 2013). It is also worth pointing out that under the assumptions of the RE model, the method provides an exact CI rather than an approximation.

The method has been suggested also by Hartung and Knapp (2005) for the RE model in the analysis of variance. Knapp *et al.* (2006) suggested a modified QP method to determine the lower bound of the interval using a different weighting scheme for the generalised *Q*-statistic. For small τ^2, the modified QP is closer to the nominal level, whereas for large τ^2 the two approaches provide very similar estimates (Knapp *et al.*, 2006).

Knapp *et al.* (2006) state that the QP method is preferable to the BT and PL methods with respect to attaining a predefined confidence level. Viechtbauer (2007) showed that the method guarantees nominal coverage probabilities even in small samples and that it yields the most accurate coverage probabilities among the BT, PL, Wt parametric, and the non-parametric bootstrap methods. The method is implemented in R software in the *metafor* (Viechtbauer, 2013) package. Bowden *et al.* (2011) also provide an R function to calculate intervals using the Q-profile method.

4.5. *Sidik and Jonkman confidence intervals (SJ)*

Sidik and Jonkman (2005b) proposed a method based on the SJ estimator and the 2.5th–97.5th quantiles of the χ^2_{k-1} distribution:

$$\left(\frac{(k-1)\hat{\tau}^2_{SJ}}{\chi^2_{k-1,0.975}}, \frac{(k-1)\hat{\tau}^2_{SJ}}{\chi^2_{k-1,0.025}}\right).$$

As $\hat{\tau}^2_{SJ}$ takes non-negative values, the interval should also be non-negative. Simulation studies suggest that the Sidik and Jonkman (SJ) intervals have very poor coverage probability when τ^2 is small, but as k and τ^2 increase the coverage probability gets close to the nominal value (Sidik and Jonkman, 2005b; Viechtbauer, 2007). Consequently, Sidik and Jonkman (2005b) recommend this method only when there is strong evidence that the between-study variance is moderate to large. However, when $\tau^2 = 0$ the method never captures the true value and Viechtbauer (2007) raised concerns about the coverage for the estimator.

4.6. *Bootstrap confidence intervals*

The bootstrap approach yields parametric (Turner *et al.*, 2000) and non-parametric (Switzer *et al.*, 1992) CIs. For any non-negative estimator of τ^2, the parametric bootstrap CIs can be constructed by generating k values from the distribution $y_i \sim N(\hat{\mu}_{RE}(\hat{\tau}^2), \hat{\tau}^2 + v_i)$. Then the between-study variance is estimated based on the bootstrap sample. After repeating this process B times, the CI is constructed by taking the 2.5th and 97.5th percentiles of the distribution of $\hat{\tau}^2$ values. The non-parametric bootstrap CIs are obtained via a similar process, where k studies are sampled with replacement from the observed set of y_i and v_i values. For each bootstrap sample, τ^2 can be estimated using any estimator. Repeating the process B times, a 95% CI is given by the 2.5th and 97.5th percentiles of the B $\hat{\tau}^2$ values. The advantage of the non-parametric bootstrap method is that it relaxes the distributional assumptions about the observed treatment effects. However, Viechtbauer (2007) showed that bootstrap methods yield CIs with coverage probabilities that deviate substantially from the nominal level.

4.7. Bayesian credible intervals

Bayesian credible intervals for the between-study variance can be obtained within a Bayesian framework using specialised software such as WinBUGS (Smith *et al*., 1995). As with the point estimate of the between-study variance, the prior selection can impact a lot on the estimated credible interval when few studies are included in the meta-analysis.

5. Illustrative examples

We illustrate the statistical methods discussed in this paper using data from a recent review by the MRC Clinical Trials Unit (Bowden *et al*., 2011). We select four meta-analyses from this review, named Sarcoma (with 14 trials), Cervix2 (with 18 trials), NSCLC1 (with 17 trials), and NSCLC4 (with 11 trials). The Sarcoma meta-analysis assessed whether adjuvant chemotherapy improves survival in patients with localised soft-tissue sarcoma; Cervix2 assessed whether chemoradiotherapy improves survival in women with cervical cancer compared with radiotherapy; NSCLC1 evaluated the effect of cytotoxic chemotherapy on survival patients with non-small cell lung cancer; and NSCLC4 compared supportive care plus chemotherapy with supportive care alone in patients with advanced non-small cell lung cancer. We selected these on the basis of values of the I^2 index, which measures the amount of between-study variance as a percentage of the total variation in point estimates of the treatment effect (Higgins and Thompson, 2002). The four meta-analyses represent very low (Sarcoma: $I^2=0\%$), low (Cervix2: $I^2=18\%$), moderate (NSCLC1: $I^2=45\%$), and substantial (NSCLC4: $I^2=75\%$) between-study variance across studies. All analyses are performed with respect to the primary outcome of overall survival, and the log hazard ratio estimates were available from each trial. We combined the aggregated data in a RE model using the different estimation methods for the between-study variance and its uncertainty.

