Methods to calculate uncertainty in the estimated overall effect size from a random-effects meta-analysis

Areti Angeliki Veroniki[1,2*]; Email: averonik@cc.uoi.gr

Dan Jackson[3]; Email: dan.jackson@mrc-bsu.cam.ac.uk

Ralf Bender[4]; Email: ralf.bender@iqwig.de

Oliver Kuss[5,6]; Email: oliver.kuss@ddz.uni-duesseldorf.de

Dean Langan[7]; Email: d.langan@ucl.ac.uk

Julian PT Higgins[8]; Email: julian.higgins@bristol.ac.uk

Guido Knapp[9]; Email: guido.knapp@tu-dortmund.de

Georgia Salanti[10]; Email: georgia.salanti@ispm.unibe.ch

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

[2] Department of Primary Education, School of Education, University of Ioannina, Ioannina, Greece

[3] MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 0SR, U.K

[4] Department of Medical Biometry, Institute for Quality and Efficiency in Health Care (IQWiG), Im Mediapark 8, 50670 Cologne, Germany

[5] Institute for Biometrics and Epidemiology, German Diabetes Center, Leibniz Institute for Diabetes Research at Heinrich Heine University, 40225 Düsseldorf, Germany

[6] Institute of Medical Statistics, Heinrich-Heine-University, Medical Faculty, Düsseldorf, Germany

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/jrsm.1319

[7] Institute of Child Health, UCL, London, WC1E 6BT, UK

[8] Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, U.K.

[9] Department of Statistics, TU Dortmund University, 44221 Dortmund, Germany

[10] Institute of Social and Preventive Medicine, University of Bern, Finkenhubelweg 11, 3012 Bern, Switzerland

*Corresponding Author:

Areti Angeliki Veroniki, PhD, MSc

Department of Primary Education, School of Education,

University of Ioannina, Ioannina, Greece

Telephone: 0030 26510 05712

Email: averonik@cc.uoi.gr

Fax: 0030 26510 05854

Keywords: meta-analysis, random effects, overall treatment effect, confidence interval, evidence synthesis

Word Count: 155 (abstract: 250 max), 9,696 (main text), 1 table, 1 supporting file

Abstract

Meta-analyses are an important tool within systematic reviews to estimate the overall effect size and its confidence interval for an outcome of interest. If heterogeneity between the results of the relevant studies is anticipated, then a random-effects model is often preferred for analysis. In this model, a prediction interval for the true effect in a new study also provides additional useful information. However, the DerSimonian and Laird method – frequently used as the default method for meta-analyses with random effects – has been long challenged due to its unfavourable statistical properties. Several alternative methods have been proposed that may have better statistical properties in specific scenarios. In this paper, we aim to provide a comprehensive overview of available methods for calculating point estimates, confidence intervals and prediction intervals for the overall effect size under the random-effects model. We indicate whether some methods are preferable than others by considering the results of comparative simulation and real-life data studies.

1 Introduction

Systematic reviews and meta-analyses provide a method for collecting and synthesizing research and are often used to inform decision making. The number of these publications has increased substantially since the 1990s.[1] Meta-analysis is a valuable technique to summarize study-specific results and often reduces bias and uncertainty from individual studies. Guidelines and Health Technology Assessment panels, as well as international organizations, including the World Health Organization, [2] the European Medicines Agency, [3] and governmental agencies worldwide, such as, the Canadian Agency for Drugs and Technologies in Health,[4] the Institute for Quality and Efficiency in Health Care (IQWiG),[5] and the National Institute for Health and Clinical Excellence,[6] recognize the need to ensure that the best available evidence informs clinical practice and health care decision making. This typically involves conducting a high-quality knowledge synthesis and meta-analysis. Quantitative results of meta-analyses of relevant studies, in the form of a point estimate and a confidence interval (CI) for the effect size parameter of interest, are invariably considered together with judgements about the quality of the evidence to produce recommendations for practice.[7]

Quantification of uncertainty in the estimated overall effect size is important in the process of drawing conclusions from a meta-analysis. This uncertainty should ideally account for between-study heterogeneity in the intervention effects across study settings and populations.[8,9] For this reason, the random-effects model is often employed, which includes a between-study variance parameter. The uncertainty of the estimate of the overall effect size can be described by the corresponding CI under the random-effects model, and its width depends on the magnitude of the between-study variance, the number of studies, the precision of the study-specific effect sizes, and the significance level. [10]

The estimation of the CI for overall effect size is often conducted with the Wald-type method using a normal distribution, with variance equal to the inverse of the sum of the study weights, and the DerSimonian and Laird [11] estimator for the between-study variance, and this has been used routinely in many meta-analyses.[12] However, numerous shortcomings of this approach have been raised, such that the CI for the overall effect size generally does not retain its coverage probability (i.e., the proportion of times that the interval includes the true value) and hence it underestimates the statistical error, producing overconfident results.[12-18] This is mainly because the Wald-type CI is based on a large-sample approximation (in terms of the number of studies) and the number of studies is usually small. Typically, the number of studies synthesised in a meta-analysis in medical research is less than 20 [19-23] suggesting that any large-sample approximation is likely to be inaccurate. Several attempts to improve the standard Wald-type CI approach have been suggested, each of which has different statistical properties.

Another important aim in decision-making is the prediction of the true effect size in an individual (future) study and setting. Higgins et al. [24] suggested the use of prediction intervals under the random-effects model for this purpose. The use of prediction intervals has been promoted and although they have not often been employed in practice they provide useful additional information. [25,26]

In this paper, we aim to provide a comprehensive overview of available methods in the methodological literature for calculating a CI for the overall effect size under the random-effects model, and to indicate whether some methods are preferable to others by considering the results of comparative simulation and real-life data studies. We also examine potential issues surrounding the computation of prediction intervals under the random-effects model.

The article is structured as follows. In section 2 we present the conventional meta-analytic models and set up our notation (section 2.1) and describe our review methodology

(section 2.2). In section 3 we describe the statistical methods found in our literature review. In section 3.1 we describe 15 identified methods to calculate a CI for the overall effect size. In section 3.2 we discuss the comparative performances of different methods for computing a CI for the overall effect size, as described in previous studies, and summarise recommendations made by their respective authors. In section 3.3 we discuss methods for computing prediction intervals. We conclude with a discussion of all intervals (confidence and prediction intervals) in section 4.

2 Methods

2.1 Meta-analysis models and notation

The conventional fixed-effect and random-effects models are the two main meta-analysis models to synthesise the study results. [27] The random-effects model accounts for two sources of variation, quantified by the within-study variance (v_i) and the between-study variance (τ^2). When $\hat{\tau}^2 = 0$, the fitted random-effects model collapses to the fixed-effect model (also known as common-effect model),[28,29] and therefore the random-effects model can be considered a generalization of the fixed-effect model (i.e., that the fixed-effect model is a special case of the random-effects model). CIs under the fixed-effect model can have poor properties even for low but non-zero heterogeneity. [13,30]

Both the conventional fixed and random-effects models require an estimated effect size y_i (such as log-odds ratio) and an estimated (within-study) variance v_i ($var(y_i) = v_i$) from every included study i, $i = 1, \ldots, k$. The choice between the two models has been widely discussed in the literature [31-33] and summarized in the Cochrane Handbook. [9] In this paper, we focus on the random-effects meta-analysis model using inverse-variance weighting. Other techniques to combine study information to calculate the overall effect size are also available, such as weighting by sample size [34,35] or using confidence distributions. [36,37] Also, dichotomous outcomes do not require inverse-variance methods, as they can be modelled

directly using one-step models (e.g., -generalised linear mixed models). [38] Alternative methods accounting for heterogeneity by the use of a multiplicative parameter, where study weights are independent of observed heterogeneity, are also available.[10,39,40] The description of these methods is beyond the scope of this review.

The conventional random-effects model assumes that the estimated effect size from the i^{th} study y_i is

$$y_i = \theta_i + \varepsilon_i,$$

where the study-specific random error (ε_i), and the underlying true effect sizes in the individual studies (θ_i) are normally distributed as

$$\varepsilon_i \sim N(0, v_i),$$

$$\theta_i \sim N(\mu, \tau^2).$$

The random-effects estimated overall effect size $\hat{\mu}_{RE}$ and its variance can be estimated as

$$\hat{\mu}_{RE} = \frac{\sum y_i w_{i,RE}}{\sum w_{i,RE}} \quad \text{and} \quad var(\hat{\mu}_{RE}) = \frac{1}{\sum w_{i,RE}},$$

with weights $w_{i,RE} = 1/(v_i + \hat{\tau}^2)$, where it can be seen that these weights are the inverse of the estimated total study variances. Similarly, the fixed-effect weights can be calculated as $w_{i,FE} = 1/v_i$, and the estimated overall effect size $\hat{\mu}_{FE}$ and its variance are given by

$$\hat{\mu}_{FE} = \frac{\sum y_i w_{i,FE}}{\sum w_{i,FE}} \quad \text{and} \quad var(\hat{\mu}_{FE}) = \frac{1}{\sum w_{i,FE}}.$$

The uncertainty in an estimated effect size for a given study, in relation to its study-specific true effect size, is expressed via the within-study variance v_i. The standard approach described above assumes that the estimated within-study variances v_i are fixed and known although they have to be estimated from the data. This assumption is justifiable when each study size is sufficiently large. The v_i estimation is not only sensitive to the study size, but also to the data type and effect size used. For example, the v_i estimator for continuous outcomes when using the standardised mean difference depends on the estimated effect size. Hence, although v_i are assumed fixed and known, v_i are in fact estimated with some

uncertainty. Several authors point out that this assumption could affect the estimation of the overall effect size, its variance, and related inferences. [13,41-45] Therefore, calculating study weights through within-study variances and assuming that they are known constants may have less desirable properties. This issue has previously been discussed and some attempts have been made to account for the uncertainty in the weights. [46-49] Hence, among other factors, the performance of the CI methods depends on how well we estimate the study weights.

Similarly, the estimation of τ^2 is performed with some uncertainty, and this uncertainty depends on the size and number of studies in the meta-analysis, as well as the size of the between-study variability. Factors such as these have implications for the accuracy of the standard statistical methods described in this paper, which has motivated many of the attempts to improve this. There are also many methods to estimate τ^2, any of which can be used in some of the CI methods described below, and we refer to a previous publication on this topic. [50] Also, for a review of the simulation studies evaluating the comparative performance of τ^2 estimators we direct the reader elsewhere. [51]

2.2 Review methods

We searched PubMed from inception until 29 April 2016 to identify full text research articles that describe or compare methods for calculating CI for μ in simulations or in real data sets. We scanned the references of the selected articles for additional relevant articles, and we conducted general internet searches using the web search engine Google. We also used our networks of professional collaborations to identify potentially relevant articles. We included all studies that report the development or comparison of methods to calculate a CI for the overall effect size under the random-effects model. We also included studies reporting on prediction interval methods identified from our internet searches and networks of collaborations. We excluded commentaries, abstracts, and studies written in languages other

than English, and studies relating to the hypothesis tests for the overall effect size. We restricted our investigation to CI methods developed under the random-effects meta-analysis model that assume the true study-specific effects are normally distributed, while we excluded CI methods developed for network meta-analysis, one-stage individual patient data meta-analysis, meta-analysis of diagnostic test accuracy studies, meta-analysis of multiple outcomes, and meta-regression analysis. One reviewer (AAV) summarised methods and studies' conclusions from each included article and recorded any conclusions from comparative articles (studies that compare at least two methods). The information extracted refers to the performance of the various methods and the judgements deducted about their related advantages, and this information was checked by all co-authors. Disagreements were resolved by discussion. The PubMed search strategy is included in Supporting File: Appendix 1.

