STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:37–52
Published online 6 February 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2514

Confidence intervals for the amount of heterogeneity in meta-analysis

Wolfgang Viechtbauer*,†

Department of Methodology and Statistics, University of Maastricht, P.O. Box 616, 6200 MD Maastricht, The Netherlands

SUMMARY

Effect size estimates to be combined in a systematic review are often found to be more variable than one would expect based on sampling differences alone. This is usually interpreted as evidence that the effect sizes are heterogeneous. A random-effects model is then often used to account for the heterogeneity in the effect sizes. A novel method for constructing confidence intervals for the amount of heterogeneity in the effect sizes is proposed that guarantees nominal coverage probabilities even in small samples when model assumptions are satisfied. A variety of existing approaches for constructing such confidence intervals are summarized and the various methods are applied to an example to illustrate their use. A simulation study reveals that the newly proposed method yields the most accurate coverage probabilities under conditions more analogous to practice, where assumptions about normally distributed effect size estimates and known sampling variances only hold asymptotically. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: meta-analysis; heterogeneity; confidence intervals; random-effects model

1. INTRODUCTION

Clinical trials examining the same treatment under identical experimental conditions are not expected to yield the exact same results due to sampling variability. Consequently, the effect size estimates derived from such a set of studies (e.g. the odds ratios, risk differences, or risk ratios) are also not expected to coincide. However, when conducting a meta-analysis, it is common to find additional variability in the effect size estimates over and beyond the amount one would expect based on sampling differences alone. This finding is typically interpreted as indicating that the effect sizes (i.e. the parameters estimated by the corresponding effect size estimates) are heterogeneous [1–5].

*Correspondence to: Wolfgang Viechtbauer, Department of Methodology and Statistics, University of Maastricht, P.O. Box 616, 6200 MD Maastricht, The Netherlands.
†E-mail: wolfgang.viechtbauer@stat.unimaas.nl

Received 9 September 2005
Accepted 21 December 2005

Heterogeneity may be a result of moderators, or in other words, due to systematic differences between the studies from which the effect size estimates were derived [2–4]. For example, a higher treatment dose, length, or intensity may yield a higher treatment effect and consequently a higher effect size. When information about the relevant characteristics of the studies are available, then it may be possible to account for the heterogeneity using appropriately formulated meta-regression models [2, 3, 6].

On the other hand, heterogeneity may also be a result of random differences between the effect sizes. In that case, the heterogeneity cannot be accounted for based on a set of moderators, but can be modelled by assuming that the effect sizes themselves are drawn from a distribution (an effect size population), characterized by its mean, indicating the expected effect size, and its variance, indicating the amount of heterogeneity [1].

Several significance tests for the presence of heterogeneity (homogeneity tests) have been proposed in the literature. The results from the so-called Q-test [1] are usually reported as part of any systematic review. Wald, likelihood ratio, and score tests for homogeneity have also been derived [7, 8] and some improvements to the Q-test have been suggested [9, 10]. However, it may be more important to actually quantify the extent of the heterogeneity than to rely on an overall statistical test to detect its presence [2]. In fact, quantifying the amount of heterogeneity and exploring its sources are among the most important aspects of systematic reviews [2–5].

Consequently, a large number of methods have been proposed for estimating the amount of heterogeneity, including two different method-of-moments estimators [1, 11], maximum-likelihood [7] and restricted maximum-likelihood estimators [12], and an empirical Bayes estimator [13]. Estimates of the amount of heterogeneity can also be obtained with fully Bayesian approaches [14]. Finally, Sidik and Jonkman [15] recently derived another promising estimator based on weighted least squares.

In addition to a point estimate, it may also be useful to report a confidence interval for the amount of heterogeneity, which not only indicates the precision of the heterogeneity estimate, but also communicates all the information contained in corresponding homogeneity tests. Various methods for constructing such confidence intervals have been proposed, including profile likelihood [7], Wald-type [16], and bootstrap [17, 18] confidence intervals. Biggerstaff and Tweedie [16] devised a method using an approximate distribution of the Q-statistic and Sidik and Jonkman [15] suggested a method for constructing confidence intervals based on their heterogeneity estimator.

In the present paper, a new method for constructing confidence intervals for the amount of heterogeneity is proposed, which guarantees nominal coverage probabilities even in small samples when model assumptions are satisfied. The proposed method may therefore constitute an improvement over the existing approaches.

The outline of the paper is as follows. In the next section, the meta-analytic random-effects model is briefly outlined. In Section 3, the new method for constructing confidence intervals for the amount of heterogeneity is described. Various existing methods for constructing such intervals are summarized in Section 4. An example is given in Section 5 to illustrate that the methods can yield noticeably divergent results. Some conjectures about the properties of the methods are discussed in Section 6. However, since the methods are not based on closed-form solutions and instead require the use of iterative procedures, it is difficult to compare their accuracy analytically. Therefore, Monte-Carlo simulations were conducted to determine whether one of the methods should be preferred over the others. The

simulation methods and results are described in Section 7. Some final remarks conclude the paper.

2. RANDOM-EFFECTS MODEL

Assume that $i = 1, \ldots, k$ independent effect size estimates have been derived from a set of studies. Each effect size estimate Y_i estimates a corresponding true effect size θ_i and therefore is subject to sampling error. Letting ε_i denote the amount by which Y_i deviates from its parameter, we can express this by writing

$$Y_i = \theta_i + \varepsilon_i$$

We assume that $\varepsilon_i \sim \mathrm{N}(0, \sigma_i^2)$, where σ_i^2 denotes the amount of sampling variability in the ith effect size estimate, and $\mathrm{Cov}(\varepsilon_i, \varepsilon_{i'}) = 0$ for $i \neq i'$ due to independence. Each effect size θ_i is assumed to be sampled from a population of effect sizes with expected value μ and variance τ^2. Denoting the difference between μ and θ_i by u_i, we can then write the meta-analytic random-effects model as

$$Y_i = \mu + u_i + \varepsilon_i$$

where it is assumed that $u_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \tau^2)$ and $\mathrm{Cov}(u_i, \varepsilon_{i'}) = 0$ for all i and i'.

While μ denotes the average effect size in the population, τ^2 denotes the amount of heterogeneity in the effect sizes. The special case where $\tau^2 = 0$ implies that the effect sizes are homogeneous ($\theta_i = \theta$, $i = 1, \ldots, k$) and the resulting model ($Y_i = \theta + \varepsilon_i$) is usually called the fixed-effects model in meta-analysis. Estimates of the σ_i^2 values can be easily computed for all effect size measures typically used in meta-analysis [19] and are treated as known constants. Therefore, the only unknown parameters in the random-effects model are μ and τ^2.