In the Sarcoma meta-analysis ($I^2=0\%$) only the SJ, BM, HM, RBp, and FB estimates yielded values larger than zero (Table 4). In examples with low to moderate between-study variance, the RBp method produced the highest $\hat{\tau}^2$ value, followed by the SJ estimator. In agreement with the simulation studies, for examples with moderate and large between-study variance (as measured by I^2), the HS and ML methods estimated lower $\hat{\tau}^2$ values compared with the other methods. For the example with large I^2, the SJ, HO, HO2, and RBp methods estimated large between-study variance values with HO yielding the largest one, followed by SJ. The differences between the estimators can be explained to some extent by the different weighting schemes they use. For example, most methods include both the within-study variances and an estimate of the between-study variance in the weights, whereas the DL and DLp methods use only the inverse of the within-study variances while the HO estimator uses

Table 4. Estimation of the between-study variance using different methods. Four meta-analyses are considered that represent zero (Sarcoma: $I^2=0\%$), low (Cervix2: $I^2=18\%$), moderate (NSCLC1: $I^2=45\%$), and high (NSCLC4: $I^2=75\%$) between-study variance

	No between-study variance	Low between-study variance	Moderate between-study variance	High between-study variance
DerSimonian and Laird (DL)	0.0000	0.0148	0.0238	0.1320
Positive DerSimonian and Laird (DLp)	0.0100	0.0148	0.0238	0.1320
Two-step DerSimonian and Laird (DL2)	0.0000	0.0130	0.0362	0.1817
Hedges and Olkin (HO)	0.0000	0.0000	0.0366	0.2243
Two-step Hedges and Olkin (HO2)	0.0000	0.0148	0.0389	0.1932
Paule and Mandel (PM)	0.0000	0.0132	0.0393	0.1897
Hartung and Makambi (HM)	0.0170	0.0305	0.0553	0.1732
Hunter and Schmidt (HS)	0.0000	0.0100	0.0190	0.1122
Maximum likelihood (ML)	0.0000	0.0151	0.0152	0.1314
Restricted maximum likelihood (REML)	0.0000	0.0201	0.0219	0.1560
Sidik and Jonkman (SJ)	0.0691	0.0469	0.0650	0.2091
Positive Rukhin Bayes (RBp)	0.1500	0.1132	0.1199	0.1970
Full Bayes * (FB)	0.0113	0.0216	0.0256	0.1838
Bayes Modal (BM)	0.0194	0.0308	0.0293	0.1649
Non-parametric Bootstrap DerSimonian and Laird (DLb)	0.0000	0.0120	0.0231	0.1250

*Half normal prior is used ($\tau \sim N(0,10^4), \tau \geq 0$).

equal weights for all trials. The PM approach yields larger estimates than the DL method for moderate to large between-study variance; REML estimation provides values similar to the DL estimator for low to moderate between-study variance and larger for high between-study variance.

In Figure 1, we present the estimated CI for τ^2 using the methods described in section 4. We display the results of the four different examples discussed above. The point estimate is calculated with the ML estimator for the PL and Wt CI methods, with the SJ estimator for the Sidik and Jonkman CI method, with the non-parametric DL estimator for the non-parametric bootstrap CI method, with the FB approach for the Bayesian credible intervals, and with the DL estimator for all other CI methods. For the method suggested by Jackson (2013), we used weights equal to the reciprocal of the within-study standard errors (i.e. with $x = 0$ and $p = 1/2$). Generally, the non-parametric bootstrap method provides the narrowest CIs, followed by the Wt approach. In all four examples, the QP method produces comparable CIs to the BJ and Jackson methods, whereas the SJ approach always produces CIs that do not include zero.

6. Comparative evaluation of the methods and recommendations

We have presented 16 methods to estimate between-study variance and seven methods to present uncertainty around the estimate. Published articles suggested that the different estimators can provide noticeably different or even conflicting results and their performance might vary regarding bias and MSE.