We describe known properties of the methods in terms of coverage probability and CI width in section 3.2. The closer the coverage probability is to the nominal level (usually 0.95) the better the CI is considered to be. A CI is exact when the actual coverage equals the nominal coverage. The coverage probability is closely related to the type I error of the hypothesis test on the overall effect size: assuming the null hypothesis is true, one minus the type I error rate is the coverage probability. A further criterion for comparing methods is that methods that provide narrower CIs, whilst retaining the correct coverage probability, are preferable because they increase precision, and hence are more informative. All statistics presented in this paper refer to two-tailed tests.

3 Results

The database search returned 5628 matches in PubMed and 20 records identified through other sources and searching reference lists. In total, 69 publications met the eligibility criteria, which are listed in Supporting File; Appendix 2. We identified 15 methods to compute a CI

for the overall effect size. The properties of those methods have been evaluated in 31 research papers, including 30 simulation studies and 32 real-life data evaluations of two or more methods. Below we present the 15 identified approaches in 7 broad categories, and as a separate section we present the comparative results of the identified simulations and studies using real data sets (see Supporting File; Appendices 3-4 for simulation scenarios and study characteristics and Supporting File; Appendix 5 for a summary of performance measures in simulation studies). In Table 1 we summarize the methods available in several software options. [52]

3.1 Confidence Intervals for the overall effect size

3.1.1 Wald-type (WT) methods

i) *Wald-type normal distribution (WTz) confidence intervals (method 1)*

The WTz approach is the most popular technique for calculating a CI for μ, [11] and a 95% CI is given by

$$\hat{\mu}_{RE} \pm z_{0.975}\sqrt{var(\hat{\mu}_{RE})},$$

where $z_{0.975}$ is the 0.975 quantile of the standard normal distribution. Any τ^2 estimator can be used when computing a WTz CI.[50,53,54] This method often has coverage probability considerably below nominal 0.95 level [14,15,45,55-61] when k is small and/or τ^2 is large.[13,45,58,59,62-67] Brockwell and Gordon [13] stated that the greatest source of error in the method is the use of a normal approximation for $\hat{\mu}_{RE}$. Despite the widespread use of the WTz method, it ignores uncertainty of the estimates of τ^2 and v_i in $var(\hat{\mu}_{RE})$.

ii) Wald-type t-distribution (WTt) confidence intervals (method 2)

A slight modification of the WTz CI is the WTt approach, where the *t*-distribution with $k-1$ degrees of freedom is used, as opposed to the normal distribution. Although the two distributions converge asymptotically, the *t*-quantile is larger than the z-quantile associated with the WTz method. Hence, the WTt approach results in a wider CI and was proposed in order to increase the coverage probability, especially when the number of studies is small.[57,68] A 95% CI can be obtained by

$$\hat{\mu}_{RE} \pm t_{k-1,0.975}\sqrt{var(\hat{\mu}_{RE})},$$

where $t_{k-1,0.975}$ is the 0.975 quantile of the *t*-distribution with $k-1$ degrees of freedom. Any τ^2 estimator can be used to compute a WTt CI.[50,53]

iii) Quantile approximation (WTqa) confidence intervals (method 3)

Brockwell and Gordon [62] proposed the WTqa method as an alternative to WTz method in an attempt to achieve better coverage. The method resembles WTz and WTt, but instead of using normal or *t* distributions it approximates the 0.025 and 0.975 quantiles of the distribution of the statistic

$$M = \frac{\hat{\mu}_{RE} - \mu}{\sqrt{var(\hat{\mu}_{RE})}},$$

that are required for the 95% CI for μ. Hence, the WTqa uses different quantiles than the ones used in the WTz and WTt approaches. Let b_k be the quantile approximation function, which monotonically decreases as a non-linear function of k, then a 95% CI is calculated as

$$\hat{\mu}_{RE} \pm b_k \sqrt{var(\hat{\mu}_{RE})}.$$

The quantiles b_k are estimated via a Monte Carlo simulation process of samples of the M statistic with b_k equal to the average of 0.025 and 0.975 absolute quantiles of the distribution, thus accounting for any small asymmetry in the distribution of M around zero.[62] To obtain the function b_k, Brockwell and Gordon fit a regression equation for the quantiles as a function of k. The resulting regression equation (for $k = 1,2, \ldots ,30$) is:

$$b_k = 2.061 + \frac{4.902}{k} + \frac{0.756}{\sqrt{k}} + \frac{0.958}{\ln(k)}.$$

However, both number of studies *k* and the magnitude of τ^2 may impact on the performance of the WTqa method, [62] and changes in the distribution of the within-study variances can importantly impact on b_k. [69] Although WTqa approach has been criticized on the grounds that it is, at best, very difficult to obtain suitable critical values b_k that apply to all meta-analyses, [69] we include it in this paper for completeness. As a conservative approach, Jackson and Bowden [69] suggested the use of the standard normal quantile instead, and to assess the robustness of the findings via a sensitivity analysis of alternative quantiles. Brockwell and Gordon [62] developed the WTqa method using the DerSimonian and Laird[11] estimator of τ^2,

but WTqa could, in principle, be implemented for any alternative τ^2 estimator. However, it is not advised to develop the WTqa CI further. [69]

3.1.2 Hartung-Knapp/Sidik-Jonkman (HKSJ) confidence intervals (methods 4 and 5)

Hartung and Knapp [14] and Sidik and Jonkman [15] independently introduced the HKSJ CI (method 4) to handle meta-analyses that include a small number of studies. This method is based on the S statistic, which follows a *t*-distribution with $k-1$ degrees of freedom,

$$S = \frac{\hat{\mu}_{RE} - \mu}{\sqrt{\sigma^2_{w,\hat{\mu}_{RE}}}} \sim t_{k-1},$$

with

$$\sigma^2_{w,\hat{\mu}_{RE}} = q \cdot \frac{1}{\sum w_{i,RE}} = q \cdot var(\hat{\mu}_{RE}),$$

and $q = \frac{Q_{gen}}{k-1}$, where Q_{gen} is the generalized Q-statistic

$$Q_{gen} = \sum w_{i,RE}(y_i - \hat{\mu}_{RE})^2.$$

A 95% CI for μ is given by

$$\hat{\mu}_{RE} \pm t_{k-1,0.975}\sqrt{\sigma^2_{w,\hat{\mu}_{RE}}}.$$

Although the method does not take into account the uncertainty in τ^2, the use of a different statistical approximation to the usual Wald-type CI may improve accuracy. Also, the HKSJ method can be applied with any τ^2 estimator, and is exact for known variance components. [14] For meta-analysis software where this method is not available yet, IntHout et al. [16] suggested an approach to convert WTz CIs easily to HKSJ CIs. The extension to meta-regression was investigated by Knapp and Hartung, [67] and a generalization of the method to multivariate meta-analysis was explored by Jackson and Riley. [70]

When all variance components, including the between-study variances, are fixed and known, the expected value of Q_{gen} is $k-1$, which equals the degrees of freedom of the associated χ^2 distribution.[71-73] Hence, the small-sample adjustment q will tend to be close to 1. However, q may in fact turn out to be much smaller than 1, such as in cases where the

effect sizes are very homogeneous or when the number and/or size of studies is small. This leads to a narrower CI than the WTt approach and can also lead to a narrower CI compared to the WTz method. [74,75] Although the *t*-quantile is always larger than the z-quantile associated with the WTz method, explaining in part why the HKSJ CI performs better than the WTz CI, in the case of $\sqrt{q} < z_{0.975}/t_{k-1;0.975}$ the HKSJ CI will be narrower than the WTz CI.

Wiksten et al. [74] show that if $\hat{\tau}^2 = 0$ then we estimate $var(\hat{\mu}_{RE}) = var(\hat{\mu}_{FE})$, and further that the variance of the estimate of μ simplifies to $\sigma^2_{w,\hat{\mu}_{RE}} = (Q/(k-1)) \cdot var(\hat{\mu}_{FE})$, with $Q = Q_{gen}\big|_{(\hat{\tau}^2=0)}$ and $Q \leq (k-1)$ when the DerSimonian and Laird[11] estimator of τ^2 is used. Therefore, the variance of the estimated effect size is always smaller or equal for the HKSJ method than the WTz method when this estimator of the between-study variance is zero. The possibility that the variance of the estimated effect size from the HKSJ method can be smaller than the variance of the WTz method was discussed by Knapp and Hartung,[67] who proposed a simple modification to the procedure. The authors suggested using q^* instead of q (method 5) [67]

$$q^* = \max\{1, q\},$$

to ensure more conservative results. However, this practice may be overly conservative, leading to loss of power.[76,77]

Sidik and Jonkman [15] recommend using the HKSJ CI, but instead of $\sigma^2_{w,\hat{\mu}_{RE}}$ they suggest applying the sandwich variance estimator:

$$\sigma^2_{SJ,\hat{\mu}_{RE}} = \frac{\sum w^2_{i,RE}(y_i - \hat{\mu}_{RE})^2}{\left(\sum w^2_{i,RE}\right)^2}.$$

This is a robust estimator of $var(\hat{\mu}_{RE})$, where the inverse of the study weights ($w^{-1}_{i,RE} = v_i + \hat{\tau}^2$) are estimated through the squared sample residuals ($(y_i - \hat{\mu}_{RE})^2$) from the data, rather than assuming $v_i + \hat{\tau}^2$.

However, Sidik and Jonkman [60] state that $\sigma^2_{SJ,\hat{\mu}_{RE}}$ is biased when k is small, and hence they suggest a bias corrected estimator of $var(\hat{\mu}_{RE})$ (see Sidik and Jonkman [60] for details). An alternative approach based on the expected information and on appropriately modified degrees of freedom of the *t*-distribution was suggested by Kenward and Roger. [78] These alternative expressions for $var(\hat{\mu}_{RE})$ could also be used in Wald-type CIs but have not been adopted in practice so we do not explore their use further here.

3.1.3 Likelihood-based methods

i) *Profile likelihood (PL) confidence intervals (method 6)*

The PL method has been established in meta-analysis by Hardy and Thompson [43] and is based on the likelihood ratio statistic, which unlike the WTz approach allows for asymmetric intervals. For $y_i \sim N(\mu, v_i + \tau^2)$, the log-likelihood function of the parameter vector (μ, τ^2) is given by

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

Maximum likelihood estimates of (μ, τ^2) can be found by maximizing $lnL(\mu, \tau^2)$ under the restriction $\tau^2 \geq 0$. The PL function is based on the log-likelihood function and uses an iterative process that provides CIs for μ that allow for the fact that τ^2 needs to be estimated as well. Since the PL approach profiles over τ^2, it accounts for, but does not fully allow for, the uncertainty in τ^2. This is because asymptotic results are required when using this method. However, the PL method is anticipated to be more accurate than the Wald-type methods in smaller samples.

The profile log-likelihood for μ is defined as

$$lnL_p(\mu) = lnL(\mu, \hat{\tau}^2_{ML}(\mu)),$$

where $\hat{\tau}^2_{ML}(\mu)$ is the maximum likelihood estimator for τ^2as μ varies. [43] A 95% CI for μ can be obtained as the values which satisfy (see Hardy and Thompson [43] equation 11):

$$lnL_p(\mu) > lnL_p(\hat{\mu}_{RE}) - \frac{\chi^2_{1,0.05}}{2},$$

where $\chi^2_{1,0.05}$ is the 0.05 quantile of the χ^2-distribution with 1 degree of freedom. It has been shown that for small k and τ^2, iterative algorithms are less likely to converge to a single value. [13]

ii) Higher-order likelihood inference methods (methods 7 and 8)

As Reid [79] explains, the main asymptotic properties of likelihood-based inference include: (a) consistent, asymptotically normal and efficient maximum likelihood estimators; (b) an asymptotically normally distributed score statistic with mean zero; and (c) an asymptotic chi-squared distributed likelihood ratio statistic. For example, the PL CI in the previous section relies upon the third of these standard results. As Reid [79] also explains, higher-order asymptotic results for likelihood based inference are also available. Some higher-order likelihood inference methods have recently been applied to meta-analysis, which is a situation where they may be thought to be especially valuable. This is because the number of studies is often small, so that the commonly used 'lower-order' asymptotic approximations to the likelihood function will be inadequate. Higher-order likelihood based methods therefore have the potential to produce more accurate results in meta-analysis and several proposals for this have been made. We briefly summarize the methods here but the details are technical and so we refer the reader to the articles cited below for more information.