3. Q-PROFILE CONFIDENCE INTERVALS FOR τ^2

Under the null hypothesis $H_0 : \tau^2 = 0$

$$Q = \sum \frac{(Y_i - \hat{\theta})^2}{\sigma_i^2} \tag{1}$$

is distributed χ^2 with $k - 1$ degrees of freedom, where $\hat{\theta} = \sum w_i Y_i / \sum w_i$ and $w_i = 1/\sigma_i^2$. Equation (1) is the usual Q-statistic that is used to test whether the effect sizes are homogeneous or not [1]. Now denote $\hat{\mu} = \sum w_i Y_i / \sum w_i$, where $w_i = 1/(\tau^2 + \sigma_i^2)$. It is easy to show under the random-effects model that a generalized Q-statistic, given by

$$Q(\tau^2) = \sum \frac{(Y_i - \hat{\mu})^2}{\tau^2 + \sigma_i^2} \tag{2}$$

also follows a χ^2 distribution with $k-1$ degrees of freedom (the proof is completely analogous to the one given by Rao [20, pp. 389–390]). Therefore, letting $\chi^2_{k-1;0.025}$ and $\chi^2_{k-1;0.975}$ denote

the 2.5th and 97.5th percentiles of a χ^2 distribution with $k-1$ degrees of freedom, it follows that $P(\chi^2_{k-1;0.025} \leqslant Q(\tau^2) \leqslant \chi^2_{k-1;0.975}) = 0.95$. Using the inversion principle as discussed by Casella and Berger [21], it follows that the lower and upper bounds of a confidence interval with 95 per cent coverage probability are given by those two $\tilde{\tau}^2$ values, where

$$(Q(\tilde{\tau}^2) = \chi^2_{k-1;0.975}, \quad Q(\tilde{\tau}^2) = \chi^2_{k-1;0.025})$$

The two $\tilde{\tau}^2$ values can be found iteratively by repeatedly computing $Q(\tau^2)$ for increasing τ^2 values (i.e. by profiling the generalized Q-statistic) until the lower and upper critical values of the χ^2 distribution are reached. Since negative values of τ^2 are outside of the parameter space, this iterative scheme should be constrained to non-negative τ^2 values and therefore always yields a non-negative lower bound. If $Q(\tau^2 = 0) < \chi^2_{k-1,0.025}$, then this implies that the upper (and therefore also the lower) bound actually falls below 0. In this case, the interval is set equal to the null set. A similar approach was recently suggested as a method for constructing confidence intervals for the among-group variance in the one-way random effects model with unequal error variances [22].

4. OTHER METHODS FOR CONSTRUCTING CONFIDENCE INTERVALS FOR τ^2

Various other methods for constructing confidence intervals for τ^2 have been proposed in the literature and are briefly summarized in the present section.

4.1. Biggerstaff–Tweedie confidence intervals

Biggerstaff and Tweedie [16] showed that the expected value and variance of (1) are equal to

$$E[Q] = (k-1) + \left(S_1 + \frac{S_2}{S_1}\right)\tau^2 \tag{3}$$

and

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

where $S_t = \sum (1/\sigma_i^2)^t$. Next, they approximated the distribution of Q with a gamma distribution with shape and scale parameters equal to

$$\gamma(\tau^2) = \frac{(E[Q])^2}{\mathrm{Var}[Q]} \quad \text{and} \quad \phi(\tau^2) = \frac{\mathrm{Var}[Q]}{E[Q]}$$

respectively (the scale parameter given by Biggerstaff and Tweedie is usually called the rate parameter, which is simply the inverse of the scale parameter [21]). Therefore, lower and upper bounds of a 95 per cent confidence interval for τ^2 can be obtained by finding those two values of $\tilde{\tau}^2$, such that

$$\int_{Q/\phi(\tilde{\tau}^2)}^{\infty} f(x|\gamma(\tilde{\tau}^2))\,\mathrm{d}x = 0.025$$

and

$$\int_0^{Q/\phi(\tilde{\tau}^2)} f(x|\gamma(\tilde{\tau}^2))\,\mathrm{d}x = 0.025 \tag{5}$$

where $f(x|\gamma(\tau^2))$ denotes the density function of a gamma distribution with shape parameter $\gamma(\tau^2)$ and scale parameter 1.

The iterative scheme used to find the two $\tilde{\tau}^2$ values is again constrained to non-negative values. Therefore, the lower bound is also always non-negative. If the integral in (5) is smaller than 0.025 for $\tilde{\tau}^2 = 0$, then the upper (and therefore also the lower) bound falls below 0 and the interval is equal to the null set.

4.2. *Profile likelihood confidence intervals*

Since $Y_i \sim \mathrm{N}(\mu, \tau^2 + \sigma_i^2)$ under the random-effects model, the log-likelihood function of μ and τ^2 is given by

$$l(\mu, \tau^2) = -\frac{1}{2}\sum \ln(\tau^2 + \sigma_i^2) - \frac{1}{2}\sum \frac{(Y_i - \mu)^2}{\tau^2 + \sigma_i^2}$$

leaving out the additive constant. Maximum-likelihood (ML) estimates of μ and τ^2 are then easily obtained by starting with an initial guess for τ^2 and iterating until convergence between the two estimating equations

$$\hat{\mu} = \frac{\sum w_i Y_i}{\sum w_i} \tag{6}$$

and

$$\hat{\tau}^2 = \frac{\sum w_i^2[(Y_i - \hat{\mu})^2 - \sigma_i^2]}{\sum w_i^2}$$

with $w_i = 1/(\tau^2 + \sigma_i^2)$ [23]. Should $\hat{\tau}^2$ converge to a negative value, it is truncated to zero. Denote the final ML estimates as $\hat{\mu}_{\mathrm{ML}}$ and $\hat{\tau}^2_{\mathrm{ML}}$.