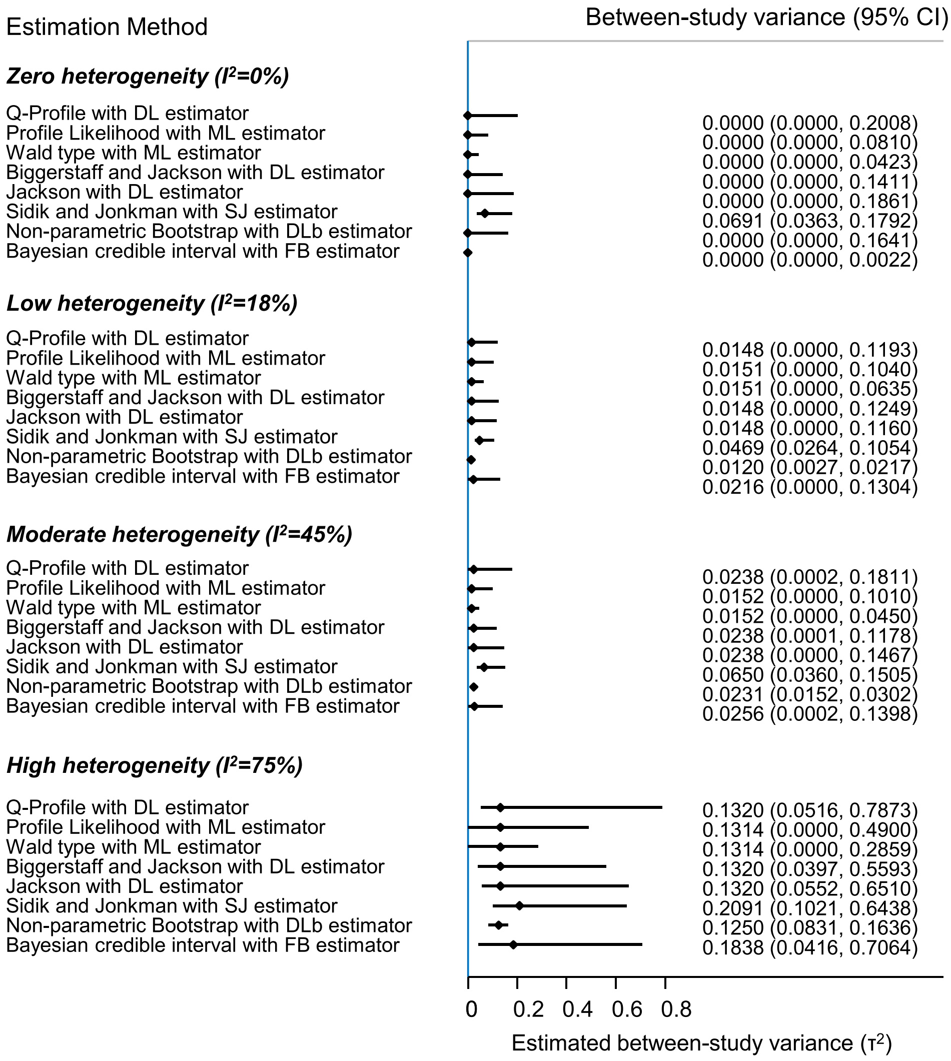

Figure 1. Confidence intervals for the between-study variance for four meta-analyses that represent zero (Sarcoma: $I^2 = 0\%$), low (Cervix2: $I^2 = 18\%$), moderate (NSCLC1: $I^2 = 45\%$), and high (NSCLC4: $I^2 = 75\%$) between-study variance. The between-study variance in the full Bayesian method was estimated using a half normal prior ($\tau \sim N(0, 10^4), \tau \geq 0$). DL: DerSimonian and Laird, DLb: Non-parametric bootstrap DerSimonian and Laird, ML: maximum likelihood, SJ: Sidik and Jonkman, FB: Full Bayes.

6.1. *Estimating the between-study variance*

Selection of the most appropriate estimator might depend on (1) whether a zero value of between-study variance is considered plausible or possible, (2) the properties of the estimators in terms of bias and efficiency, which may themselves depend on the number of studies included and the magnitude of the true between-study variance, and (3) the ease of application (which generally favours non-iterative methods).

The DLp, HM, SJ, RBp, and BM estimators always yield a positive estimate of the true between-study variance. Simulation studies suggest that this results in overestimation of τ^2 when the between-study variance is small to moderate or when the number of studies included in the meta-analysis is small. In contrast, the non-negative likelihood estimators and DL method tend to underestimate τ^2. Consequently, although the DL, ML, and REML estimator have smaller MSEs than the always-positive SJ estimator, they are only recommended when the true between-study variance in effect measures is relatively small (Sidik and Jonkman, 2007). An empirical study (Kontopantelis *et al*., 2013) showed that non-negative methods perform well on average, but produce biased results for meta-analyses with few studies where positive between-study variance methods are to be preferred. Novianti *et al*. (2014) compared the estimators DL, DL2, PM, HO, REML, and SJ in a simulation study for sequential meta-analysis when true between-study variance is zero, and showed that all methods overestimate τ^2, with the DL, PM, and REML estimators having the best properties. Thompson and Sharp (1999) compared the DL, ML, REML, and FB estimators and concluded that the FB method may produce inflated estimates when τ^2 is close to zero and that REML estimation is the most appropriate estimation technique.

A standard assumption in meta-analysis is that the within-study variances are fixed and known, but in reality they need to be estimated. The non-negative DL, HO, REML, and PM estimators are approximately unbiased (Panityakul *et al*., 2013; Viechtbauer, 2005) when the within-study variances v_i are assumed to be known. The HO estimator is also approximately unbiased when the within-study variances are not known, but can be estimated unbiasedly. Still, over-estimates of within-study variances can lead to negative estimates of the between-study variance, requiring truncation. On the other hand, when the sampling variances are exactly known, truncation might also be needed for small *k*, so that estimates of τ^2 have a large variance. Truncation introduces positive bias which possibly counteracts the negative bias because of overestimation of the sampling variances. Hence, bias because of truncation depends on how well we estimate the within-study sampling variances and the number of studies included in the meta-analysis.

Generally, studies have found that bias increases for small number of studies and large within-study variances. Viechtbauer (2005) recommended using REML estimation as the best approach in terms of bias and efficiency compared with the DL, ML, HS, and HO methods. The simulation study by Panityakul *et al*. (2013) suggested that the PM estimator is less biased than the DL and REML methods. Empirical evidence illustrates that the between-study variance might vary with different effect measures (Deeks, 2002; Engels *et al*., 2000; Friedrich *et al*., 2011). It is therefore possible that the performance of the estimators might differ according to the type of outcome data. Novianti *et al*. (2014) recommended the use of the PM estimators for both dichotomous and continuous outcome data, while stated that DL2 for both outcome data and REML for continuous data are valid alternatives as well. All between-study variance estimators described in this review can be applied with any type of outcome data and effect measure. Additional estimators have been proposed that can be applied to specific effect measures, e.g. the estimator suggested by Malzahn *et al*. (2000) and the special form of HO estimator (Hedges and Olkin, 1985), which can be applied under the standardised mean difference only.