The Bartlett-type correction of the likelihood ratio statistic was first introduced by Bartlett (method 7). [80] Noma [17] explains how to apply this to random-effects meta-analysis, and so use a higher-order approximation than the PL method above. Noma [17] also explains how to use the score statistic to compute CIs, and subsequently derives a higher-order Bartlett type adjustment to this score. Skovgaard proposed an alternative higher-order approximation to the profile log-likelihood (method 8) [81] and Guolo [65] explains how to apply this to random-

effects meta-analysis. For details on the method we direct the reader elsewhere.[65,82] The higher-order asymptotic methods have higher degree of accuracy, but in some cases (e.g., when the between-study variance is close to zero) they may produce numerically unstable maximum likelihood estimates due to the discontinuity of the statistic.[65,82-84] In such cases, a bias reduction approach is suggested.[85] Hence, the Bartlett-type correction (method 7) and the Skovgaard statistic (method 8) are the two main proposals for higher-order approximations when using methods based on the PL.[65]

3.1.4 Henmi and Copas (HC) confidence intervals (method 9)

Henmi and Copas [30] propose an alternative strategy for obtaining intervals for μ that are less sensitive to publication bias than the widely used WTz method. Since the fixed-effect estimates assign larger weight to bigger studies, and study size is one component among others that is associated with the overall effect size in the presence of publication bias, this method centres the CI on a fixed-effect estimate. This is because the fixed-effect estimates are less sensitive to publication bias than the random-effects estimates. To allow for heterogeneity, they first estimate the variance of the fixed-effect estimate under the random-effects model as

$$\hat{V} = \frac{\hat{\tau}^2 \sum w_{i,FE}^2 + \sum w_{i,FE}}{\left(\sum w_{i,FE}\right)^2}.$$

Henmi and Copas [30] then derive an approximation to the resulting pivot G that is used for making inferences about μ

$$G = \frac{\hat{\mu}_{FE} - \mu}{\sqrt{\hat{V}}},$$

assuming that the DerSimonian and Laird [11] estimator of the between-study variance is used. Hence approximate CIs can be computed. This can be thought of as a hybrid approach, where the fixed-effect estimate is accompanied by a CI that allows for between-study heterogeneity

under the assumptions made in the random-effects model. A limitation of the approach is that the fixed-effect estimate is not fully efficient under the random-effects model, but Henmi and Copas [30] argue that it is "better to use a method that is more robust to publication bias, even if this means sacrificing some efficiency under the standard setting". A much simpler, but less accurate, way to implement Henmi and Copas' idea would be to assume that the pivot G approximately follows a standard normal distribution, but this would ignore all uncertainty in the between-study variance. This simpler approach could also be used with alternative estimators of the between-study variance. Alternatively, one could apply the IVher model suggested by Doi et al.[86] which uses quasi-likelihood approaches and is performed under the fixed-effect assumption. Doi et al.[86] show that the IVher model favours larger trials, retains the nominal coverage probability, and exhibits lower variance of the overall effect size as opposed to the random-effects model irrespective of the degree of the between-study heterogeneity.

3.1.5 Biggerstaff and Tweedie (BT) confidence intervals (method 10)

Biggerstaff and Tweedie [87] proposed the use of different study-specific weights to those more conventionally used in the random-effects model, and estimated μ along with its variance using the weight $w_{i,RE}^{BT}$, so as to acknowledge for τ^2 variability in the computation of CI for μ. Acknowledging the uncertainty in the weights allows greater uncertainty in the estimation of μ. [87] The $w_{i,RE}^{BT}$ weights are the expected value of the random-effects weights (calculated using the estimated τ^2) rather than the usual random-effects observed weights:

$$w_{i,RE}^{BT} = E\left(w_{i,RE}\right).$$

The $w_{i,RE}^{BT}$ weights depend on the density form of τ^2, and were derived using the DerSimonian and Laird[11] estimator for the between-study variance. Alternative estimators could also be used, in principle, when using this method, provided that their distribution, and

so the expected weights used by the method, can be evaluated. The variance of $\hat{\mu}_{RE}^{BT}$ is estimated as

$$var(\hat{\mu}_{RE}^{BT}) = \frac{1}{(\sum w_{i,RE}^{BT})^2} \sum (w_{i,RE}^{BT})^2 (v_i + \hat{\tau}^2).$$

Assuming normality, a 95% CI can be obtained as

$$\hat{\mu}_{RE}^{BT} \pm z_{0.975} \sqrt{var(\hat{\mu}_{RE}^{BT})},$$

Biggerstaff and Tweedie [87] use an approximate distribution to obtain the expected weights, but this has been improved upon by Preuß and Ziegler [88] who used the exact weights through the exact cumulative distribution function of Q, where $Q = Q_{gen}\big|_{(\hat{\tau}^2=0)}$.[89] Biggerstaff and Tweedie [87] provided the algorithm to implement the method in SAS.

3.1.6 Resampling methods

i) *Zeng and Lin (ZL) confidence intervals (method 11)*

Zeng and Lin [90] examine the distribution of the estimated overall effect size under the random-effects model and find that it is not asymptotically normally distributed for a finite number of studies *k*. This makes intuitive sense, because the textbook result that a linear combination of normal random variables is normally distributed requires that the coefficients in this linear combination are constants. When estimating the overall effect size however these coefficients are proportional to the weights and so are functions of the estimated between-study variance. We require a large number of studies in order to estimate this variance accurately enough to treat the weights as fixed constants.

Recognising that the estimated overall effect size is not asymptotically normally distributed for small *k*, Zeng and Lin [90] suggest a resampling procedure to obtain the distribution of this estimate, assuming that the DerSimonian and Laird[11] estimator of τ^2 is to be used. Briefly, they simulate values of τ^2 using the DerSimonian and Laird[11] estimating equation (where the individual study results used in this estimation are simulated from the

fitted random-effects model). They then simulate estimated average effect sizes using the sampled τ^2 to calculate the weights in the estimating equation for $\hat{\mu}_{RE}$ (as given in section 2.1, where the individual study results used in this estimation are simulated from the random-effects model centred at the estimated overall effect, and where the between study variance is taken to be the sampled value used to compute the weights). By repeating both aspects of this sampling process B times, B^2 estimates provide an empirical distribution of estimated overall effects that can be used to compute confidence intervals and make inferences.[90] This re-sampling procedure could be modified to accommodate alternative estimators of τ^2, by instead calculating alternative estimates at the first stage, but this idea would need to be critically evaluated before it could be accepted.

ii) Bootstrap confidence intervals (methods 12 and 13)

Non-parametric bootstrapping is a way to approximate the sampling distribution of a statistic by resampling, from the sample itself, with replacement. Parametric bootstrapping instead samples from a fitted model. Both forms of bootstrapping can be used to make a variety of inferences but are most usually used to quantify the uncertainty in point estimates through the computation of standard errors and CIs. Briefly, bootstrap datasets are sampled (either non-parametrically (method 12) [91,92] or parametrically (method 13)), [93] from which the bootstrap statistics (the statistic of interest calculated using the bootstrap datasets) are calculated. Then the empirical distribution of the bootstrap statistics is taken to approximate the distribution of the statistic of interest. Hence, measures of the uncertainty in the statistic, such as standard errors and CIs, can be calculated from this empirical distribution. In our context, this statistic is the estimated overall effect size.

There is a variety of ways in which the bootstrap samples can be sampled under the random-effects model. For example, we could either sample estimated effect sizes and their standard errors directly, or instead sample the individual patient data in situations where this

is available (this can readily be derived for dichotomous outcome data from the frequency and sample size). A full discussion of all the possibilities is beyond the scope of this paper, but Van Den Noortgate and Onghena [94] describe four different bootstrapping procedures, where two of these are parametric and the other two are non-parametric. We refer the reader to this paper for full details of the sampling methods used. Parametric bootstrap CIs have also been advocated by Turner et al. [93] and non-parametric bootstrap CIs by Efron.[95]

iii) Follmann and Proschan (FP) confidence intervals (method 14)

Permutation tests have been suggested primarily to assess the true statistical significance of an observed finding under the null hypothesis of the absence of effect, especially in meta-analyses with a small number of studies.[96] This method can be extended and used for calculating CIs for the effect size. These tests are especially appropriate when the included studies in a meta-analysis may not be considered randomly sampled from a larger population of studies. Confidence intervals can be constructed by inverting hypothesis tests, where parameter values that are not rejected by the hypothesis test lie within the corresponding CI.

Follmann and Proschan [57] begin by considering a permutation method for testing the null hypothesis $H_0: \mu = 0$. Their argument assumes that the distributions of the outcome data y_i are symmetric [57] and this is implied by the random-effects model. Under the null hypothesis, the sign of y_i, is equally likely to be positive or negative for a symmetric y_i. There are 2^k possible permutations of the signs of the values of the outcome data y_i. We take $\boldsymbol{Z^p}$=(Z_1^p, Z_2^p ,…Z_k^p) to be the pth of these 2^k permutations; for example with $k = 5$ studies, $(+1, +1, -1, -1, +1)$ is one of the 32 possible permutations. We define $\boldsymbol{X^p} = (Z_1^p\,|y_1|, \ldots\, Z_k^p\,|y_k|)$ to be the pth permutation of the outcome data corresponding to $\boldsymbol{Z^p}$. The central idea is that, under the null hypothesis $H_0: \mu = 0$ and because the distributions of

the y_i are symmetric, all 2^k permutations $\boldsymbol{X^p}$ are equally likely. Hence, Follman and Proschan [57] propose the null distribution where all values of

$$\hat{\mu}^p = \sum w_i^p Z_i^p |y_i|$$

are equally likely, where $\widehat{w}_i^p$ are the normalised (sum to one) random-effects weights described in section 2.1 where the between-study variance is estimated using $\boldsymbol{X^p}$ as outcome data. Hence, $\hat{\mu}^p$is the estimated average effect size using the p[th] permutation of signs of the outcome data. The 2-sided p-value based on the group permutation method proposed by Follmann and Proschan [57] is simply the proportion of the absolute values of $\hat{\mu}^p$ that are more than or equal to the absolute value of the estimated average effect under the random-effects model using the observed data. Follman and Proschan [57] describe their procedure in terms of the DerSimonian and Laird[11] estimator of the between-study variance, but, in principle, alterative estimators could be used. If *k* is too large for all permutations to be evaluated then the permutation distribution can be approximated by instead simulating a large number of permutations.[57]

The procedure described above can be extended so that we instead test the hypothesis H_0: $\mu = c$, and invert this hypothesis test to give the bounds of CIs. For further details on the FP method, we refer the reader to Follman and Proschan.[57] This method has been suggested as an alternative approach to the Wald-type and likelihood-based approaches which assume normality of the observed effects, but it can be computationally demanding. The discrete nature of the permutation distribution will ensure that the CI maintains the desired coverage probability, but in general this coverage probability will be larger than the nominal level.