A confidence interval for τ^2 can now be obtained by profiling the likelihood ratio statistic [7]. Denote $\tilde{\mu}$ as that value of (6) with $w_i = 1/(\tilde{\tau}^2 + \sigma_i^2)$. A 95 per cent confidence interval for τ^2 is then given by the set of $\tilde{\tau}^2$ values satisfying

$$l(\tilde{\mu}, \tilde{\tau}^2) > l(\hat{\mu}_{\mathrm{ML}}, \hat{\tau}^2_{\mathrm{ML}}) - 3.84/2$$

Alternatively, one can base the confidence interval on the restricted log-likelihood, which is given by

$$l_R(\tau^2) = -\frac{1}{2}\sum \ln(\tau^2 + \sigma_i^2) - \frac{1}{2}\ln \sum \frac{1}{\tau^2 + \sigma_i^2} - \frac{1}{2}\sum \frac{(Y_i - \hat{\mu})^2}{\tau^2 + \sigma_i^2}$$

leaving out additive constants. The restricted maximum-likelihood (REML) estimate of τ^2 is obtained by iterating through

$$\hat{\tau}^2 = \frac{\sum w_i^2[(Y_i - \hat{\mu})^2 - \sigma_i^2]}{\sum w_i^2} + \frac{1}{\sum w_i}$$

until convergence, with $\hat{\mu}$ and w_i defined as before. A negative τ^2 estimate is again truncated to zero. Denoting the REML estimate as $\hat{\tau}^2_{\text{REML}}$, a 95 per cent confidence interval for τ^2 is then given by the set of $\tilde{\tau}$ values satisfying

$$l_R(\tilde{\tau}^2) > l_R(\hat{\tau}^2_{\text{REML}}) - 3.84/2$$

Since the ML and REML estimates of τ^2 are constrained to be non-negative, the lower bound of the profile likelihood intervals is also always non-negative, while the upper bound must consequently be positive.

4.3. Wald-type confidence intervals

The asymptotic sampling variances of the ML and REML estimates of τ^2 can be obtained by taking the inverse of the Fisher information and are equal to

$$\text{Var}[\hat{\tau}^2_{\text{ML}}] = 2(\textstyle\sum w_i^2)^{-1} \tag{7}$$

and

$$\text{Var}[\hat{\tau}^2_{\text{REML}}] = 2\left(\sum w_i^2 - 2\frac{\sum w_i^3}{\sum w_i} + \frac{(\sum w_i^2)^2}{(\sum w_i)^2}\right)^{-1} \tag{8}$$

respectively. Estimates of the sampling variances are obtained by setting $w_i = 1/(\hat{\tau}^2_{\text{ML}} + \sigma_i^2)$ and $w_i = 1/(\hat{\tau}^2_{\text{REML}} + \sigma_i^2)$ in (7) and (8), respectively.

Based on the asymptotic normality assumption of ML and REML estimates, 95 per cent Wald-type confidence intervals [16] for τ^2 are then given by

$$\hat{\tau}^2_{\text{ML}} \pm 1.96\sqrt{\text{Var}[\hat{\tau}^2_{\text{ML}}]}$$

and

$$\hat{\tau}^2_{\text{REML}} \pm 1.96\sqrt{\text{Var}[\hat{\tau}^2_{\text{REML}}]}$$

One can either leave a lower bound that falls below 0 unchanged (which has the advantage that the interval provides a better indication of the precision of the τ^2 estimate) or truncate it to zero (which avoids a bound that falls outside the parameter space). Since the interval is always constructed around a non-negative τ^2 estimate, the upper bound must be positive.

4.4. Sidik–Jonkman confidence intervals

Sidik and Jonkman [15] recently suggested a new heterogeneity estimator and, based on it, a method for obtaining confidence intervals for τ^2. The proposed method works as follows. First, a rough estimate of τ^2 is calculated with

$$\hat{\tau}^2_0 = \frac{1}{k}\sum (Y_i - \bar{Y})^2$$

where $\bar{Y}$ is the unweighted average of the Y_i values. Next, calculate $\hat{\mu}_0$ with (6), where $w_i = 1/(\hat{\tau}^2_0 + \sigma_i^2)$. The heterogeneity estimator is then given by

$$\hat{\tau}^2_{\text{SH}} = \frac{\hat{\tau}^2_0}{k-1}\sum w_i(Y_i - \hat{\mu}_0)^2$$

Finally, based on the assumption that $(k-1)\hat{\tau}^2_{\mathrm{SH}}/\tau^2$ approximately follows a χ^2 distribution with $k-1$ degrees of freedom, Sidik and Jonkman suggested that a 95 per cent confidence interval for τ^2 can be obtained with

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

Note that the heterogeneity estimator $\hat{\tau}^2_{\mathrm{SH}}$ is always greater than zero. This implies that the lower and upper bounds of the confidence interval are also always greater than zero.

4.5. *Parametric bootstrap confidence intervals*

Any consistent heterogeneity estimator can be used in conjunction with parametric bootstrapping [24] to obtain confidence intervals for τ^2 [18]. Let $\hat{\tau}^2$ denote the value obtained from any non-negative and consistent estimator of τ^2 and $\hat{\mu}$ the value of (6) with $w_i = 1/(\hat{\tau}^2 + \sigma_i^2)$. Then parametric bootstrap confidence intervals are obtained as follows. Generate k values of Y_i from $\mathrm{N}(\hat{\mu}, \hat{\tau}^2 + \sigma_i^2)$. Next, estimate τ^2 based on this bootstrap sample and denote the estimate as $\hat{\tau}^2_b$. Repeat this process $b = 1, \ldots, B$ times. A 95 per cent parametric bootstrap interval using the percentile method [24] is then given by the 2.5th and 97.5th empirical percentiles of the $\hat{\tau}^2_b$ values (which are approximately equal to the $(B \times 0.025)$th and $(B \times 0.975)$th ordered $\hat{\tau}^2_b$ values).

4.6. *Non-parametric bootstrap confidence intervals*

Non-parametric bootstrapping [24] can also be applied in this context to obtain confidence intervals for τ^2 [17]. For this, we sample k times with replacement from the Y_i and corresponding σ_i^2 values and estimate τ^2 based on the bootstrap sample. Denoting each bootstrap estimate as $\hat{\tau}^2_b$, we repeat this process $b = 1, \ldots, B$ times. The confidence interval is then again given by the 2.5th and 97.5th empirical percentiles of the $\hat{\tau}^2_b$ values.

If the heterogeneity estimate used in conjunction with the bootstrapping can yield negative estimates, then one has the option of either leaving negative estimates unchanged (which can yield negative confidence interval bounds) or to truncate them to zero (thereby also constraining the bounds to be non-negative).