Empirical studies have shown that most meta-analyses using the log-odds ratio effect measure yield $\tau^2 \leq 0.4$, and that the majority of meta-analyses are informed by fewer than ten studies (Pullenayegum, 2011; Rhodes *et al*., 2015; Turner *et al*., 2012). In such cases, evidence from simulation studies shows that the HM, HO, and SJ methods overestimate τ^2 (Panityakul *et al*., 2013; Sidik and Jonkman, 2007); the BM estimator performs worse than the DL estimator (Chung *et al*., 2014); the ML method is associated with substantial negative bias (Panityakul *et al*., 2013; Sidik and Jonkman, 2007); REML estimation is less downwardly biased than the DL and ML estimators with greater MSE though (Sidik and Jonkman, 2007); and the PM estimator is less downwardly biased than the DL or REML methods (Panityakul *et al*., 2013). After joint consideration of all empirical and simulation studies, we conclude that for both dichotomous and continuous outcome data the PM estimator has good performance in most studied scenarios, and for continuous data the REML estimator appears to be preferable compared to other alternatives (Bowden *et al*., 2011; DerSimonian and Kacker, 2007; Novianti *et al*., 2014; Panityakul *et al*., 2013).

6.2. *Estimating confidence intervals for the between-study variance*

Jackson (2013), Knapp *et al*. (2006), and Viechtbauer (2007) are, to the best of our knowledge, the only studies that compare methods for estimating the uncertainty around τ^2, applicable for any outcome data type. Tian (2008), suggested an additional method to the approaches presented in section 4, based on generalised inference, that can only be applied for continuous outcome data with scaled χ^2—distributed within-study variances. The method yields adequate coverage irrespective the magnitude of the between-study variance, and the number of studies included in a meta-analysis. The same simulation study showed that Tian's (2008) method yields on average comparable coverage and CI length with the modified QP proposed by Knapp *et al*. (2006), but for small τ^2 values the modified QP tends to have larger CI length and coverage rate.

Viechtbauer (2007) compared the QP, PL, Wt, BT, SJ, parametric, and non-parametric bootstrap methods. He showed that the bootstrap methods have less than adequate coverage probabilities and that the QP and BT methods yield the most accurate coverage probabilities even for small meta-analyses (Viechtbauer, 2007). However, both the QP and BT methods can result in null sets, especially when there is low between-study variance and the number of studies included in the meta-analysis is small (Viechtbauer, 2007). Viechtbauer (2013) suggested employing the PM estimator of τ^2 in order to avoid estimating a CI obtained with the QP method that does not include the point estimate. The PL, Wt, and bootstrap CIs, in contrast to the QP, BT, and SJ CIs, only work well for large meta-analyses as they are based on asymptotic results. Knapp *et al*. (2006) simulated data to compare the Wt, BT, PL, and QP CIs, and showed that the BT method was rather conservative for large τ^2, whereas the PL approach was extremely conservative for small τ^2. Jackson (2013) found that the QP method produces narrower CIs than the BJ approach for moderate to large between-study variance, whereas for lower τ^2, CIs obtained with the BJ method should be preferable. The Jackson method with weights equal to the reciprocal of the within-study standard errors appears to be a reasonable alternative that outperforms the QP and BJ CIs for small τ^2. However, the assumption of known and fixed within-study variances impacts on the study weights estimation and the chi-square approximation, and hence on the performance of these CI methods, especially for small studies.

It should be noted that the various methods for calculating confidence intervals are not all appropriate for all of the available between-study variance estimators. In Table 5, we summarise all combinations between the

Table 5. Summary of our proposals for appropriate combinations of approaches for estimating and calculating confidence intervals for the between-study variance.

	Confidence interval for the between-study variance methods						
Between-study variance estimators	Profile Likelihood (PL)	Wald-type (Wt)	Biggerstaff, Tweedie and Jackson (BT, BJ, Jackson)	Q-Profile (QP)	Sidik and Jonkman (SJ)	Bootstrap	Bayesian Credible Intervals
Method of moments estimators							
DerSimonian and Laird (DL)	—	✓	✓	(✓)	—	✓	—
Positive DerSimonian and Laird (DLp)	—	✓	✓	(✓)	—	✓	—
Two-step DerSimonian and Laird (DL2)	—	✓	✓	(✓)	—	✓	—
Hedges and Olkin (HO)	—	✓	✓	(✓)	—	✓	—
Two-step Hedges and Olkin (HO2)	—	✓	✓	(✓)	—	✓	—
Paule and Mandel (PM)	—	✓	(✓)	✓	—	✓	—
Hartung and Makambi (HM)	—	✓	✓	(✓)	—	✓	—
Hunter and Schmidt (HS)	—	✓	(✓)	(✓)	—	✓	—
Maximum Likelihood estimators							
Maximum likelihood (ML)	✓	✓	(✓)	(✓)	—	✓	—
Restricted maximum likelihood (REML)	✓	✓	(✓)	(✓)	—	✓	—
Approximate restricted maximum likelihood (AREML)	✓	✓	(✓)	(✓)	—	✓	—
Model error variance estimator							
Sidik and Jonkman (SJ)	—	✓	(✓)	(✓)	✓	✓	—
Bayes estimators							
Rukhin Bayes (RB)	—	✓	(✓)	(✓)	—	✓	✓
Positive Rukhin Bayes (RBp)	—	✓	(✓)	(✓)	—	✓	—
Full Bayes (FB)	—	—	—	—	—	—	✓
Bayes Modal (BM)	—	✓	(✓)	(✓)	—	✓	—
Bootstrap estimator							
Non-parametric bootstrap DerSimonian and Laird (DLb)	—	✓	(✓)	(✓)	—	✓	—

Pairwise combinations are categorised in three groups: a) confidence intervals naturally paired with the between-study variance estimator, ✓; b) confidence intervals paired in principle with the between-study variance estimator, but not naturally, (✓); c) confidence intervals considered unlikely compatible with the between-study variance estimator, —.