3.1.7 Bayesian credible intervals (method 15)

Bayesian credible intervals (CrIs) for the overall effect size can be obtained within a Bayesian framework using specialised software and the Markov Chain Monte Carlo

(MCMC) technique, such as WinBUGS [97] or SAS PROC MCMC. Some advantages of the Bayesian approach include: 1) incorporation of uncertainty in model parameters (μ, τ^2), 2) derivation of CrIs from the posterior distribution, and 3) use of informative prior distributions on the model parameters. However, the use of informative priors for the effect size parameters has been discouraged by some researchers due to potential inclusion of bias.[98] The use of vague priors allows the analysis to remain data driven. On the contrary, the use of informative priors for the between-study variance has been suggested to increase confidence in the overall effect size, especially when few studies are included in a meta-analysis.[20,21] Informative priors for τ^2 under several treatment comparison types and outcome settings are available for dichotomous data [20] and for continuous data. [21] Friede et al. [99] suggest Bayesian CrIs perform well even in rare diseases with a small number of studies when the appropriate prior for τ^2 is applied. In rare diseases and small populations, the use of half-normal priors, with expectation 0 and variance 0.25 or 1 for τ^2, has been recommended when log-odds ratios are used to measure the effect size.[99] Vague priors can also be applied for τ^2, but caution is needed as results are sensitive to the prior specification, especially when the number of studies is small.[100] This is because the choice of prior may impact on the estimation of the between-study variance and consequently on the estimated overall effect size and the width of its CI. Other difficulties that have been associated with the derivation of Bayesian CrIs include the complication of determining whether convergence is achieved, the need to burn-in when using MCMC, and the impact of MC error. Alternative methods to implement a Bayesian meta-analysis are available by using numerical integration, importance sampling and data augmentation as described by Turner et al. and Rhodes et al. [101,102] For practical application the R package *bayesmeta* is available.[103]

3.2 Comparative evaluation of the methods

The properties of the 15 CI approaches have been evaluated in 31 research papers, including 30 simulation studies and 32 real data evaluations of the methods (for simulation scenarios and study characteristics see Supporting File; Appendices 3-4). Published articles suggested that the different approaches can provide noticeably different or even conflicting results and their performance can vary regarding coverage and CI width. Below we discuss the comparative results as presented in the identified studies. However, it is hard to compare simultaneously all 15 CI approaches, as they have never all been compared under the same conditions and simulation scenarios. The presentation of results follows the same CI presentation order with section 3.1.

Wald-type methods (methods 1, 2, and 3)

The performance of the popular WTz method has been assessed in several studies and it is poor when compared with other methods. Simulations suggest that the WTz performs worse in terms of coverage for small numbers of studies ($k < 16$) compared with the PL and the WTt methods, whereas for large *k* all three methods perform well.[104] The performance of the WTz method though does not only depend on the number of studies, but also on the τ^2 estimator employed and its magnitude.[45] The WTz coverage has been found to differ by up to 0.05 between different τ^2 estimators, up to 0.30 between meta-analysis samples, and up to 0.20 across between-study variance values ranging from small to large τ^2. Coverage has been found to be as low as 0.65 (at 0.95 nominal level) when I^2 (defined as the percentage of the total variability in a set of effect sizes that is due to between-study variability beyond what is expected by within-study random error) is 90% and two or three studies are included in a meta-analysis, but it tends towards the nominal level as the number of studies increases.[53]

To increase coverage, the *t*-distribution can be used, which produces wider CIs than those obtained by the standard normal distribution, especially when τ^2 and k are small.[15,57] The coverage probability is therefore higher with the WTt approach, but it depends on the estimator and the magnitude of τ^2, as well as on the number of studies.[45] Simulation showed that the WTt CI is less affected by the number of studies compared to the WTz CI.[66] Although WTt coverage may be more robust to changes in the τ^2 magnitude compared with WTz when few studies are included in a meta-analysis, it has been found to differ by up to 0.05 depending on the τ^2 estimator used and the number of studies.[53] For large meta-analysis samples (e.g., $k \geq 20$), the coverage of the 95% WTt CI may be below the nominal level, but it becomes conservative (close to 1) when k is small.[53,62,104]

Alternatively, the WTqa method is easy to implement and produces intervals with better coverage in comparison to the WTz method.[62] A simulation study [45] showed that different estimators of the between-study variance may impact on coverage and that the WTqa method is associated with higher coverage than WTz and HKSJ CIs, but the HKSJ method produced values closer to the nominal level. The same study showed that the WTqa method has similar coverage to the WTt method. For small k, coverage of the WTt method is well above the nominal level and higher than that for the WTqa method, but as k increases the differences in coverage are not so important.[62] Simulations have also shown that WTqa outperforms WTz, PL, and ZL approaches, but it is very conservative.[90]

Hartung-Knapp/Sidik-Jonkman methods (method 4 and 5)

The HKSJ approach (method 4) is often preferred, as in case of small k it is conservative and on average produces wider CIs with more adequate type I error compared with the WTz method.[16,59,83,96,105] The HKSJ method provides exact inference when all study sizes are equal and the random-effects model is true, resulting in better inference than WTz, [106] but also provides more accurate inference in small meta-analyses with different study

sizes. [14,15,56] Several studies suggested that the HKSJ method has coverage close to the nominal level, and that it is not influenced by the magnitude or estimator of τ^2. [16,45,53,55,59,61,67,74,77,99] Nevertheless, Knapp and Hartung [67] recommend using the PM [107] estimator for the between-study variance along with the HKSJ method to obtain CIs for μ so as to get a cohesive approach based on Q_{gen}. [107,108] Sanchez-Meca and Marin Martinez [45] recommend using the HKSJ method as it is additionally insensitive to the number of trials. Simulation studies suggest that HKSJ has good coverage when the effect measure is the log-odds ratio, [15,59] the standardised mean difference, [61] the mean difference and the risk difference. [58] The coverage of the 95% HKSJ CI is generally better than the WTz and WTt coverages, but it is suboptimal in meta-analyses with binary outcomes and rare events, as shown in simulated meta-analyses where the odds-ratio was used as the measure of effect. [53]

A real-life data study of 920 Cochrane meta-analyses with $k \geq 3$, showed that the WTz method yielded more often statistically significant results compared with the HKSJ method (45% vs. 35% of meta-analyses).[109] IntHout et al.[16] found similar results in their real-life data study with 434 Cochrane meta-analyses with dichotomous data (43% vs. 34%) and 255 Cochrane meta-analyses with continuous data (51% vs. 40%). It is recommended that caution is needed when fewer than five studies of unequal sizes are included in the meta-analysis. [16] Wiksten et al. [74] in their empirical evaluation including 157 meta-analyses with dichotomous data and $k \geq 4$, found that in the presence of heterogeneity (using the DerSimonian and Laird[11] estimator [$\hat{\tau}^2 > 0$] or the Cochran's Q statistic [$p < 0.10$] [11,110]) the p-value for the overall effect size was typically greater when using the HKSJ than the WTz method. However, they comment that the HKSJ method is not always more conservative when $\hat{\tau}^2 = 0$.

It has been shown that in the absence of heterogeneity the coverage of HKSJ may be smaller than the WTz coverage providing narrower CIs. [15,55,58,61,74,75,111] This was more

prevalent in cases with rare events. [74] Jackson et al. [75] raise a variety of concerns about the use of the HKSJ method, including 1) the implications of the modification for any given meta-analysis are hard to predict , 2) HKSJ can result in shorter CIs for the overall effect size than the WTz method, and 3) the coverage of the HKSJ CI might be anticipated to be low when $\hat{\tau}^2 = 0$. However, in simulation studies conducted by Röver et al. [77], Viechtbauer et al.[76] , and Sanchez-Meca and Marin Martinez [45] HKSJ worked well even in the absence of heterogeneity. This is in line with the simulations by Gonnermann et al. [112], but in the presence of only two studies for $\tau^2 = 0$, HKSJ is associated with very low power compared with WTz (15% vs. 60%), which may be due to the wider CI, whereas for mild to moderate τ^2 both methods have poor control of type I error. A simulation study compared HKSJ with the small sample modification suggested by Knapp and Hartung [67] and indicated that the use of the modified HKSJ (method 5) is preferable when few studies of varying size and precision are available. [77] Another simulation study suggested the use of the modified HKSJ approach instead of the common HKSJ and WTz approaches when dichotomous data are considered. [113] However, for few studies (and particularly for $k = 2$) and as the between-study variance decreases, the modified HKSJ tends to be over-conservative, and selection between the methods is a matter of power vs. type I error. [75-77]

Likelihood-based methods (methods 6, 7, and 8)

The PL method is often preferred to the WTz method, as it is associated with a higher coverage closer to the nominal level, even when k is relatively small. [62,85] Jackson et al. [104] showed that the PL method performed well and better than the WTz and WTt methods in meta-analyses with few studies ($k \leq 8$) with coverage close to the nominal level. However, coverage decreases as τ^2 increases and/or k decreases. [43] Simulations suggest that the PL CI is less affected by the number of studies in a meta-analysis compared to the WTz CI, but both WTz and PL have poor coverage control, as they yield values below the nominal level. [66]

Simulations found that the Bartlett-type correction CI (method 7) improves coverage properties over the WTz, WTt, and PL methods that their coverage deviates the nominal level as τ^2 increases and/or k decreases. [17,66] Although the Bartlett-type correction CI has a satisfactory power compared to the WTz, WTt, and PL CIs, [66] and performs well when $\tau^2 = 0$, [84] caution is needed for $k \leq 5$ as it tends to be over-conservative. [17] The Skovgaard statistic CI (method 8) is associated with coverage closer to the nominal level compared with the WTz and PL CIs, which is remarkable for small k. [17,65,83] Both the Skovgaard statistic CI and the Bartlett-type correction CI perform very satisfactorily regarding coverage and yield similar results. [83]

Henmi and Copas and Biggerstaff and Tweedie methods (methods 9 and 10)

Simulations showed that in the absence of publication bias and for $k > 10$ the HC method yields better coverage than WTz, HKSJ, PL, and BT methods, whereas for $k < 10$ the HKSJ and PL methods perform best. [30] The same study showed that when publication bias is present and for $k > 10$, HC improved coverage compared to WTz, HKSJ, PL, and BT methods, and showed less bias than the fixed-effect model. Also, the WTz and BT methods have comparable coverage probabilities with coverage below the nominal level, [62,88] but coverage is increased for the exact weights. [88]

Resampling methods (methods 11, 12, 13 and 14)

Zeng and Lin [90] showed that the ZL CI outperforms both WTz and PL CIs for small k in terms of coverage. Another simulation study showed that the FP CI controls coverage better than WTz, WTt, PL, and is closely followed by the Bartlett-type correction CI, but the latter is slightly more powerful especially for small k. [66] The same study showed that the FP CI and the Bartlett-type correction CI were less affected by the number of studies than WTz, PL, and WTt methods.