5. EXAMPLE

Consider Table I, which provides the results from 9 clinical trials on the effectiveness of taking diuretics for preventing pre-eclampsia during pregnancy [25] (the data given here were adapted from References [7, 16]). The log of the odds ratios (Y_i) is used as the effect size measure, since its sampling distribution is approximately normally distributed. The estimated sampling variances of the effect size estimates ($\hat{\sigma}_i^2$) are also given in the table. With $Q = 27.3$ ($\mathrm{df} = 8$, $p < 0.001$), there is little doubt that heterogeneity is present in the effect sizes.

Table II shows that the ML and REML estimates of τ^2 are equal to 0.24 and 0.30, respectively. The fact that the MLE is smaller is not surprising: ML estimates of variance components are typically negatively biased [26, 27] and the MLE of τ^2 in the random-effects

Table I. Results for 9 trials on the effects of diuretics on pre-eclampsia.

Study	Cases/Total		Odds ratio	Y_i	$\hat{\sigma}_i^2$
	Treated	Control			
1	14/131	14/136	1.04	0.04	0.160
2	21/385	17/134	0.40	−0.92	0.118
3	14/57	24/48	0.33	−1.12	0.178
4	6/38	18/40	0.23	−1.47	0.299
5	12/1011	35/760	0.25	−1.39	0.114
6	138/1370	175/1336	0.74	−0.30	0.015
7	15/506	20/524	0.77	−0.26	0.121
8	6/108	2/103	2.97	1.09	0.686
9	65/153	40/102	1.14	0.14	0.068

Table II. Point estimates and confidence intervals for τ^2 for the diuretic and pre-eclampsia data in Table I.

Method	Point estimate	Confidence interval
Q-profile	$\hat{\tau}^2_{\text{DL}} = 0.23$	(0.07, 2.20)
Biggerstaff–Tweedie	$\hat{\tau}^2_{\text{DL}} = 0.23$	(0.05, 2.36)
Profile likelihood (ML)	$\hat{\tau}^2_{\text{ML}} = 0.24$	(0.03, 1.13)
Profile likelihood (REML)	$\hat{\tau}^2_{\text{REML}} = 0.30$	(0.04, 1.47)
Wald-type (ML)	$\hat{\tau}^2_{\text{ML}} = 0.24$	(−0.10, 0.58)
Wald-type (REML)	$\hat{\tau}^2_{\text{REML}} = 0.30$	(−0.13, 0.73)
Sidik–Jonkman	$\hat{\tau}^2_{\text{SH}} = 0.46$	(0.21, 1.67)
Parametric bootstrap	$\hat{\tau}^2_{\text{DL}} = 0.23$	(−0.02, 0.70)
Non-parametric bootstrap	$\hat{\tau}^2_{\text{DL}} = 0.23$	(0.03, 0.49)

did not truncate estimate for parametric bootstrapping; hence the negative lower bound

model is no exception to this rule [28]. The frequently used DerSimonian–Laird estimator [1], given by

$$\hat{\tau}^2_{\text{DL}} = \frac{Q - (k-1)}{\sum w_i + \dfrac{\sum w_i^2}{\sum w_i}}$$

with Q as defined in (1) and $w_i = 1/\hat{\sigma}_i^2$, was also calculated and is equal to 0.23. Finally, estimating τ^2 with the method suggested by Sidik and Jonkman [15] yields a value of 0.46.

The confidence intervals obtained with the various methods discussed earlier are also given in Table II. The DerSimonian–Laird estimator was used for the parametric and non-parametric bootstrap methods due to its ubiquitous use and ease of calculation. It is quite apparent that there are some substantial discrepancies between the various methods. Some lower bounds include the value zero, implying that the effect sizes may be homogeneous. On the other hand, the lower bound falls above zero for other methods, suggesting that heterogeneity is present. There are also large differences with respect to the upper bounds.

6. PROPERTIES OF CONFIDENCE INTERVALS FOR τ^2

The large discrepancies observed in the previous example raise the question whether we should give more credence to the bounds obtained with one of the methods. Based on previous research and theoretical considerations, it is possible to make some general conjectures about the accuracy of the various methods. First of all, profile likelihood confidence intervals for μ in the random-effects model have been shown to be quite accurate [23]. However, whether this is also true when constructing intervals for τ^2 has not been verified. Based on research in related contexts [29], we may also expect adequate coverage probabilities for τ^2, with intervals based on REML estimation possibly having a slight advantage [8, 30]. However, we can already predict that profile likelihood intervals will capture the parameter too often when $\tau^2 = 0$ or close to it. Specifically, when $\tau^2 = 0$, then the asymptotic distribution of the likelihood ratio statistic is not χ^2 with 1 degree of freedom (as used in the construction of the intervals), but a 50:50 mixture of a degenerate random variable with all of its probability mass concentrated at 0 and a χ^2 random variable with 1 degree of freedom [31, 32]. By ignoring this fact, we expect to obtain a coverage probability around 97.5 per cent instead of the nominal 95 per cent when $\tau^2 = 0$.

Wald-type intervals are not expected to yield adequate coverage probabilities. Based on a simulation study examining the statistical properties of the Wald test for homogeneity [8], we can expect the coverage probabilities to be well above the nominal 95 per cent level when $\tau^2 = 0$. Although the properties of Wald-type intervals for values of $\tau^2 > 0$ are unknown at this point, it is generally acknowledged that the normal distribution provides a poor approximation to the distribution of ML and REML estimates of variance components [29].

The accuracy of the method suggested by Biggerstaff and Tweedie [16] depends on how well the gamma distribution approximates the true distribution of the Q-statistic. When $\tau^2 = 0$, then the gamma distribution used in the Biggerstaff and Tweedie method simplifies to a χ^2 distribution with $k - 1$ degree of freedom, which is the exact distribution of Q in that case. Therefore, the coverage probability should be nominal when $\tau^2 = 0$. However, since the distribution of Q for $\tau^2 > 0$ is quite complicated, one cannot determine analytically how well the method works in general (one may suspect that Q then follows a non-central χ^2 distribution, but (3) implies that the non-centrality parameter would then have to be $(S_1 + S_2/S_1)\tau^2$, which in turn implies that the variance of Q would have to be $2(k-1) + 4(S_1 + S_2/S_1)\tau^2$, but that result does not match (4)).

As discussed earlier, the lower and upper bounds of the confidence interval obtained with the method proposed by Sidik and Jonkman [15] are always greater than zero. Therefore, the coverage probability of this method must actually be zero when $\tau^2 = 0$, since the interval can never capture the parameter in this case. The coverage probability does appear to approach the nominal value as τ^2 increases [15], but it is unknown how this method compares to the other methods in terms of accuracy.

