Confidence intervals for the between-study variance in random effects meta-analysis using generalised Cochran heterogeneity statistics

Dan Jackson*†

1. Introduction

Meta-analysis is the statistical process of pooling the results from separate studies that investigate the same treatment or issue. One major difficulty when attempting this in practice however is that the study results may be too disparate to assume that they measure a common underlying effect. Indeed, some between-study variation may be thought to be inevitable due to different study practices and population mixes. The random effects model, described in detail in Section 2, is commonly used to model this between-study heterogeneity and estimate a meaningful average effect. Here, an unobserved random effect provides any necessary additional variation in the studies’ results.

Although the average effect is the parameter of primary interest, quantifying the extent of the between-study variation is also important (Higgins and Thompson, 2002; Higgins et al., 2009). In particular, if this variability appears to be considerable, then investigating the reasons for this is often encouraged (Thompson, 1994). However, the uncertainty in the estimate of the between-study variance is often large (Biggerstaff and Jackson, 2008) and a measure of this uncertainty is an important aid to inference. Here, a procedure for calculating confidence intervals is developed for this purpose.

More specifically, generalised Cochran heterogeneity statistics (DerSimonian and Kacker, 2007) are used to provide confidence intervals in a similar way to Biggerstaff and Jackson (2008), who used the more conventional heterogeneity statistic. Here, it is also proved that the calculations result in ‘well-behaved’ confidence intervals with the correct coverage. Methods based on these heterogeneity statistics are however not the most efficient (Jackson et al., 2010). Likelihood-based methods are asymptotically efficient, but these present problems. For example, inferences from the asymptotic theory of maximum likelihood (Hardy and Thompson, 1996; Biggerstaff and Tweedie 1997) cannot be expected to be accurate in situations where there are a small number of studies, as
is commonly the case. Small sample correction methods, such as those suggested for the average effect (Noma, 2011; Guolo, 2012), could be usefully employed to provide more accurate inference, but finding accurate likelihood-based methods for variance components with very few observations is inevitably challenging. Furthermore, inferences from Bayesian analyses are sensitive to the prior specification (Higgins et al., 2009; Lambert et al., 2005; Pullenayegum, 2011). In situations where an informative prior distribution for the between-study variance is available, then this could be used, but the resulting analysis requires more assumptions than conventional meta-analyses. A Bayesian analysis using informative prior distributions will also be open to immediate criticism by those who do not find the priors plausible.

Less efficient, but exact and easily computed, methods with good frequentist properties are therefore a desirable alternative. The Q profile method has recently been developed for this purpose (Knapp et al., 2006; Viechtbauer 2007), where Cochran type heterogeneity statistics are used with weights related to the total (within plus the unknown between) study variances. This gives rise to a pivotal test statistic that follows an \( \chi^2 \) distribution that may be inverted to give confidence sets. A similar approach was also suggested by Tian (2008) for normally distributed data, and Bowden et al. (2011) present some closely related ideas. Here, however, an alternative procedure is developed where the weights are instead fixed constants. As demonstrated in a simulation study in Section 4, this can provide shorter exact confidence intervals than the alternatives. The methods applied in this paper provide exact confidence intervals under the assumptions of the random effects model for meta-analysis, but this exactness requires normally distributed study outcomes with within-study variances that are treated as fixed and known. Hence, the lengths of the confidence intervals will be used to determine which method provides the most informative inference and so makes best use of the data.

The rest of the paper is set out as follows. In Section 2, the random effects model for meta-analysis and Cochran’s conventional heterogeneity statistic are described. In Section 3, DerSimonian and Kacker’s generalised heterogeneity statistic is presented and all the properties needed to ensure that these can be used to provide confidence intervals for the between-study variance are derived. In Section 4, a simulation study is performed that compares the proposed method with the Q profile method. Three variations of the methods are applied to some example datasets in Section 5, and the paper concludes in with a short discussion in Section 6.

2. The univariate random effects model and Cochran’s heterogeneity statistic

The random effects model assumes that the outcome from each of \( n \) studies, \( y_i \) for \( i = 1, 2, \ldots, n \), may be modelled using the equation

\[
Y_i = \mu + u_i + e_i
\]

where \( u_i \sim N(0, \tau^2) \), \( e_i \sim N(0, \sigma^2_i) \) and all \( u_i \) and \( e_i \) are mutually independent. It is conventional in the meta-analysis setting to assume that the within-study variances \( \sigma^2_i \) are fixed and known, but they are estimated in practice. This convention is followed but methodology, which reflects the fact that the \( \sigma^2_i \) are estimated, has also been developed (Malzahn et al., 2000). The random variable \( u_i \) denotes the random, study-specific deviation from the mean effect, and the parameter \( \tau^2 \) represents the between-study variance: \( \tau^2 > 0 \) reflects underlying study differences in a formal sense, and if \( \tau^2 = 0 \), then all studies have the same underlying average effect \( \mu \), providing a fixed effect model.