Bayesian credible intervals (method 15)

Simulation studies showed that Bayesian intervals produce intervals with coverage closer to the nominal level compared to the HKSJ, modified HKSJ, and PL CIs, [99,114] and they tend to be smaller than the HKSJ CI even in situations with similar or larger coverage. [99] However, the performance of the Bayesian CrIs may vary depending on the prior assigned to the between-study variance. [100]

3.3 Prediction intervals

One of the most important aims in clinical decision-making is the prediction of the possible effect size in an individual setting. A prediction interval provides a predicted range for the true effect size in a new study, and its calculation is recommended to be conducted under the random-effects model. [24,97] Assuming the random effects are normally distributed an *ad hoc* 95% prediction interval can be obtained by

$$\hat{\mu}_{RE} \pm t_{k-1,0.975}\sqrt{\hat{\tau}^2 + var(\hat{\mu}_{RE})}.$$

To date, this is the standard prediction interval approach used in meta-analysis. We call this prediction interval *ad hoc*, because τ^2 is unknown and currently there is no exact distributional form available. The use of a *t*-distribution instead of a normal distribution reflects the uncertainty resulting from the estimation of the heterogeneity. Higgins et al. [24] also presented this *ad hoc* 95% prediction interval but instead using quantiles from the t_{k-2} distribution. However, we suggest using the *t*-distribution with consistent degrees of freedom for both CIs (e.g., see the WTt and HKSJ methods) and prediction intervals. This is because when there is truly no heterogeneity ($\tau^2 = 0$) the overall effect size and the true effect size in a new study are identical, so that CIs (for the overall effect size) and prediction intervals (for the true effect in a new study) should be identical. Taking the estimated between-study variance of zero to be the true value therefore gives rise to the intuition that CIs and

prediction intervals should be identical when $\hat{\tau}^2 = 0$. In fact, in the *metafor* R package, [115] the CIs and prediction intervals are always computed in a consistent manner: when a WTz CI is computed then a prediction interval is also calculated using a standard normal distribution, whereas when a HKSJ CI is computed then both CI and prediction interval are computed using the t_{k-1} distribution. Hence, when $\hat{\tau}^2 = 0$, the CI and prediction interval will coincide, as intuition suggests that they should. To date, other routines, including the *meta* [116] R package, the Stata *metan* [117] and Stata *mvmeta* [118] commands, calculate a prediction interval using a *t*-distribution with $k - 2$ degrees of freedom. Another advantage of using the t_{k-1} distribution, is that for $k = 2$, where a CI for the overall effect size is available, prediction intervals can be calculated. However, this is not the case when the t_{k-2} distribution is used. Prediction intervals come especially naturally from a Bayesian approach, but at the price of specifying priors. [100]

It is worth noting that the prediction interval does not inform the statistical significance of $\hat{\mu}_{RE}$, it instead describes the region within which the true study effects of new studies are expected to be found. A prediction interval can help understand the uncertainty about whether an intervention is expected to work and reflects the potential effect in future study participants. [119] Prediction intervals are particularly helpful when excess between-study heterogeneity exists, and the combination of individual studies into an overall effect size would not be advisable. IntHout et al.[120] found that in more than 70% of the statistically significant meta-analyses in the Cochrane Database of Systematic Reviews with $\hat{\tau}^2 > 0$, the 95% prediction interval suggested that the effect size in a new study could be null or even in the opposite direction from the overall result in some patient populations. The prediction interval can also be used to calculate the probability that a new trial will have a negative result and to improve the calculations of the power of a new trial. Conclusions drawn from a prediction interval are based on the assumption the study-effects are normally distributed.

The prediction interval estimation will be imprecise if the estimates of the overall effect size and τ^2 are away from the true parameter. Partlett and Riley [121] assessed the performance of the *ad hoc* 95% prediction interval in a simulation study and they concluded that the method is only accurate when heterogeneity is large ($I^2 > 30\%$) and the study sizes are similar. However, for small heterogeneity and different study sizes the coverage of prediction interval can be as low as 78% depending on the between-study variance estimator. [121] Lee and Thompson [122] highlight the importance when calculating a prediction interval to allow for potential skewing and heavy tails in the random-effects distributions. Prediction intervals can be implemented in several software, such as R (using for example *metafor,* [115] and *meta* [116] packages), Stata (using for example *metan* [117] and *mvmeta* [118] commands).

4 Discussion

The estimation of the overall effect size is one of the primary aims in meta-analysis. Therefore, the computation of a confidence/credible interval is crucial in order to interpret the uncertainty in the estimated overall effect size. Wald-type methods, and in particular the WTz CI using the DerSimonian and Laird [11] estimator for the between-study variance, are commonly used and are the default option in several meta-analysis software (e.g., RevMan).[132] However, the accuracy of these standard CI methods is not optimal, as the coverage probability associated with these CIs can deviate considerably from the nominal coverage probability in small meta-analyses. [12-18] This is not surprising as the Wald-type methods rely upon large-sample approximations requiring many studies to be included in a meta-analysis. However, meta-analyses often include a small number of studies, and large-sample approximations can be inaccurate. [19-23] Perhaps because of this property, several other CI methods have been proposed to improve the standard Wald-type CIs, including likelihood-based and resampling methods, and more recently, higher-order likelihood inference

methods. In the present study, we provide a comprehensive review of the CIs for the overall effect size under the random-effects model.

Our review identified 15 methods for calculating a CI for the overall effect size, each of which has different statistical properties. The selection of a method for computing a CI should be based on its statistical performance according to the corresponding meta-analysis' characteristics, as well as on the method's computational and conceptual complexity. Usually one of these comes at the price of another. For example, the likelihood-based methods are associated with coverage closer to the nominal level compared to the commonly used WTz method but are computationally more demanding than the WTz CI. Also, the use of some methods (e.g., ZL) is limited in meta-analyses, due to complex calculations with standard software or their unavailability in statistical software. Simulations have assessed the performance of various methods and showed that it mostly depends on the magnitude of the between-study variance and number of studies in a meta-analysis. However, additional items should be considered when selecting a CI. These may include the type of outcome data and the study size.

The selection of the most preferable methods to calculate a CI for the overall effect size can be mostly based on coverage, as this measure was the only one consistently reported across the identified studies. The 15 methods identified in this review have never all been compared in one simulation study under the same conditions, and hence making any clear recommendations about these methods would be difficult. Also, none of the methods had an optimal coverage across all settings. Therefore, we can only offer tentative recommendations based on the available evidence, but these depend on the study findings, their simulation scenarios, and the CIs examined. It would require an extensive simulation study to assess the performance of all of these methods, under the same, realistic settings. Future studies should evaluate the CIs for all relevant properties, including coverage, precision, complexity, and

power of the corresponding tests. In addition, further research is necessary to make judgements on the performance of CIs. In particular, a comprehensive simulation study informed by real-life data included in a meta-analysis would help determine the factors that impact the performance of the CI methods. Factors to consider in this analysis may include: number and size of studies, baseline risk variability, magnitude and estimator of the between-study variance, frequency of events in dichotomous outcome data, type of outcome data, choice of effect size, distribution of effect sizes, sensitivity to small-study effects or publication bias, and different meta-analytical approaches (e.g., Mantel-Haenszel, Peto or one-step methods).

To date, limited evidence exists to inform which method performs best, especially when studies are few in number (<5), and given that the Bayesian intervals have not been assessed extensively in comparative studies. Overall, studies suggest that the HKSJ method has one of the best performance profiles. It performs well even in meta-analyses with fewer than 10 studies, [28] and is robust to the use of different estimators for the between-study variance and to changes in the magnitude of the between-study variance. [45,53] However, it should be considered that HKSJ is not always conservative compared to a fixed-effect meta-analysis. [74] If the estimated between-study variance is zero, the variance of the estimated overall effect size can be inaccurately small, and hence the HKSJ CI will be too narrow. [74] In such cases, it has been suggested to use the modified HKSJ to avoid inaccurate narrow CIs. [67] Also, caution is needed in meta-analyses with rare events, where the HKSJ coverage has been found to be as low as 85% [53] and meta-analyses with fewer than 5 studies. [28] In the case of few studies, the modified HKSJ has been suggested, [77] but in the case of $k = 2$ the modified HKSJ tends to be overly conservative. [77,112] The likelihood based methods, and in particular the higher order methods Skovgaard statistic CI and Bartlett-type correction CI, are also associated with good coverage properties. [83] However, the higher order likelihood methods

have never been compared directly to HKSJ, which would help make informed decisions on the CI selection. Alternatively, Bayesian intervals may be considered preferable to frequentist intervals in situations where prior information is available and can be considered suitable for use in the meta-analysis. Bender et al. [28] recommend the use of the HKSJ method as a standard approach, but in case studies have considerably different precisions the modified HKSJ should be preferred. The same authors also suggest that when reliable prior information on the between-study variance is available, then the Bayesian intervals with (weakly) informative prior distributions for the heterogeneity should be preferred.

The computation of prediction intervals in meta-analysis is also valuable, as they provide additional information about the overall effect size and we believe that they should be used more frequently. We propose to use $k-1$degrees of freedom rather than $k-2$ to calculate prediction intervals, so that the CIs using a *t*-distribution (e.g., WTt and HKSJ CIs) and prediction intervals are identical when $\hat{\tau}^2 = 0$. Although some concerns have been raised about prediction intervals, including their actual coverage probability of the true effect in a new study and their sensitivity to distributional assumptions, [121,122] their advantages outweigh their disadvantages as they are a nice and easy way for people to interpret the implications of the between-study heterogeneity implied by their fitted model. A comprehensive simulation study assessing the different types of prediction intervals under a variety of meta-analytical scenarios and different between-study variance estimators would help critically examine the issues associated with the calculation and interpretation of prediction intervals.

In conclusion, there are multiple methods to compute a CI for the overall effect size, and none of the methods clearly performs best across all meta-analytical settings. We hope that bringing them all together in one place will facilitate investigators in forming their own judgements about the most appropriate method for their needs. Overall, based on the existing literature and consensus among the co-authors of this paper, we tentatively suggest the

application of the Hartung-Knapp-Sidik-Jonkman method as standard approach, at least in a meta-analysis with 5 or more studies. We recommend conducting a sensitivity analysis using a variety of methods (with at least 2 to 3 methods) to assess the robustness of findings and conclusions, especially in a meta-analysis with fewer than 10 studies. It should be highlighted that these results refer to normally distributed true study-specific effects, and simulation studies are necessary to compare the performance of the 15 methods described in this review. For example, Kontopantelis and Reeves [133] used various non-normal distributions for the effect sizes and compared the WTz, WTqa, HKSJ, PL, BT, and FP methods. The authors showed that simulation results were broadly consistent across different effect size distributions (normally distributed, skew-normal, and extreme non-normal study-effects) with PL providing the best coverage, but with wide CI. Also, the FP method provided coverage close to the nominal level, regardless the included number of studies, at the expense of highly lengthy CIs. The HKSJ method had a consistent 94% coverage for non-normal study-effects and small heterogeneity, but with larger heterogeneity the FP method performed better than the HKSJ CI. For a small number of studies (≤5), the WTz and PL methods performed best, with WTz outperforming PL only when the between-study variance was small. However, more simulation studies with non-normal true study-specific effects are required to draw robust conclusions for the 15 CIs across different meta-analytical scenarios.

We also recommend the calculation of prediction intervals as a supplement to a CI to illustrate the degree of heterogeneity, particularly when large between-study heterogeneity is present. However, caution is needed for small between-study heterogeneity and unequal study sizes. In this case it is advisable to prefer prediction intervals derived in Bayesian framework using for example informative prior distributions. [20,21] Should any new methods become available, we recommend that these are compared to most, or ideally all, of the methods described in this review, and under the same circumstances both using real-life data

and simulation studies. In Appendices 6 to 10 we present a selection of the identified methods for computing a CI to four illustrative, and contrasting, real data examples. We hope that our codes presented in Appendices 8A, B, C and 9 can help to make this possible. This will help obtain a clearer picture about the performance of these methods when these are compared to each other.

List of Abbreviations

BT: Biggerstaff and Tweedie; CI: Confidence interval; CrI: Credible intervals; FP: Follmann and Proschan; HC: Henmi and Copas; HKSJ: Hartung-Knapp/Sidik-Jonkman; IQWiG: Institute for Quality and Efficiency in Health Care; PL: Profile likelihood; WT: Wald-type; WTqa: Quantile approximation; WTt: Wald-type with a t distribution; WTz: Wald-type with a normal distribution; ZL: Zeng and Lin

Availability of Data and Materials

All data are from a published study and are available in the supporting file.