Bootstrap confidence intervals have been recommended, because their use relaxes certain distributional assumptions [18]. Specifically, the parametric bootstrap intervals do not assume normally distributed heterogeneity estimates and the non-parametric bootstrap intervals have the added advantage of not assuming normally distributed effect size estimates. However, the accuracy of these methods in the present context has not been established yet.

The method by Biggerstaff and Tweedie [16] relies on the gamma distribution approximation, which is only exact when $\tau^2 = 0$. On the other hand, the Q-profile method does not

rely on an approximation at all, since the generalized Q-statistic given in (2) is exactly χ^2 distributed, assuming the assumptions of the model hold and that the sampling variances are known constants. Under these conditions, we are guaranteed to obtain confidence intervals with nominal coverage probabilities when using this method.

Finally, it should be noted that the profile likelihood, Wald-type, and bootstrap intervals are based on asymptotic results, relying on large k for their nominal performance. On the other hand, the Q-profile, Biggerstaff–Tweedie, and Sidik–Jonkman intervals make no assumptions about the size of k. To what extent this is relevant for practice (i.e. how large is 'large'?) is difficult to say without further analysis.

7. MONTE CARLO SIMULATIONS

All of the proposed methods, except the one suggested by Sidik and Jonkman, require the use of iterative techniques to obtain the confidence interval bounds. Consequently, a general comparison between the various methods either requires some simplifying assumptions or the use of simulation methods. The latter option was chosen in the present paper. In fact, the discussion so far has already made use of a simplifying assumption, namely that the sampling variances of the effect size estimates are known. This is only approximately true when the within-study sample sizes are large (in this case, $\hat{\sigma}_i^2 \approx \sigma_i^2$). On the other hand, when the within-study sample sizes are small, then the error in the $\hat{\sigma}_i^2$ values cannot be simply ignored. Moreover, most effect size estimates are not exactly normally distributed, as assumed under the random-effects model (however, the approximation usually becomes more accurate as the within-study sample sizes increase). In a Monte-Carlo simulation, we can examine how well the different methods perform when these assumptions do not hold.

7.1. Design

The data were simulated in a manner analogous to the one described by Sidik and Jonkman [15] and Platt *et al.* [33], with the log odds ratio as the chosen effect size measure. First, a value of θ_i was generated from $\mathrm{N}(\mu, \tau^2)$. The size of the control and treatment groups ($n_i = n_i^{\mathrm{C}} = n_i^{\mathrm{T}}$) was then generated from $\mathrm{N}(n, n/4)$ rounded to the nearest integer. The number of cases in the control group (x_i^{C}) was simulated from a binomial ($n_i^{\mathrm{C}}, p_i^{\mathrm{C}}$) distribution, with p_i^{C} randomly chosen from a uniform distribution on the interval $(0.05, 0.65)$. The number of cases in the treatment group was obtained from a binomial ($n_i^{\mathrm{T}}, p_i^{\mathrm{T}}$) distribution, where $p_i^{\mathrm{T}} = p_i^{\mathrm{C}} \exp(\theta_i)/\{1 - p_i^{\mathrm{C}} + p_i^{\mathrm{C}} \exp(\theta_i)\}$. Repeating this process k times, we can generate k 2×2 tables, where the study specific log-odds ratios are given by the θ_i values and the overall log odds ratio by μ.

The effect size estimate in the ith study is equal to the log of the observed odds ratio, namely $Y_i = \log[\{x_i^{\mathrm{T}}/(n_i^{\mathrm{T}} - x_i^{\mathrm{T}})\}/\{x_i^{\mathrm{C}}/(n_i^{\mathrm{C}} - x_i^{\mathrm{C}})\}]$. The sampling variance of Y_i can be estimated with $\hat{\sigma}_i^2 = 1/(x_i^{\mathrm{T}} + 0.5) + 1/((n_i^{\mathrm{T}} - x_i^{\mathrm{T}}) + 0.5) + 1/(x_i^{\mathrm{C}} + 0.5) + 1/((n_i^{\mathrm{C}} - x_i^{\mathrm{C}}) + 0.5)$. Adding $1/2$ to each cell avoids complications introduced by cells with a zero count. The distribution of $Y_i|\theta_i$ is approximately normal with mean θ_i and variance $\hat{\sigma}_i^2$ [19].

The DerSimonian–Laird estimator [1] was used in conjunction with the two bootstrapping methods. The number of bootstrap iterations was set to $B = 1000$. Various adjustments to the standard percentile method (which was described earlier) can be used, most notably the

so-called BC_α method [24], which can improve the coverage probability of bootstrap intervals considerably. The accuracy of the standard percentile and the BC_α method was examined in the simulations.

7.2. Conditions

The following conditions were included in the simulations: $k = (10, 20, 30, 50, 80)$, $n = (10, 20, 40, 80, 160)$, τ^2 between 0 and 1 in steps of 0.1, and μ was set to 0.5. A total of 10 000 iterations were run for each condition. The standard error of the empirical coverage probabilities was therefore at most 0.005.

7.3. Results

Figures 1 and 2 show the coverage probabilities of the various methods as a function of τ^2 for selected values of k and n. In particular, values of k and n were chosen to illustrate the small, medium, and large sample behaviour of the methods. The horizontal dotted line at 0.95 indicates the nominal coverage probability.

The Q-profile method yielded the most accurate coverage probabilities, closely followed by the method suggested by Biggerstaff and Tweedie. In fact, the coverage probabilities of the two methods were identical when $\tau^2 = 0$, but did diverge slightly as the amount of heterogeneity increased. The difference between the two methods was most notable when k was large. Here, the Biggerstaff–Tweedie method yielded coverage probabilities that fell below the nominal level as τ^2 increased. As Figure 2 shows, the coverage probabilities of the Q-profile method also departed somewhat from the nominal level in the $k = 80$, $n = 10$ case (i.e. for large-scale meta-analyses with very small studies) when τ^2 was large. However, this is a rather unlikely condition to be encountered in practice and the alternative methods fared even worse in this case.

As expected, the coverage probabilities of the profile likelihood intervals were too high when τ^2 was equal to or close to zero. As τ^2 increased, the coverage probabilities approached the nominal level as long as $n/k > 1$. On the other hand, a meta-analysis of a large number of studies with small sample sizes yields coverage probabilities that deviate quite substantially from the nominal level. Finally, it is worthwhile to note that the REML-based profile intervals were slightly more accurate, especially for small k.