approaches to estimate the uncertainty around the between-study variance and estimators for the between-study variance, that we suggest are likely to be considered appropriate in practice. We use three different groups to classify each pairwise combination. The classification includes combinations of approaches for estimating and calculating CIs for the between-study variance that (1) are based on the same statistical principle and can be naturally paired, (2) could be paired in principle, but not naturally, and (3) are unlikely to be considered compatible. It is possible some pairwise combinations that we consider compatible in Table 5 yield CIs not containing the estimate of the between-study variance, and we would not recommend presenting them together in such instances. The QP method is the default for all frequentist methods in the *metafor* package (Viechtbauer, 2013) (except when using some version of the generalised method of moments estimator for τ^2) and so this is deemed widely appropriate in Table 5; because Jackson's method is a competitor to the QP method, this alternative is deemed appropriate for the same range of possibilities as the QP method (and is used by default when using the generalised method of moments estimator). The suggestions for compatible methods for point and interval estimation in Table 5 are only tentative and we anticipate that further refinement is both possible and desirable.

6.3. Recommendations

To conclude, we offer tentative recommendations for practice based on the existing simulation and empirical studies, and informed by the consensus of the authors. The results of the available studies depend on the particular scenarios they investigated and properties of the methods they examined. Many approaches presented in this review have not been compared under the same simulation settings, and hence making any clear recommendations about these methods is difficult. Also, the selection of the most preferable methods to calculate a confidence interval for the between-study variance is mostly based on coverage, because this was consistently reported in the identified simulation studies. Further research is required to evaluate the important properties (i.e. bias, efficiency, complexity, coverage, and precision) of all promising methods under the same, realistic, scenarios through a comprehensive simulation study.

In Appendix Table 1, we summarise scenarios examined by studies that compare between-study variance estimators, and in Appendix Table 2, we present the results as described on average for each pairwise comparison of the estimation methods. There is limited evidence to inform which estimator performs best, in particular when the number of studies is low (<5) and when the normality assumption does not hold; and the fully Bayesian estimator has not been evaluated extensively in comparative studies. When estimation of the between-study variance is the aim of the meta-analysis, a sensitivity analysis using a variety of suitable methods for estimating τ^2 and its CI might be needed, particularly when studies are few in number. In Appendix Table 3, we present the simulation results with respect to the properties of all CI methods as described in the three comparative studies (Jackson, 2013; Knapp *et al*., 2006; Viechtbauer, 2007).

Overall, the popular semi-parametric DL method appears to perform adequately. However, according to the summarised findings of the simulation studies, the PM method appears to have a more favourable profile among other estimators of the between-study variance in meta-analysis, including DL. It is easy to calculate, does not require distributional assumptions, (Bowden *et al*., 2011; DerSimonian and Kacker, 2007), and has been shown to be less biased and more efficient than many alternatives. For estimation of a confidence interval for the between-study variance, a good option for large τ^2 appears to be the QP method, whereas for small τ^2 the BJ method offers a reasonable approach (Jackson, 2013). The PM method in conjunction with the QP approach for obtaining CIs fit naturally together and appear to perform satisfactorily under several realistic scenarios. For continuous outcomes, we also advocate the use of REML estimation as a preferable alternative to the DL estimator, in agreement with other recent recommendations (Novianti *et al*., 2014; Viechtbauer, 2005).

Our recommendations on the estimation of the between-study variance are mostly based on non-Bayesian approaches, as Bayesian estimators have not been fully investigated in simulations. However, Bayesian estimators might be considered preferable to classical estimators in situations where appropriate prior information is available and is considered suitable for use in analysis.

Our review has identified gaps and deficiencies in the existing literature, and hopefully facilitates investigators in forming their own judgements about the most appropriate method for their needs. When additional evidence becomes available, we plan to update this review and our recommendations accordingly.

7. Competing of interests

The authors declare that they have no competing interests.

8. Authors' contributors

AAV, DJ, WV, RB, JB, GK, OK, JH, DL, and GS contributed to the conception and design of the study, and helped to draft the manuscript. AAV conducted the statistical analysis. All authors read and approved the final manuscript.

9. Funding

AAV is funded by the CIHR Banting Postdoctoral Fellowship Program. JB is supported by an MRC Methodology Research Fellowship grant (code MR/L012286/1). DJ is employed by the UK Medical Research Council (code U105260558). DL is funded by the Centre for Reviews and Dissemination, University of York. RB is employed by the Institute for Quality and Efficiency in Health Care, Cologne, Germany. GS receives funding from the European Research Council (IMMA, grant Nr 260559).

Acknowledgements

The authors would like to thank the associate editor Christopher H. Schmid for his insightful comments and support during the review process. They would also like to thank reviewer Kepher Makambi and an anonymous referee for their constructive suggestions and comments that greatly contributed to improving the final version of the manuscript.

References

Bax L. 2011. MIX 2.0—professional software for meta-analysis in Excel. BiostatXL.

Berkey CS, Hoaglin DC, Mosteller F, Colditz GA. 1995. A random-effects regression model for meta-analysis. *Statistics in Medicine* **14**: 395–411. DOI:10.1002/sim.4780140406.

Bhaumik DK, Amatya A, Normand S-L, Greenhouse J, Kaizar E, Neelon B, Gibbons RD. 2012. Meta-analysis of rare binary adverse event data. *Journal of the American Statistical Association* **107**: 555–567. DOI:10.1080/01621459.2012.664484.