Details of Contributors

AAV, DJ, RB, OK, DL, JPTH, GK, GS contributed to the conception and design of the study, and helped to draft the manuscript. AAV searched PubMed, screened identified articles for eligibility, contacted original study authors when needed and conducted the statistical analysis. All authors read and approved the final manuscript.

Funding

This work did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

DJ is employed by the UK Medical Research Council (code U105260558). RB is employed by the Institute for Quality and Efficiency in Health Care, Cologne, Germany. DL is funded by the Centre for Reviews and Dissemination, University of York

Acknowledgements

We thank Dr. Jack Bowden and Dr. Wolfgang Viechtbauer for providing their feedback on a previous draft of the paper. We thank Annamaria Guolo and Michael Preuss for

responding to our questions regarding their approaches. We also thank Shazia Siddiqui and Myanca Rodrigues for formatting the manuscript.

Competing Interest Declaration

The authors declare that there is no conflict of interest.

Supporting Information

File name: Supporting File

Tile of data: Methods to calculate uncertainty in the estimated overall effect size from a random-effects meta-analysis

Description of Data: The appendices includes all supporting data

Appendix 1: Search Strategy

Appendix 2: Articles included in the review

Appendix 3: Summary of scenarios and trial characteristics used in simulation studies that compared different confidence interval methods

Appendix 4: Real data set studies that compared different confidence interval methods

Appendix 5: Summary of performance measures reported in the 30 included simulation studies

Appendix 6: Illustrative examples

Appendix 7: Data used in illustrative examples

Appendix 8: Codes used for illustrative examples

- Appendix 8A: R
- Appendix 8B: Stata
- Appendix 8C: OpenBUGS

Appendix 9: SAS codes

Appendix 10: Figures produced in illustrative examples

References

1. da Costa BR, Juni P. Systematic reviews and meta-analyses of randomized trials: Principles and pitfalls. *Eur Heart J.* 2014;35(47):3336-3345.
2. WHO. World Health Organization. http://www.who.int/en/. Accessed 08 February, 2017.
3. EMA. European Medicines Agency. http://www.ema.europa.eu/ema/. Accessed 08 February, 2017.
4. CADTH. Canadian Agency for Drugs and Technologies in Health. https://www.cadth.ca/. Accessed 08 February, 2017.
5. IQWiG. Institute for Quality and Efficiency in Health Care. https://www.iqwig.de/. Accessed 06 October 2017.
6. NICE. National Institute for Health and Clinical Excellence. https://www.nice.org.uk/. Accessed 08 February, 2017.
7. Atkins D, Best D, Briss PA, et al. Grading quality of evidence and strength of recommendations. *BMJ.* 2004;328(7454):1490.
8. Chess LE, Gagnier JJ. Applicable or non-applicable: Investigations of clinical heterogeneity in systematic reviews. *BMC Med Res Methodol.* 2016;16:19.
9. Higgins JPT, Green S. *Cochrane Handbook for Systematic Reviews of Interventions.* Version 5.1.0 [updated March 2011] ed: The Cochrane Collaboration; 2011.
10. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: A comparison of methods. *Stat Med.* 1999;18(20):2693-2708.
11. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials.* 1986;7(3):177-188.
12. Cornell JE, Mulrow CD, Localio R, et al. Random-effects meta-analysis of inconsistent effects: A time for change. *Ann Intern Med.* 2014;160(4):267-270.
13. Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. *Stat Med.* 2001;20(6):825-840.
14. Hartung J, Knapp G. A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Stat Med.* 2001;20(24):3875-3889.
15. Sidik K, Jonkman JN. A simple confidence interval for meta-analysis. *Stat Med.* 2002;21(21):3153-3159.
16. IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Med Res Methodol.* 2014;14:25.
17. Noma H. Confidence intervals for a random-effects meta-analysis based on Bartlett-type corrections. *Stat Med.* 2011;30(28):3304-3312.
18. Poole C, Greenland S. Random-effects meta-analyses are not always conservative. *Am J Epidemiol.* 1999;150(5):469-475.
19. Longford NT. Estimation of the effect size in meta-analysis with few studies. *Stat Med.* 2010;29(4):421-430.
20. Turner RM, Davey J, Clarke MJ, Thompson SG, Higgins JP. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane Database of Systematic Reviews. *Int J Epidemiol.* 2012;41(3):818-827.
21. Rhodes KM, Turner RM, Higgins JP. Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continuous outcome data. *J Clin Epidemiol.* 2015;68(1):52-60.

22. Davey J, Turner RM, Clarke MJ, Higgins JP. Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: A cross-sectional, descriptive analysis. *BMC Med Res Methodol.* 2011;11:160.
23. Kontopantelis E, Springate DA, Reeves D. A re-analysis of the Cochrane Library data: The dangers of unobserved heterogeneity in meta-analyses. *PLoS One.* 2013;8(7):e69930.
24. Higgins JP, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *J R Stat Soc Ser A Stat Soc.* 2009;172(1):137-159.
25. Riley RD, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. *BMJ.* 2011;342:d549.
26. Borenstein M, Hedges L, Higgins J, Rothstein H. *Introduction to meta-analysis.* Chichester, UK: Wiley; 2009.
27. McKenzie JE, Beller EM, Forbes AB. Introduction to systematic reviews and meta-analysis. *Respirology.* 2016;21(4):626-637.
28. Bender R, Friede T, Koch A, et al. Methods for evidence synthesis in the case of very few studies. *Res Synth Methods.* 2018;epub ahead of print:1-11.
29. Laird NM, Mosteller F. Some statistical methods for combining experimental results. *Int J Technol Assess Health Care.* 1990;6(1):5-30.
30. Henmi M, Copas JB. Confidence intervals for random effects meta-analysis and robustness to publication bias. *Stat Med.* 2010;29(29):2969-2983.
31. Schmidt FL, Oh IS, Hayes TL. Fixed- versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *Br J Math Stat Psychol.* 2009;62(Pt 1):97-128.
32. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res Synth Methods.* 2010;1(2):97-111.
33. Hunter JE, Smith FL. Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *Int J Select Assess* 2000;8:275-292.
34. Marín-Martínez F, Sánchez-Meca J. Weighting by inverse variance or by sample size in random-effects meta-analysis. *Educ Psychol Meas.* 2009;70(1):56-73.
35. Sanchez-Meca J, Marín-Martínez F. Weighting by inverse variance or by sample size in meta-analysis: A simulation study. *Educ Psychol Meas.* 2016;58(2):211-220.
36. Singh K, Xie M, Strawderman W. Combining information from independent sources through confidence distribution. *Ann Stat.* 2005;33(1):159-183.
37. Xie M, Singh K, Strawderman WE. Confidence distributions and a unifying framework for meta-analysis. *J Am Stat Assoc.* 2011;106(493):320-333.
38. Simmonds MC, Higgins JP. A general framework for the use of logistic regression models in meta-analysis. *Stat Methods Med Res.* 2016;25(6):2858-2877.
39. Mawdsley D, Higgins JP, Sutton AJ, Abrams KR. Accounting for heterogeneity in meta-analysis using a multiplicative model-an empirical study. *Res Synth Methods.* 2017;8(1):43-52.
40. Stanley TD, Doucouliagos H. Neither fixed nor random: Weighted least squares meta-analysis. *Stat Med.* 2015;34(13):2116-2127.
41. Bellio R, Guolo A. Integrated likelihood inference in small sample meta-analysis for continuous outcomes. *Scand J Statist.* 2016;43:191-201.
42. Hoaglin DC. We know less than we should about methods of meta-analysis. *Res Synth Methods.* 2015;6(3):287-289.
43. Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. *Stat Med.* 1996;15(6):619-629.

44. Sidik K, Jonkman JN. A comparison of heterogeneity variance estimators in combining results of studies. *Stat Med.* 2007;26(9):1964-1981.
45. Sanchez-Meca J, Marin-Martinez F. Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychol Methods.* 2008;13(1):31-48.
46. Kulinskaya E, Dollinger MB, Bjørkestøl K. Testing for homogeneity in meta-analysis I. The one-parameter case: Standardized mean difference. *Biometrics.* 2011;67(1):203-212.
47. Kulinskaya E, Dollinger MB, Bjorkestol K. On the moments of Cochran's Q statistic under the null hypothesis, with application to the meta-analysis of risk difference. *Res Synth Methods.* 2011;2(4):254-270.
48. Bohning D, Malzahn U, Dietz E, Schlattmann P, Viwatwongkasem C, Biggeri A. Some general points in estimating heterogeneity variance with the DerSimonian-Laird estimator. *Biostatistics.* 2002;3(4):445-457.
49. Malzahn U, Böhning D, Holling H. Nonparametric estimation of heterogeneity variance for the standardised difference used in meta-analysis. *Biometrika.* 2000;87(3):619-632.
50. Veroniki AA, Jackson D, Viechtbauer W, et al. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Res Synth Methods.* 2016;7(1):55-79.
51. Langan D, Higgins JP, Simmonds M. Comparative performance of heterogeneity variance estimators in meta-analysis: A review of simulation studies. *Res Synth Methods.* 2017;8(2):181-198.
52. Polanin JR, Hennessy EA, Tanner-Smith EE. A review of meta-analysis packages in R. *J Educ Behav Stat.* 2017;42(2):206 - 242.
53. Langan D, Higgins JPT, Jackson D, et al. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Res Synth Methods.* 2018;Accepted.
54. Petropoulou M, Mavridis D. A comparison of 20 heterogeneity variance estimators in statistical synthesis of results from studies: A simulation study. *Stat Med.* 2017;36(27):4266-4280.
55. Hartung J. An alternative method for meta-analysis. *Biom J* 1999;41(8):901-916.
56. Hartung J, Knapp G. On tests of the overall treatment effect in meta-analysis with normally distributed responses. *Stat Med.* 2001;20(12):1771-1782.
57. Follmann DA, Proschan MA. Valid inference in random effects meta-analysis. *Biometrics.* 1999;55(3):732-737.
58. Hartung J, Makambi KH. Reducing the number of unjustified significant results in meta-analysis. *Commun Stat Simul Comput* 2003;32(4):1179–1190.
59. Makambi KH. The effect of the heterogeneity variance estimator on some tests of treatment efficacy. *J Biopharm Stat.* 2004;14(2):439-449.
60. Sidik K, Jonkman JN. Robust variance estimation for random effects meta-analysis. *Comput Stat Data Anal* 2006;50(12):3681–3701.
61. Sidik K, Jonkman JN. On constructing confidence intervals for a standardized mean difference in meta-analysis. *Commun Stat Simul Comput.* 2003;32(4):1191–1203.
62. Brockwell SE, Gordon IR. A simple method for inference on an overall effect in meta-analysis. *Stat Med.* 2007;26(25):4531-4543.
63. Sidik K, Jonkman JN. Simple heterogeneity variance estimation for meta-analysis. *J Roy Stat Soc* 2005;54(2):367–384.
64. Jackson D. The significance level of the standard test for a treatment effect in meta-analysis. *Stat Biopharm Res.* 2009;1(1):92-100.