The performance of the Wald-type intervals was unsatisfactory, with close to or exactly 100 per cent coverage when the effect sizes were homogeneous and coverage probabilities substantially below the nominal level for larger values of τ^2. Only when k and n are both very large do the coverage probabilities begin to approach the nominal level.

The Sidik–Jonkman method also yielded unacceptable coverage probabilities. For values of τ^2 close to zero, this was to be expected for reasons outlined earlier. However, the coverage probabilities were still too small, often substantially so, even when τ^2 was large.

Finally, the bootstrapping methods yielded coverage probabilities that were also less than adequate. The BC_α method did help to improve the accuracy of the bootstrapping methods and these are the results shown in Figure 2, but the coverage probabilities were still usually too low when using non-parametric bootstrapping and, depending on the value of τ^2, above (for small τ^2) or below (for large τ^2) the nominal level when using parametric bootstrapping.

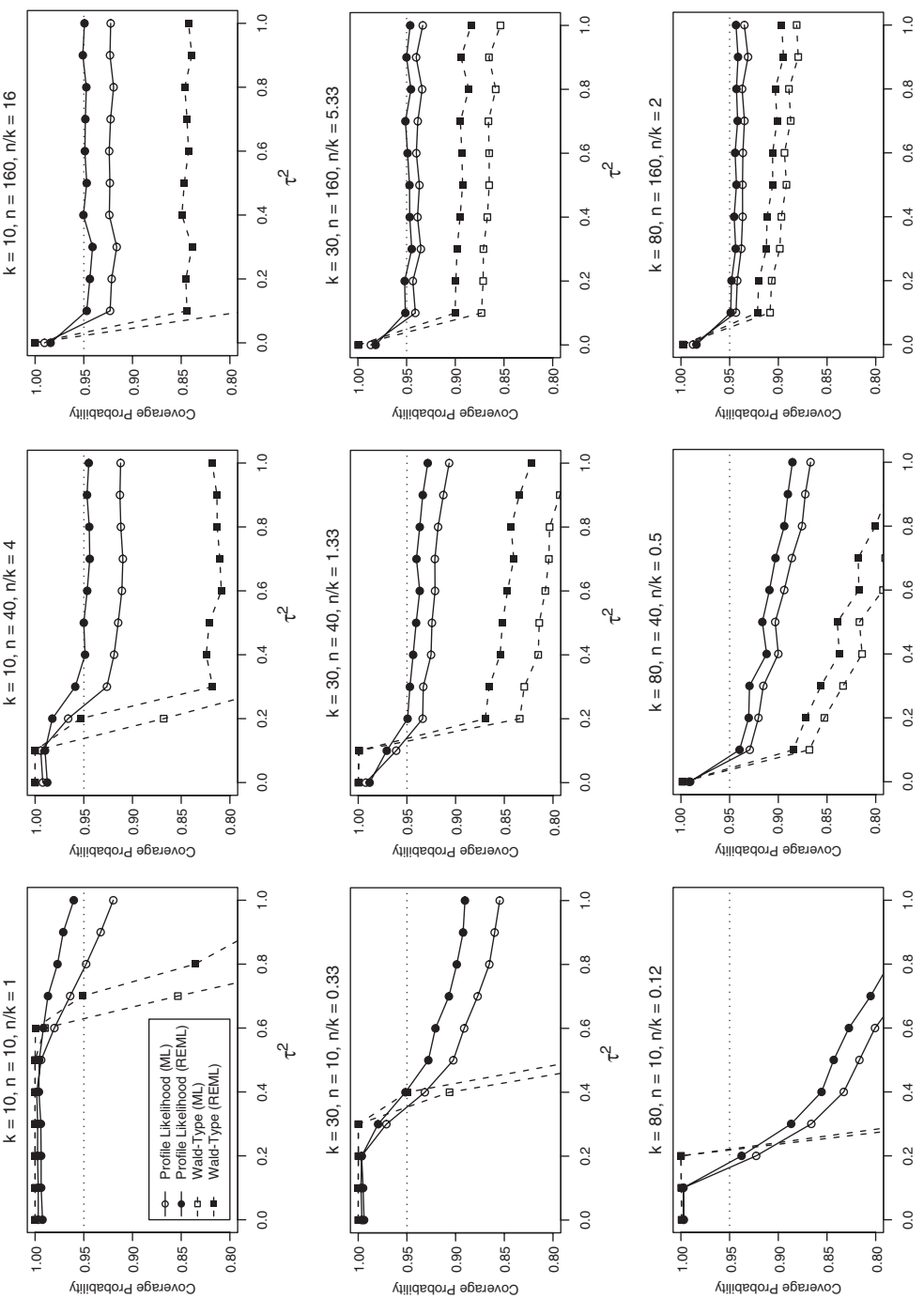

Figure 1. Coverage probabilities of the profile likelihood and Wald-type confidence intervals as a function of τ^2.