Biggerstaff BJ, Jackson D. 2008. The exact distribution of Cochran's heterogeneity statistic in one-way random effects meta-analysis. *Statistics in Medicine* **27**: 6093–6110. DOI:10.1002/sim.3428.

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

Böhning D, Malzahn U, Dietz E, Schlattmann P, Viwatwongkasem C, Biggeri A. 2002. Some general points in estimating heterogeneity variance with the DerSimonian–Laird estimator. *Biostat. Oxf. Engl.* **3**: 445–457. DOI:10.1093/biostatistics/3.4.445.

Borenstein M, Hedges L, Higgins J, Rothstein H. 2005. *Comprehensive Meta-Analysis Version 2*. Englewood, NJ: Biostat **104**.

Borenstein M, Hedges LV, Higgins J, Rothstein H. 2009. *Introduction to Meta-Analysis*. John Wiley & Sons, Ltd, Chichester, UK.

Bowden J, Tierney JF, Copas AJ, Burdett S. 2011. Quantifying, displaying and accounting for heterogeneity in the meta-analysis of RCTs using standard and generalised Q statistics. *BMC Medical Research Methodology* **11**: 41. DOI:10.1186/1471-2288-11-41.

Bradburn MJ, Deeks JJ, Berlin JA, Russell Localio A. 2007. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Statistics in Medicine* **26**: 53–77. DOI:10.1002/sim.2528.

Brockwell SE, Gordon IR. 2001. A comparison of statistical methods for meta-analysis. *Statistics in Medicine* **20**: 825–840. DOI:10.1002/sim.650.

Cheung MWL. 2014. metaSEM: an R package for meta-analysis using structural equation. *Frontiers in Psychology* **5**: 1521. Available from: https://www.yumpu.com/en/document/view/17500684/metasem-an-r-package-for-meta-analysis-using-strutural-equation- (Accessed 4.8.14).

Chung Y, Rabe-Hesketh S, Choi I-H. 2014. Avoiding zero between-study variance estimates in random-effects meta-analysis. *Statistics in Medicine* . DOI:10.1002/sim.5821.

Chung Y, Rabe-Hesketh S, Dorie V, Gelman A, Liu J. 2013. A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika* **78**: 685–709. DOI:10.1007/s11336-013-9328-2.

Cochran WG. 1954. The combination of estimates from different experiments. *Biometrics* **10**: 101. DOI:10.2307/3001666.

Cornell JE, Mulrow CD, Localio R, Stack CB, Meibohm AR, Guallar E, Goodman SN. 2014. Random-effects meta-analysis of inconsistent effects: a time for change. *Annals of Internal Medicine* **160**: 267–270. DOI:10.7326/M13-2886.

Deeks JJ. 2002. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine* **21**: 1575–1600. DOI:10.1002/sim.1188.

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

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

Duchesne P, Lafaye De Micheaux P. 2010. Computing the distribution of quadratic forms: further comparisons between the Liu–Tang–Zhang approximation and exact methods. *Computational Statistics and Data Analysis* **54**: 858–862. DOI:10.1016/j.csda.2009.11.025.

Engels EA, Schmid CH, Terrin N, Olkin I, Lau J. 2000. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. *Statistics in Medicine* **19**: 1707–1728.

Farebrother RW. 1984. Algorithm AS 204: the distribution of a positive linear combination of $\chi2$ random variables. *Journal of the Royal Statistical Society: Series C: Applied Statistics* **33**: 332–339. DOI:10.2307/2347721.

Friedman L. 2000. Estimators of random effects variance components in meta-analysis. *Journal of Educational and Behavioral Statistics* **25**: 1–12. DOI:10.3102/10769986025001001.

Friedrich JO, Adhikari NKJ, Beyene J. 2011. Ratio of means for analyzing continuous outcomes in meta-analysis performed as well as mean difference methods. *Journal of Clinical Epidemiology* **64**: 556–564. DOI:10.1016/j.jclinepi.2010.09.016.

Gardiner JC, Luo Z, Roman LA. 2009. Fixed effects, random effects and GEE: what are the differences? *Statistics in Medicine* **28**: 221–239. DOI:10.1002/sim.3478.

Gelman A. 2006. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis* **1**: 515–534. DOI:10.1214/06-BA117A.

Goldstein H. 1995. *Multilevel Statistical Models* (2nd edn). Edward Arnold: New York, Halstead Press, London.

Goldstein H, Rasbash J. 1996. Improved approximations for multilevel models with binary responses. *J. R. Stat. Soc. Ser. A Stat. Soc.* **159**: 505. DOI:10.2307/2983328.

Gutierrez RG, Carter S, Drukker DM. 2001. sg160: On boundary-value likelihood-ratio tests. Stata Technical Bulletin 60: 15–18. *Reprinted in Stata Technical Bulletin Reprints* (pp. 269–273). College Station, TX: Stata.

Harbord RM, Higgins JPT. 2008. Meta-regression in Stata. *Stata Journal* **8**: 493–519.

Hardy RJ, Thompson SG. 1996. A likelihood approach to meta-analysis with random effects. *Statistics in Medicine* **15**: 619–629. DOI:10.1002/(SICI)1097-0258(19960330)15:6<619::AID-SIM188>3.0.CO;2-A.

Harris R, Bradburn M, Deeks J, Harbord R, Altman D, Sterne J. 2008. metan: fixed- and random-effects meta-analysis. *Stata Journal* **8**: 3–28.