65. Guolo A. Higher-order likelihood inference in meta-analysis and meta-regression. *Stat Med.* 2012;31(4):313-327.
66. Huizenga HM, Visser I, Dolan CV. Testing overall and moderator effects in random effects meta-regression. *Br J Math Stat Psychol.* 2011;64(Pt 1):1-19.
67. Knapp G, Hartung J. Improved tests for a random effects meta-regression with a single covariate. *Stat Med.* 2003;22(17):2693-2710.
68. Hartung J, Makambi KH. Positive estimation of the between-study variance in meta-analysis : Theory and methods. *S Afr Stat J.* 2002;36(1):55-76.
69. Jackson D, Bowden J. A re-evaluation of the 'quantile approximation method' for random effects meta-analysis. *Stat Med.* 2009;28(2):338-348.
70. Jackson D, Riley RD. A refined method for multivariate meta-analysis and meta-regression. *Stat Med.* 2014;33(4):541-554.
71. Jackson D, Bowden J. Confidence intervals for the between-study variance in random-effects meta-analysis using generalised heterogeneity statistics: Should we use unequal tails? *BMC Medical Research Methodology.* 2016;16(1):118.
72. Jackson D. Confidence intervals for the between-study variance in random effects meta-analysis using generalised Cochran heterogeneity statistics. *Res Synth Methods.* 2013;4(3):220-229.
73. Viechtbauer W. Confidence intervals for the amount of heterogeneity in meta-analysis. *Stat Med.* 2007;26(1):37-52.
74. Wiksten A, Rücker G, Schwarzer G. Hartung-Knapp method is not always conservative compared with fixed-effect meta-analysis. *Stat Med.* 2016;35(15):2503-2515.
75. Jackson D, Law M, Rucker G, Schwarzer G. The Hartung-Knapp modification for random-effects meta-analysis: A useful refinement but are there any residual concerns? *Stat Med.* 2017;36(25):3923-3934.
76. Viechtbauer W, López-López JA, Sanchez-Meca J, Marin-Martinez F. A comparison of procedures to test for moderators in mixed-effects meta-regression models. *Psychol Methods.* 2015;20(3):360-374.
77. Röver C, Knapp G, Friede T. Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. *BMC Med Res Methodol.* 2015;15:99.
78. Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics.* 1997;53(3):983-997.
79. Reid N. The 2000 Wald Memorial Lectures. Aymptotics and the theory of inference. *Ann Stat.* 2003;31(6):1695–1731.
80. Bartlett MS. Properties of sufficiency and statistical tests. *Proc Royal Soc A.* 1937;160:268-282.
81. Skovgaard IM. An explicit large-deviation approximation to one-parameter tests. *Bernoulli.* 1996;2:145-165.
82. Sharma G, Mathew T. Higher order inference for the consensus mean in inter-laboratory studies. *Biom J.* 2011;53(1):128-136.
83. Guolo A, Varin C. Random-effects meta-analysis: The number of studies matters. *Stat Methods Med Res.* 2015;26(3):1500-1518.
84. Rukhin AL. Confidence regions and intervals for meta-analysis model parameters. *Technometrics.* 2015;57(4):547-558.
85. Kosmidis I, Guolo A, Varin C. Improving the accuracy of likelihood-based inference in meta-analysis and meta-regression. *Biometrika.* 2017;asx001.

86. Doi SA, Barendregt JJ, Khan S, Thalib L, Williams GM. Advances in the meta-analysis of heterogeneous clinical trials I: The inverse variance heterogeneity model. *Contemp Clin Trials.* 2015;45(Pt A):130-138.
87. Biggerstaff BJ, Tweedie RL. Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Stat Med.* 1997;16(7):753-768.
88. Preuß M, Ziegler A. A simplification and implementation of random-effects meta-analyses based on the exact distribution of Cochran's Q. *Methods Inf Med.* 2014;53(1):54-61.
89. Biggerstaff BJ, Jackson D. The exact distribution of Cochran's heterogeneity statistic in one-way random effects meta-analysis. *Stat Med.* 2008;27(29):6093-6110.
90. Zeng D, Lin DY. On random-effects meta-analysis. *Biometrika.* 2015;102(2):281-294.
91. Switzer FS, Paese PW, Drasgow F. Bootstrap estimates of standard errors in validity generalization. *Journal of Applied Psychology.* 1992;77:123-129.
92. Efron B, Tibshirani R. *An introduction to the bootstrap.* New York: Chapman & Hall; 1993.
93. Turner RM, Omar RZ, Yang M, Goldstein H, Thompson SG. A multilevel model framework for meta-analysis of clinical trials with binary outcomes. *Stat Med.* 2000;19(24):3417-3432.
94. Van Den Noortgate W, Onghena P. Parametric and nonparametric bootstrap methods for meta-analysis. *Behav Res Methods.* 2005;37(1):11-22.
95. Efron B. 10. Nonparametric confidence intervals. In: *The Jackknife, the Bootstrap and Other Resampling Plans.* Philadelphia: SIAM1982:75-90.
96. Higgins JP, Thompson SG. Controlling the risk of spurious findings from meta-regression. *Stat Med.* 2004;23(11):1663-1682.
97. Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: A comparative study. *Stat Med.* 1995;14(24):2685-2699.
98. Thorlund K, Thabane L, Mills EJ. Modelling heterogeneity variances in multiple treatment comparison meta-analysis-are informative priors the better solution? *BMC Med Res Methodol.* 2013;13:2.
99. Friede T, Röver C, Wandel S, Neuenschwander B. Meta-analysis of few small studies in orphan diseases. *Res Synth Methods.* 2017;8(1):79-91.
100. Lambert PC, Sutton AJ, Burton PR, Abrams KR, Jones DR. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Stat Med.* 2005;24(15):2401-2428.
101. Turner RM, Jackson D, Wei Y, Thompson SG, Higgins JP. Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. *Stat Med.* 2015;34(6):984-998.
102. Rhodes KM, Turner RM, White IR, Jackson D, Spiegelhalter DJ, Higgins JP. Implementing informative priors for heterogeneity in meta-analysis using meta-regression and pseudo data. *Stat Med.* 2016;35(29):5495-5511.
103. Röver C, Friede T. Bayesian random-effects meta-analysis. 2015; version 1.1:https://cran.r-project.org/web/packages/bayesmeta/index.html. Accessed October 27, 2017.
104. Jackson D, Bowden J, Baker R. 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 Infer.* 2010;140(4):961-970.
105. López-López JA, Botella J, Sánchez-Meca J, Marín-Martínez F. Alternatives for mixed-effects meta-regression models in the reliability generalization approach: A simulation study. *J Educ Behav Stat.* 2013;38(5):443 - 469.

106. Sidik K, Jonkman JN. Authors reply. *Stat Med.* 2004;23:159-162.
107. Paule RC, Mandel J. Consensus values and weighting factors. *J Res Natl Inst Stand Technol.* 1982;87(5):377-385.
108. Rukhin AL, Biggerstaff BJ, Vangel MG. Restricted maximum-likelihood estimation of a common mean and the Mandel-Paule algorithm. *J Stat Plan Infer.* 2000;83(2):319-330.
109. Thorlund K, Wetterslev J, Awad T, Thabane L, Gluud C. 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.* 2011;2(4):238-253.
110. Cochran WG. The combination of estimates from different experiments. *Biometrics.* 1954;10:101-129.
111. Veroniki AA, Mavridis D, Higgins JP, Salanti G. Characteristics of a loop of evidence that affect detection and estimation of inconsistency: A simulation study. *BMC Med Res Methodol.* 2014;14:106.
112. Gonnermann A, Framke T, Grosshennig A, Koch A. No solution yet for combining two independent studies in the presence of heterogeneity. *Stat Med.* 2015;34(16):2476-2480.
113. Friedrich T, Knapp G. Generalised interval estimation in the random effects meta regression model. *Comput Stat Data Anal.* 2013;64:165–179.
114. Bodnar O, Link A, Arendacka B, Possolo A, Elster C. Bayesian estimation in random effects meta-analysis using a non-informative prior. *Stat Med.* 2017;36(2):378-399.
115. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw.* 2010;36(3):1–48.
116. Schwarzer G. meta: General package for meta-analysis. 2015; version 4.1-0:https://cran.r-project.org/web/packages/meta/meta.pdf. Accessed 1 September 2017.
117. Harris R, Bradburn M, Deeks J, Harbord R, Altman D, Sterne J. metan: Fixed- and random-effects metaanalysis. *Stata Journal.* 2008;8:3–28.
118. White IR. Multivariate random-effects meta-analysis. *Stata Journal.* 2009;9:40–56.
119. Graham PL, Moran JL. Robust meta-analytic conclusions mandate the provision of prediction intervals in meta-analysis summaries. *J Clin Epidemiol.* 2012;65(5):503-510.
120. IntHout J, Ioannidis JP, Rovers MM, Goeman JJ. Plea for routinely presenting prediction intervals in meta-analysis. *BMJ Open.* 2016;6(7):e010247.
121. Partlett C, Riley RD. Random effects meta-analysis: Coverage performance of 95% confidence and prediction intervals following REML estimation. *Stat Med.* 2017;36(2):301-317.
122. Lee KJ, Thompson SG. Flexible parametric models for random-effects distributions. *Stat Med.* 2008;27(3):418-434.
123. Bowden J, Tierney JF, Copas AJ, Burdett S. Quantifying, displaying and accounting for heterogeneity in the meta-analysis of RCTs using standard and generalised Q statistics. *BMC Med Res Methodol.* 2011;11:41.
124. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med.* 2002;21(11):1539-1558.
125. Guolo A, Varin C. The R package metaLik for likelihood inference in meta-analysis. *J Stat Softw.* 2012;50(7):1–14.
126. *boot: Bootstrap R (S-Plus) functions* [computer program]. 2016.
127. Davison A, Hinkley D. Bootstrap methods and their applications. *Cambridge University Press.* 1997;Cambridge.
128. Harbord RM, Higgins JPT. Meta-regression in Stata. *Stata J.* 2008;8(4):493–519.

129. Thomas N. OpenBUGS website. Overview. 2010; http://www.openbugs.net/w/Overview Accessed 3 September 2017.
130. Lunn DJ, Thomas AB, N., Spiegelhalter D. WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Stat Comput.* 2000;10:325-337.
131. Guddat C, Grouven U, Bender R, Skipka G. A note on the graphical presentation of prediction intervals in random-effects meta-analyses. *Syst Rev.* 2012;1:34.
132. Centre TNC. Review Manager (RevMan). 2014; www.cochrane.org/. Accessed 28 October 2017.
133. Kontopantelis E, Reeves D. Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: A simulation study. *Stat Methods Med Res.* 2012;21(4):409-426.
134. Borenstein M, Hedges L, Higgins JPT, Rothstein H. Comprehensive Meta-Analysis. *104* 2005; version 2:https://www.meta-analysis.com/. Accessed 24 October 2017.
135. Kontopantelis E, Reeves D. MetaEasy: A meta-analysis add-in for Microsoft Excel. *J Stat Softw*. 2009;30(7):1–25.
136. Raudenbush S, Byrk A, Congdon R. HLM 6 for Windows. 2004; http://www.ssicentral.com/hlm/references.html. Accessed 28 September 2017.
137. *Meta-DiSc statistical methods* [computer program]. 2006.
138. *MetaWin: Statistical software for meta-analysis. Version 2* [computer program]. Sunderland, MA.: Sinauer Associates; 2000.
139. Bax L, Yu LM, Ikeda N, Tsuruta H, Moons KG. Development and validation of MIX: Comprehensive free software for meta-analysis of causal research data. *BMC Med Res Methodol.* 2006;6:50.
140. Rasbash J, Charlton C, Browne W, Healy M, Cameron B. MLwiN. Centre forMultilevelModelling. *MLwin: A software package for fitting multilevel models* 2014; http://www.bristol.ac.uk/cmm/software/mlwin/. Accessed 28 September 2017.
141. Wallace BC, Dahabreh IJT, T.A., Lau J, Trow P, Schmid CH. Closing the gap between methodologists and end-users: R as a computational back-end. *J Stat Softw.* 2012;49:1-15.
142. What is R? *The R Project for Statistical Computing* 2008; https://www.r-project.org/. Accessed 29 September 2017.
143. Chung Y, Rabe-Hesketh S, Dorie V, Gelman A, Liu J. A nondegenerate penalized likelihood estimator for variance parameters in multilevel model. *Psychometrika.* 2013;78(4):685-709.
144. Thomas A, O'Hara B, Ligges U, Sturtz S. Making BUGS Open. *R News.* 2006;6(1):12-17.
145. Del Re A, Hoyt W. MAd: Meta-Analysis with mean differences. *R package version 0.8-2* 2014; http://cran.r-project.org/web/packages/MAd. Accessed 23 May 2018.
146. Doebler P. mada: Meta-analysis of diagnostic accuracy. *R package version 0.5.8* 2015; https://cran.r-project.org/web/packages/mada/mada.pdf. Accessed 23 May 2018.
147. van Valkenhoef G, Kuiper J. gemtc: Network meta-analysis using Bayesian methods. *R package version 0.8-2* 2016; https://CRAN.R-project.org/package=gemtc. Accessed 23 May 2018.
148. Möbius T. metagen: Inference in meta analysis and meta regression. *R package version 1.0* 2014; https://CRAN.R-project.org/package=metagen. Accessed 23 May 2018.
149. Debray T, de Jong V. metamisc: Diagnostic and prognostic meta-analysis. *R package version 0.1.* 2017; https://CRAN.R-project.org/package=metamisc. Accessed 23 May 2018.