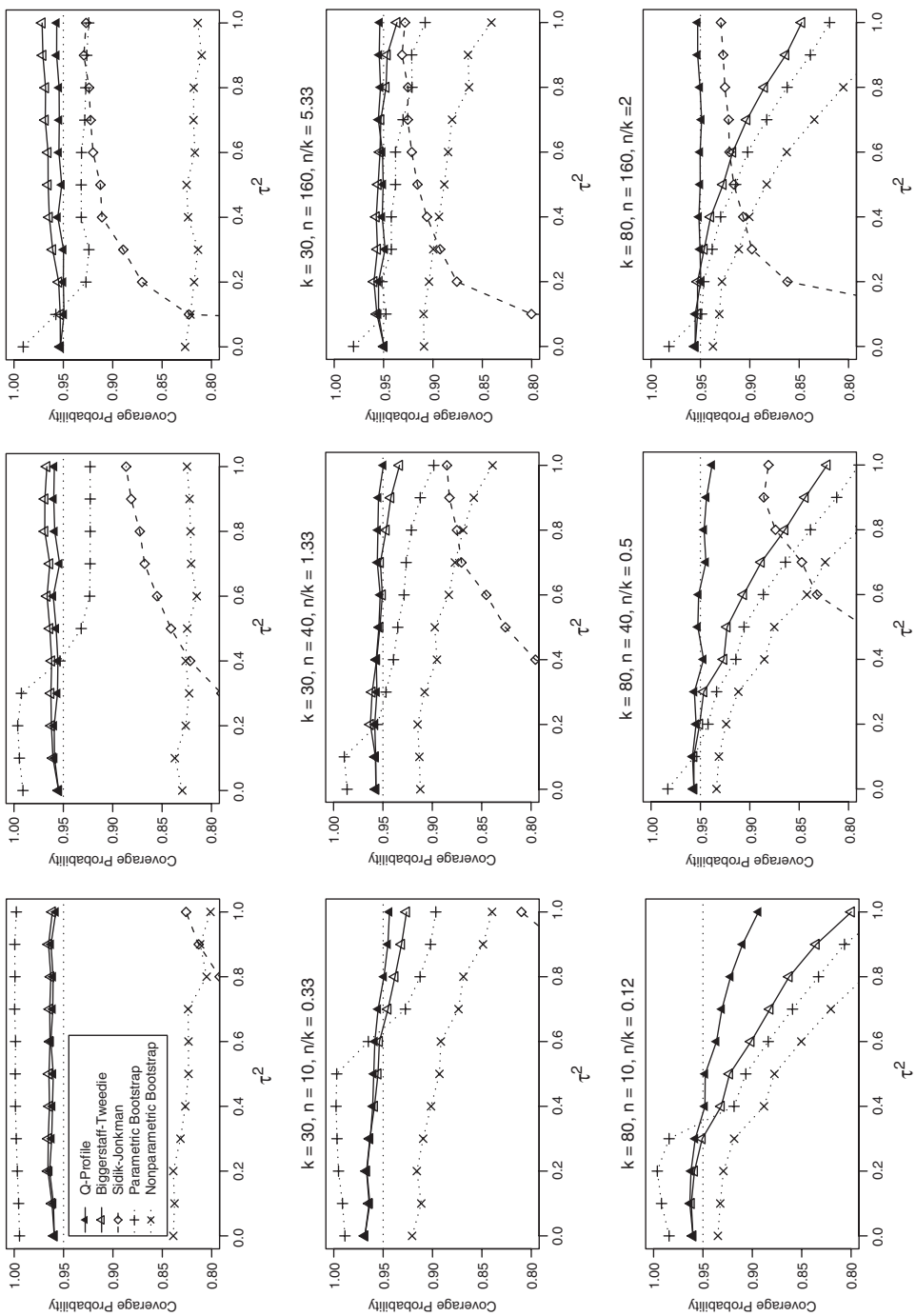

Figure 2. Coverage probabilities of the Q-profile, Biggerstaff–Tweedie, Sidik–Jonkman, and bootstrap confidence intervals as a function of τ^2.

8. CONCLUSIONS

Meta-analysts have clearly recognized that an estimate of the overall effect size should be accompanied by a confidence interval to indicate the precision with which the overall effect size has been estimated. The accuracy of various methods for calculating confidence intervals for the overall effect size has also been examined in previous research [23, 34, 35]. On the other hand, reporting of confidence intervals for the amount of heterogeneity still appears to be relatively uncommon. And although a large number of methods for obtaining confidence intervals for the amount of heterogeneity has been suggested in the literature, no systematic comparison between the various methods had been conducted so far.

The present results reveal some notable differences in the accuracy of the various methods and suggest that some methods are preferable over others. Specifically, the newly proposed Q-profile method yielded the most accurate coverage probabilities of all the methods considered. In fact, assuming that the assumptions of the random-effects model are satisfied (including the assumption of normally distributed effect size estimates and known sampling variances), the method guarantees nominal coverage levels. There are cases where tighter intervals can be obtained with the other methods (cf. Table II), but this may also come at a loss of coverage accuracy, sometimes drastically so.

An added advantage of the Q-profile method is its simplicity. The only method with a closed-form solution for the interval bounds was the one suggested by Sidik and Jonkman [15], but this method generally yielded unsatisfactory results. While all other methods require iterative procedures, the iterative scheme for the Q-profile method could even be applied with a pocket calculator. One simply has to increase τ^2 until one finds those two values of τ^2 where $Q(\tau^2)$ equals the appropriate lower and upper bounds of a χ^2 distribution with $k-1$ degrees of freedom. An R/S-Plus function to obtain Q-profile confidence intervals has also been made available at the author's website at http://www.wvbauer.com/.

Finally, it may be useful to elaborate a bit more on how a confidence interval for τ^2 may be used in a meta-analysis. First of all, a confidence interval for the amount of heterogeneity allows researchers to assess the precision of the corresponding point estimate. Reporting confidence intervals for the amount of heterogeneity would also highlight an aspect of meta-analysis that may have remained somewhat underappreciated. The width of such intervals is often quite large, indicating that τ^2 is estimated with little precision. This, in turn, underscores the need for sensitivity analyses, to show, for example, how the confidence interval for the overall effect size (μ) can change substantially in width as τ^2 changes.

For example, consider Figure 3, which shows the overall effect size estimated with (6) as a function of τ^2 for the diuretic and pre-eclampsia data in Table I. Dashed lines indicate the lower and upper bounds of the corresponding 95 per cent confidence interval for μ, typically calculated with

$$\hat{\mu} \pm 1.96\sqrt{1/\textstyle\sum w_i}$$

where $w_i = 1/(\hat{\tau}^2 + \hat{\sigma}_i^2)$. Finally, vertical dotted lines are drawn at the lower and upper bounds of the confidence interval for τ^2 as obtained with the Q-profile method ($\hat{\tau}^2 = 0.07$ and $\hat{\tau}^2 = 2.20$, respectively) and at $\hat{\tau}^2_{\mathrm{DL}} = 0.23$, the DerSimonian–Laird estimate of the amount of heterogeneity.