Hartung J. 1999. An alternative method for meta-analysis. *Biometrical Journal* **41**: 901–916. DOI:10.1002/(SICI)1521-4036(199912)41:8<901::AID-BIMJ901>3.0.CO;2-W.

Hartung J, Knapp G. 2005. On confidence intervals for the among-group variance in the one-way random effects model with unequal error variances. *J. Stat. Plan. Inference* **127**: 157–177. DOI:10.1016/j.jspi.2003.09.032.

Hartung J, Makambi KH. 2003. Reducing the number of unjustified significant results in meta-analysis. *Commun. Stat.—Simul. Comput.* **32**: 1179–1190. DOI:10.1081/SAC-120023884.

Hartung J, Makambi KH. 2002. Positive estimation of the between-study variance in meta-analysis. *S. Afr. Stat. J.* **36**: 55–76.

Hedges LV. 1983. A random effects model for effect sizes. *Psychological Bulletin* **93**: 388–395. DOI:10.1037/0033-2909.93.2.388.

Hedges LV, Olkin I. 1985. *Statistical Methods for Meta-Analysis*. Orlando: Academic Press.

Higgins JPT, Thompson SG. 2002. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* **21**: 1539–1558. DOI:10.1002/sim.1186.

Hunter JE, Schmidt FL. 2004. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. Thousand Oaks, Calif: Sage.

IBM Corp 2013. *IBM SPSS for Windows*. Armonk, NY: IBM Corp.

Ioannidis JPA, Patsopoulos NA, Evangelou E. 2007. Uncertainty in heterogeneity estimates in meta-analyses. *BMJ* **335**: 914–916. DOI:10.1136/bmj.39343.408449.80.

Jackson D. 2013. Confidence intervals for the between-study variance in random effects meta-analysis using generalised Cochran heterogeneity statistics. *Res. Synth. Methods* **4**: 220–229. DOI:10.1002/jrsm.1081.

Jackson D, Bowden J, Baker R. 2010. How does the DerSimonian and Laird procedure for random effects meta-analysis compare with its more efficient but harder to compute counterparts? *J. Stat. Plan. Inference* **140**: 961–970. DOI:10.1016/j.jspi.2009.09.017.

Knapp G, Biggerstaff BJ, Hartung J. 2006. Assessing the amount of heterogeneity in random-effects meta-analysis. *Biometrical Journal* **48**: 271–285. DOI:10.1002/bimj.200510175.

Knapp G, Hartung J. 2003. Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine* **22**: 2693–2710. DOI:10.1002/sim.1482.

Kontopantelis E, Reeves D. 2012. Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: a simulation study. *Statistical Methods in Medical Research* **21**: 409–426. DOI:10.1177/0962280210392008.

Kontopantelis E, Reeves D. 2010. metaan: random-effects meta-analysis. *Stata Journal* **10**: 395–407.

Kontopantelis E, Reeves D. 2009. MetaEasy: a meta-analysis add-in for Microsoft Excel. *Journal of Statistical Software* **30**(7): 1–25.

Kontopantelis E, Springate DA, Reeves D. 2013. A re-analysis of the Cochrane Library data: the dangers of unobserved heterogeneity in meta-analyses. *PLoS One* **8**, e69930. doi: 10.1371/journal.pone.0069930

Kulinskaya E, Dollinger MB, Bjørkestøl K. 2011a. Testing for homogeneity in meta-analysis I. The one-parameter case: standardized mean difference. *Biometrics* **67**: 203–212. DOI:10.1111/j.1541-0420.2010.01442.x.

Kulinskaya E, Dollinger MB, Bjørkestøl K. 2011b. On the moments of Cochran's Q statistic under the null hypothesis, with application to the meta-analysis of risk difference. *Res. Synth. Methods* **2**: 254–270. DOI:10.1002/jrsm.54.

Kuss O. 2014. Statistical methods for meta-analyses including information from studies without any events-add nothing to nothing and succeed nevertheless. *Statistics in Medicine* . DOI:10.1002/sim.6383.

Laird NM, Mosteller F. 1990. Some statistical methods for combining experimental results. *International Journal of Technology Assessment in Health Care* **6**: 5–30.

Lambert PC, Sutton AJ, Burton PR, Abrams KR, Jones DR. 2005. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine* **24**: 2401–2428. DOI:10.1002/sim.2112.

Lunn DJ, Thomas A, Best N, Spiegelhalter D. 2000. WinBUGS—a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* **10**: 325–337.

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

Morris CN. 1983. Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association* **78**: 47. DOI:10.2307/2287098.

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

Panityakul T, Bumrungsup C, Knapp G. 2013. On estimating residual heterogeneity in random-effects meta-regression: a comparative study. *J. Stat. Theory Appl.* **12**: 253. DOI:10.2991/jsta.2013.12.3.4.

Paule RC, Mandel J. 1982. Consensus values and weighting factors. National Institute of Standards and Technology.

Preuß M, Ziegler A. 2014. A simplification and implementation of random-effects meta-analyses based on the exact distribution of Cochran's Q. *Methods of Information in Medicine* **53**: 54–61. DOI:10.3414/ME13-01-0073.

Pullenayegum EM. 2011. An informed reference prior for between-study heterogeneity in meta-analyses of binary outcomes. *Statistics in Medicine* **30**: 3082–3094. DOI:10.1002/sim.4326.

Rabe-Hesketh S, Skrondal A, Pickles A. 2003. Maximum likelihood estimation of generalized linear models with covariate measurement error. *Stata Journal* **3**: 386–411.

Rasbash J, Charlton C, Browne WJ, Healy M, Cameron B. 2014. MLwiN. Centre for Multilevel Modelling, University of Bristol.