150. Beath K. metaplus: An R package for the analysis of robust meta-analysis and meta-regression. *R Journal* 2016; https://journal.r-project.org/archive/2016-1/beath.pdfS. Accessed 23 May 2018.
151. Cheung MWL. Meta-analysis using structural equation modeling. 2016; version 0.9.8:https://cran.r-project.org/web/packages/metaSEM/metaSEM.pdf. Accessed 1 October 2017.
152. Huizenga H, Visser I, Dolan C. Hypothesis testing in random effects meta-regression. *Br J Math Stat Psychol.* 2011;64:1-19.
153. Gasparrini A. Multivariate and univariate meta-analysis and meta-regression. 2015; version 0.4.7:https://cran.r-project.org/web/packages/mvmeta/mvmeta.pdf. Accessed 1 October 2017.
154. Chen H. mvtmeta: Multivariate meta-analysis. *R package version 1.0* 2012; https://CRAN.R-project.org/package=mvtmeta. Accessed 23 May 2018.
155. Rücker G, Schwarzer G, Krahn K, König J. Network meta-analysis using frequentist methods. 2016; version 0.9-0:https://cran.r-project.org/web/packages/netmeta/netmeta.pdf. Accessed 29 September 2017.
156. Sturtz S, Ligges U, Gelman A. R2WinBUGS: A package for running WinBUGS from R. *J Stat Softw.* 2005;12(3):1-16.
157. Lumley T. rmeta: Meta-analysis. *R package version 2.16* 2012; https://CRAN.R-project.org/package=rmeta. Accessed 23 May 2018.
158. SAS Institute Inc. SAS Software. 2003; http://www.sas.com/technologies/analytics/statistics/stat/. Accessed 3 October 2017.
159. marandom.sas. SAS® meta-analysis macros. http://www.senns.demon.co.uk/SAS%20Macros/SASMacros.html. Accessed 4 October 2017.
160. PROC IML. SAS/IML 9.22 User's guide. http://support.sas.com/documentation/cdl/en/imlug/63541/PDF/default/imlug.pdf. Accessed 29 October 2017.
161. Littell RC, Milliken GA, Stroup WW, Wolfinger RD. SAS system for mixed models. *SAS Inst Inc, Cary, NC.* 1996:31-63.
162. PROC MIXED. SAS/STAT 9.2 User's guide The MIXED procedure (Book Excerpt). https://support.sas.com/documentation/cdl/en/statugmixed/61807/PDF/default/statugmixed.pdf. Accessed 29 October 2017.
163. PROC GLIMMIX. SAS/STAT 9.2 User's guide The MIXED procedure (Book Excerpt). https://support.sas.com/documentation/cdl/en/statugmixed/61807/PDF/default/statugmixed.pdf. Accessed 29 October 2017.
164. Zhang Z, McArdle JJ, Wang L, Hamagami F. A SAS interface for Bayesian analysis with WinBUGS. *Structural Equation Modeling: A Multidisciplinary Journal.* 2008;15(4):705-728.
165. RASmacro. 2017; https://github.com/rsparapa/rasmacro.
166. PROC MCMC. SAS Support site *The power to know* https://support.sas.com/en/support-home.html.
167. STATA. *Stata Statistical Software* 2013; https://www.stata.com/. Accessed 1 October 2017.
168. Huang L, Dziak J, Wagner A, Lanza S. LCA bootstrap Stata function users' guide (version 1.0). *University Park: The Methodology Center, Penn State.* 2016.
169. Rabe-Hesketh S, Skrondal A, Pickles A. Maximum likelihood estimation of generalized linear models with covariate measurement error. *Stata Journal.* 2003;3(4):386-411.

170. Kontopantelis E, Reeves D. metaan: Random-effects meta-analysis. *Stata Journal*. 2010;10:395–407.
171. Gutierrez RG, Carter S, Drukker DM. sg160: On boundary-value likelihood-ratio tests. *College Station, TX: Stata*. 2001;10:269-273.
172. IBM Corp. IBM SPSS for Windows. 2013; http://www.spss.co.in/. Accessed 29 October 2017.
173. Wilson B D. Meta-analysis stuff. *meanes.sps,* 2010; http://mason.gmu.edu/~dwilsonb/ma.html. Accessed 3 October 2017.
174. Wilson B D. Meta-analysis stuff. *metaf.sps* 2010; http://mason.gmu.edu/~dwilsonb/ma.html. Accessed 3 October 2017.
175. Wilson B D. Meta-analysis stuff. *metareg.sps,* 2010; http://mason.gmu.edu/~dwilsonb/ma.html. Accessed 3 October 2017.
176. Thomas A. BUGS: a statistical modelling package. *RTA/BCS Modular Languages Newsletter* 1994; https://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-latest-news/the-winbugs-project-references-to-use-when-citing-bugs/. Accessed 4 October 2017.

Table 1 *Software options (with packages or macros) for each CI method. To our knowledge, routines for WTqa (method 3), Bartlett-type correction (method 7), and ZL[‖] (method 11) CIs are not available in any of the software options listed below.*

Software	**License Type**	**Confidence/Credible interval Methods**										
		WTz (method 1)	*WTt (method 2)*	*HKSJ* (method 4)*	*Modified HKSJ (method 5)*	*PL (method 6)*	*Skovgaard (method 8)*	*HC[‖] (method 9)*	*BT[†] (method 10)*	*Bootstrap (methods 12, 13)*	*FP (method 14)*	*Bayes (method 15)*
Comprehensive Meta-Analysis	Commercial	Yes	-	Yes	-	-	-	-	-	-		-
Excel - MetaEasy AddIn	Freeware	Yes	Yes	-	-	Yes	-	-	-	-	Yes	-
Excel - MetaXL AddIn	Freeware	Yes	-	-	-	-	-	-	-	-	-	-
HLM	Commercial	-	-	-	-	Yes	-	-	-	-	-	-
Meta-DiSc	Freeware	Yes	-	-	-	Yes	-	-	-	-	-	-
Metawin	Commercial	Yes	-	-	-	-	-	-	-	Yes	-	-
MIX	Commercial	Yes		-	-	-	-	-	-	-	-	-
MLwin	Freeware	Yes		-	-	Yes	-	-	-	Yes	-	Yes
Open Meta Analyst	Freeware	Yes	-	-	-	-	-	-	-	-		-
RevMan	Freeware	Yes	-	-	-	-	-	-	-	-	-	-
R	Freeware	Yes *MAd, meta, metafor, metagen, metalik, metamisc, metaSEM, metatest, metaplus, mvmeta, mvtmeta, netmeta, rmeta*	Yes *metaplus*	Yes *MAd, meta, metafor, metamisc, metatest*	-	Yes *metaLik, metaplus*	Yes *metaLik, metatest*	Yes *metafor*	Yes *metaxa*[††]	Yes *metaplus, boot*	Yes *metafor*	Yes *bayesmeta*[§], *blme, BRugs, gemtc, metamisc R2WinBUGS, SASBUGS rjugs*

SAS	Commercial	Yes *marandom.sas, PROCs GLM and MIXED*	Yes *PROCs GLM and MIXED*	-	-	Yes *marandom.sas, PROC NLP*	-	-	-	-	-	Yes *PROC MCMC*
Stata	Commercial	Yes *metaan, metan, metareg, mvmeta, xtreg*	-	-	Yes *metareg*	Yes *gllamm, metaan*	-	-	-	Yes *bootstrap*	Yes *metaan*	-
SPSS	Commercial	Yes *meanes.sps, metaf.sps, metareg.sps*	-	-	-	-	-	-	-	-	-	-
BUGS, OpenBUGS, WinBUGS	Freeware	-		-		-	-	-	-	-		Yes

[I] A resampling test is available in the R package *metatest* [66].
[II] Henmi and Copas [30] provide an R code to implement the HC method.
* IntHout et al [16] provide an approach to easily convert WTz CIs to HKSJ CIs.
† Biggerstaff and Tweedie [87] provide a SAS code to implement the BT method.
†† This package uses the exact random-effects weights in the Biggerstaff and Tweedie approach. [88]
§ Bayesian approaches can be implemented using the Markov Chain Monte Carlo (MCMC) techniques in several software, such as OpenBUGS,[129] WinBUGS[130] or without MCMC as described by Turner et al, [101] in the R package *bayesmeta.* [103]

Comprehensive Meta-Analysis[134] www.meta-analysis.com/
Excel using the MetaEasy AddIn[135] https://www.jstatsoft.org/article/view/v030i07 or MetaXL AddIn http://www.epigear.com/
HLM [136] http://www.ssicentral.com/hlm/
Meta-DiSc[137] ftp://ftp.hrc.es/pub/programas/metadisc/
Metawin[138] http://www.metawinsoft.com/
MIX[139] www.mix-for-meta-analysis.info/
MLwin[140] http://www.bristol.ac.uk/cmm/software/mlwin/
Open Meta Analyst[141] http://www.cebm.brown.edu/openmeta/
RevMan[132] www.cochrane.org/
R[142]http://www.r-project.org/ Packages: *bayesmeta,* [103] *blme,* [143] *boot,* [126,127] *BRugs,* [144] *Mad,* [145] *mada,* [146] *meta,* [116] *gemtc,* [147] *metafor,* [115] *metagen,* [148] *metaLik,* [65,125] *metamisc,* [149] *metaplus,* [150] *metaSEM,* [151] *metatest,* [66,152] *metaxa,* [88] *mvmeta,* [153] *mvtmeta,* [154] *netmeta,* [155]) *R2WinBUGS,* [156] *rjugs,* [144] *rmeta* [157]
SAS[158] http://www.sas.com/technologies/analytics/statistics/stat/ Macros: *marandom.sas,* [159] *PROC IML,* [160] *PROC MIXED,* [161,162] *PROC GLIMMIX,* [163] *SASBUGS,*[164] *RASmacro,* [165] *PROC MCMC*[166]
Stata[167] www.stata.com / Routines: *bootstrap,* [168] *gllamm,*[169] *metaan,*[170] *metareg,*[128] *metan,*[117] *mvmeta,*[118] *xtreg,* [171]
SPSS[172] http://www.spss.co.in/ Macros: *meanes.sps,* [173] *metaf.sps,* [174] *metareg.sps* [175]
BUGS, [176] OpenBUGS, [129] WinBUGS[130] www.mrc-bsu.cam.ac.uk/bugs/

ABBREVIATIONS: BT, Biggerstaff and Tweedie; FP, Follmann and Proschan; HC, Henmi and Copas; HKSJ, Hartung-Knapp/Sidik-Jonkman; PL, Profile likelihood; WTz, Wald-type with a normal distribution; WTt, Wald-type with a t distribution; WTqa, Quantile approximation; ZL,