Several things are to be noted about this figure. First of all, as has been shown before [7], the estimate of the overall effect is relatively insensitive to changes in τ^2. On the other hand,

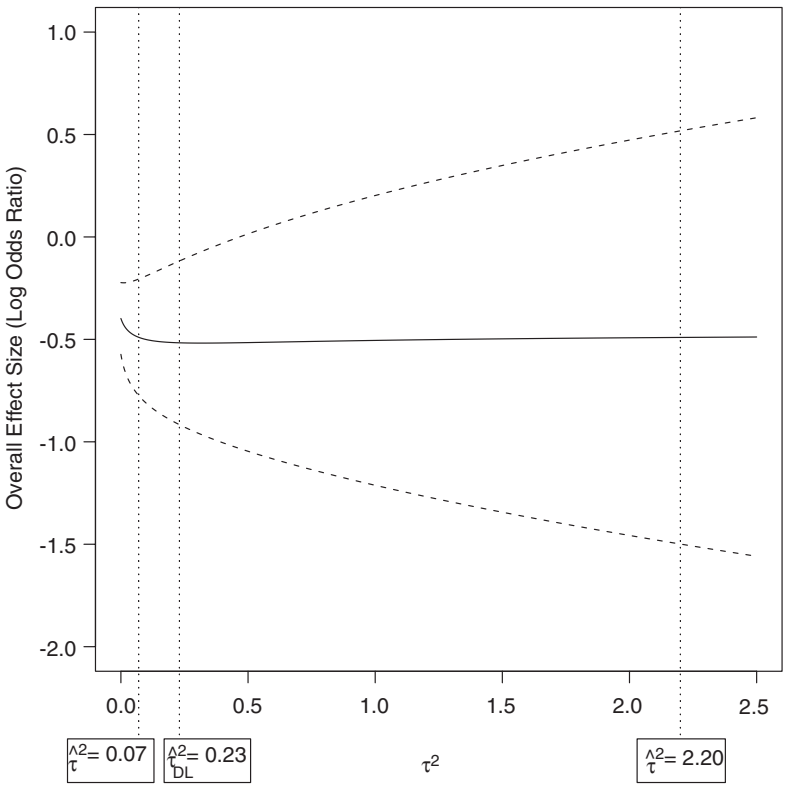

Figure 3. Overall effect size estimate (solid line) and corresponding 95 per cent confidence interval bounds (dashed lines) as a function of τ^2 for the diuretic and pre-eclampsia data (the vertical dotted lines indicate the lower and upper bounds of a 95 per cent confidence interval for τ^2 obtained with the Q-profile method and the point estimate of τ^2 obtained with the DerSimonian–Laird estimator).

the width of the confidence interval for μ increases considerably as a function of τ^2. Finally, given the imprecision in the estimate of τ^2, we may be severely overstating the precision of the estimated overall effect size. Confidence intervals for τ^2 could facilitate such sensitivity analyses by suggesting a possible range of τ^2 values one should consider.

REFERENCES

1. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**(3):177–188.
2. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. *British Medical Journal* 1994; **309**(6965):1351–1355.
3. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in Medicine* 1999; **18**(20):2693–2708.
4. Glasziou PP, Sanders SL. Investigating causes of heterogeneity in systematic reviews. *Statistics in Medicine* 2002; **21**(11):1503–1511.
5. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* 2002; **21**(11):1539–1558.
6. van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine* 2002; **21**(4):589–624.

7. Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. *Statistics in Medicine* 1996; **15**(6):619–629.
8. Viechtbauer W. Hypothesis tests for population heterogeneity in meta-analysis. *British Journal of Mathematical and Statistical Psychology* 2005, in press.
9. Lipsitz SR, Dear KBG, Laird NM, Molenberghs G. Tests for homogeneity of the risk difference when data are sparse. *Biometrics* 1998; **54**(1):148–160.
10. Takkouche B, Cadarso-Suárez C, Spiegelman D. Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis. *American Journal of Epidemiology* 1999; **150**(2):206–215.
11. Hedges LV. A random effects model for effect sizes. *Psychological Bulletin* 1983; **93**(2):388–395.
12. Raudenbush SW, Bryk AS. Empirical Bayes meta-analysis. *Journal of Educational Statistics* 1985; **10**(2):75–98.
13. Morris CN. Parametric empirical Bayes inference: theory and practice. *Journal of the American Statistical Association* 1983; **78**(381):47–55.
14. Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: a comparative study. *Statistics in Medicine* 1995; **14**(24):2685–2699.
15. Sidik K, Jonkman JN. Simple heterogeneity variance estimation for meta-analysis. *Journal of the Royal Statistical Society*, *Series C* 2005; **54**(2):367–384.
16. Biggerstaff BJ, Tweedie RL. Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in Medicine* 1997; **16**(7):753–768.
17. Switzer FS III, Paese PW, Drasgow F. Bootstrap estimates of standard errors in validity generalization. *Journal of Applied Psychology* 1992; **77**(2):123–129.
18. Turner RM, Omar RZ, Yang M, Goldstein H, Thompson SG. A multilevel model framework for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine* 2000; **19**(24):3417–3432.
19. Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. *Methods for Meta-Analysis in Medical Research*. Wiley: New York, 2000.
20. Rao CR. *Linear Statistical Inference and its Application*. McGraw-Hill: New York, 1973.
21. Casella G, Berger RL. *Statistical Inference*. Duxbury: Belmont, CA, 2001.
22. Hartung J, Knapp G. On confidence intervals for the among-group variance in the one-way random effects model with unequal error variances. *Journal of Statistical Planning and Inference* 2005; **127**(1–2):157–177.
23. Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. *Statistics in Medicine* 2001; **20**(6):825–840.
24. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman & Hall: New York, 1993.
25. Collins R, Yusuf S, Peto R. Overview of randomised trials of diuretics in pregnancy. *British Medical Journal* 1985; **290**(6461):17–23.
26. Corbeil RR, Searle SR. Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Technometrics* 1976; **18**(1):31–38.
27. Patterson HD, Thompson R. Maximum likelihood estimation of components of variance. *Proceedings of the 8th International Biometrics Conference* 1974; 197–207.
28. Viechtbauer W. Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics* 2005; **30**(3):261–293.
29. Stern SE, Welsh AH. Likelihood inference for small variance components. *Canadian Journal of Statistics* 2000; **28**(3):517–532.
30. Morrell CH. Likelihood ratio testing of variance components in the linear mixed-effects model using restricted maximum likelihood. *Biometrics* 1998; **54**(4):1560–1568.
31. Self SG, Liang KY. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* 1987; **82**(398):605–610.
32. Stram DO, Lee JW. Variance components testing in the longitudinal mixed effects model. *Biometrics* 1994; **50**(4):1171–1177.
33. Platt RW, Leroux BG, Breslow N. Generalized linear mixed models for meta-analysis. *Statistics in Medicine* 1999; **18**(6):643–654.
34. Sidik K, Jonkman JN. A simple confidence interval for meta-analysis. *Statistics in Medicine* 2002; **21**(21): 3153–3159.
35. Sidik K, Jonkman JN. On constructing confidence intervals for a standardized mean difference in meta-analysis. *Communications in Statistics*, *Simulation and Computation* 2003; **32**(4):1191–1203.