Raudenbush SW. 2009. Analyzing effect sizes: random-effects models. In Cooper H, Hedges LV, Valentine JC (eds.). *The Handbook of Research Synthesis and Meta-Analysis* (pp. 295–315). New York: Russell Sage Foundation.

Raudenbush SW, Bryk A, Congdon R. 2004. *HLM 6 for Windows*. Skokie, IL: Scientific Software International, Inc..

R Development Core Team 2008. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rhodes KM, Turner RM, Higgins JPT. 2015. Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continuous outcome data. *Journal of Clinical Epidemiology* **68**: 52–60. DOI:10.1016/j.jclinepi.2014.08.012.

Rosenberg MS, Adams DC, Gurevitch J. 2000. MetaWin: statistical software for meta-analysis. Sinauer Associates.

Rukhin AL. 2013. Estimating heterogeneity variance in meta-analysis. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **75**: 451–469. DOI:10.1111/j.1467-9868.2012.01047.x.

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

SAS Institute Inc. 2003. *SAS Software*. USA: Cary, NC.

StataCorp 2013. *Stata Statistical Software*. College Station, TX: StataCorp LP.

Senn S. 2007. Trying to be precise about vagueness. *Statistics in Medicine* **26**: 1417–1430. DOI:10.1002/sim.2639.

Sidik K, Jonkman JN. 2007. A comparison of heterogeneity variance estimators in combining results of studies. *Statistics in Medicine* **26**: 1964–1981. DOI:10.1002/sim.2688.

Sidik K, Jonkman JN. 2005a. A note on variance estimation in random effects meta-regression. *Journal of Biopharmaceutical Statistics* **15**: 823–838. DOI:10.1081/BIP-200067915.

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

Smith TC, Spiegelhalter DJ, Thomas A. 1995. Bayesian approaches to random-effects meta-analysis: a comparative study. *Statistics in Medicine* **14**: 2685–2699.

Stern SE, Welsh AH. 2000. Likelihood inference for small variance components. *The Canadian Journal of Statistics* **28**: 517–532. DOI:10.2307/3315962.

Stijnen T, Hamza TH, Özdemir P. 2010. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. *Statistics in Medicine* **29**: 3046–3067. DOI:10.1002/sim.4040.

Sweeting MJ, Sutton AJ, Lambert PC. 2004. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in Medicine* **23**: 1351–1375. DOI:10.1002/sim.1761.

Switzer FS, Paese PW, Drasgow F. 1992. Bootstrap estimates of standard errors in validity generalization. *Journal of Applied Psychology* **77**: 123–129. DOI:10.1037/0021-9010.77.2.123.

The Nordic Cochrane Centre. 2014. Review Manager (RevMan). The Cochrane Collaboration.

Thomas A. 1994. BUGS: a statistical modelling package. *RTA/BCS Modular Languages Newsletter* **2**: 36–38.

Thomas N. 2010. OpenBUGS website. Overview. Available at: http://www.openbugs.net/w/Overview [Accessed July 2015]

Thompson SG, Sharp SJ. 1999. Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in Medicine* **18**: 2693–2708.

Thorlund K, Wetterslev J, Awad T, Thabane L, Gluud C. 2011. Comparison of statistical inferences from the DerSimonian–Laird and alternative random-effects model meta-analyses—an empirical assessment of 920 Cochrane primary outcome meta-analyses. *Res. Synth. Methods* **2**: 238–253. DOI:10.1002/jrsm.53.

Tian L. 2008. Inferences about the between-study variance in meta-analysis with normally distributed outcomes. *Biom. J. Biom. Z.* **50**: 248–256. DOI:10.1002/bimj.200710408.

Turner RM, Davey J, Clarke MJ, Thompson SG, Higgins JP. 2012. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane Database of Systematic Reviews. *International Journal of Epidemiology* **41**: 818–827. DOI:10.1093/ije/dys041.

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

Viechtbauer W. 2013. Metafor: meta-analysis package for R.

Viechtbauer W. 2007. Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine* **26**: 37–52. DOI:10.1002/sim.2514.

Viechtbauer W. 2005. Bias and efficiency of meta-analytic variance estimators in the random-effects *Model*. *Journal of Educational and Behavioral Statistics* **30**: 261–293. DOI:10.3102/10769986030003261.

Villar J, Mackey ME, Carroli G, Donner A. 2001. Meta-analyses in systematic reviews of randomized controlled trials in perinatal medicine: comparison of fixed and random effects models. *Statistics in Medicine* **20**: 3635–3647.

Wallace BC, Dahabreh IJ, Trikalinos TA, Lau J, Trow P, Schmid CH. 2012. Closing the gap between methodologists and end-users: R as a computational back-end. *Journal of Statistical Software* **49**: 1–15.

White IR. 2009. Multivariate random-effects meta-analysis. *Stata Journal* **9**: 40–56.

Swallow WH, Monahan JF. 1984. Monte Carlo comparison of ANOVA, MIVQUE, REML, and ML estimators of variance components. *Technometrics* **26**: 47–57. DOI:10.1080/00401706.1984.10487921.

Zamora J, Muriel A, Abraira V. 2006. Meta-DiSc statistical methods.

Zhang Z, McArdle JJ, Wang L, Hamagami F. 2008. A SAS interface for Bayesian analysis with WinBUGS. *Struct. Equ. Model. Multidiscip. J.* **15**: 705–728. DOI:10.1080/10705510802339106.

Supporting information

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