Cochran’s heterogeneity or Q statistic is frequently used in conjunction with this model and is conventionally written as the weighted sum of squares

\[
Q = \sum_{i=1}^{n} w_i(Y_i - \bar{\mu})^2
\]

where \( w_i = 1/\sigma^2_i \) and \( \bar{\mu} = \left( \sum_i w_i Y_i / \left( \sum_i w_i \right) \right) \). Commonly used inferential (DerSimonian and Laird, 1986; Biggerstaff and Jackson, 2008) and descriptive (Higgins and Thompson, 2002) statistics use Cochran’s heterogeneity statistic, but inference may also be performed without it (Jackson et al., 2010; Hardy and Thompson, 1996).

3. Generalised Cochran heterogeneity statistics and confidence intervals for \( \tau^2 \)

DerSimonian and Kacker (2007) present a generalised form of Cochran’s heterogeneity statistic. They denote the weights by \( a_i \) and this same notation will be adopted here. Their generalised heterogeneity statistic takes the same form as (1), but \( w_i = \sigma^2_i^{-2} \) is replaced by \( a_i \), where \( a_i \) may be any positive constant. It is however important to note that the \( a_i \) are fixed constants, chosen by the analyst prior to examining the data or performing the analysis, and that these values are not estimated. However, a fixed function of the within-study variances, \( a_i = f(\sigma^2_i) \), may be chosen prior to examining the data and performing the analysis. This is because the random
effects model treats the $\sigma^2_i$ as fixed constants, rather than random quantities, and hence, any fixed function of the $\sigma^2_i$ is also treated as a constant by this model. The conventional choice for this function is the reciprocal function. Estimation of $\tau^2$ and then inference $\mu$ may be performed as described by DerSimonian and Kacker.

3.1. A generalisation of Biggerstaff and Jackson’s result

In this section, a generalisation of the result from Biggerstaff and Jackson (2008) is proved. If $a_i = w_i$ for all $i$, then the generalised heterogeneity statistic reduces to Cochran’s heterogeneity statistic, and Biggerstaff and Jackson’s result is recovered as a special case.

Write DerSimonian and Kacker’s generalised heterogeneity statistic in matrix form. Let $Y$ be the vector containing the $Y_i$ and let $B = A - \frac{1}{2}aa^T$, where $A$ is the diagonal matrix containing the $a_i$, $a$ is the vector containing the $a_i$, $a_i = \sum_i a_i$ and $t$ denotes matrix transpose. The matrix representation of DerSimonian and Kacker’s generalised heterogeneity statistic, $Q_a$, is then given by

$$Q_a = Y'BY$$

Next, let $\Sigma$ denote the variance of $Y$, the diagonal matrix with entries $(\sigma^2_i + \tau^2)$, and let $Z$ denote a standard $n$ dimensional multivariate normal vector. Noting that $Q_a$ is location invariant, we can write

$$Q_a = Y'BY \triangleq Z'\Sigma^{1/2}B\Sigma^{1/2}Z = Z'SZ$$

defining $S = \Sigma^{1/2}B\Sigma^{1/2}$. $B$ is symmetric and hence, so is $S$. Following the same procedure described by Biggerstaff and Jackson (2008), writing $S$ in terms of its spectral decomposition, we obtain

$$Q_a = \sum_{i=1}^{n} \lambda_i(S) \chi^2(1)$$

(2)

where $\chi^2(1)$ are mutually independent chi-squared random variables with 1 degree of freedom and $\lambda_1(S) \geq \lambda_2(S) \geq \cdots \geq \lambda_n(S)$ are the ordered eigenvalues of $S$.

3.2. The parameters that the distribution of $Q_a$ depend on

The $\lambda_i(S)$ are functions of $\tau^2$ through their dependence on $S$ but do not depend on $\mu$. The eigenvalues $\lambda_i(S)$ also depend upon the $\sigma^2_i$ and $a_i$ but these are taken to be fixed constants: the $\sigma^2_i$ are treated as fixed constants in the random effects model and the $a_i$ are fixed values, possibly a fixed function of the $\sigma^2_i$, chosen by the analyst prior to analysis. Hence, the only unknown parameter that the distribution of $Q_a$ depends on is $\tau^2$. Next, some other important properties possessed by $Q_a$ are proven.

3.3. Property 1: $Q_a$ is distributed as a positive linear combination of $\chi^2(1)$ random variables with exactly $(n-1)$ positive coefficients

Biggerstaff and Jackson (2008) proved that the conventional $Q$ statistic (1) is distributed as a positive linear combination of $\chi^2(1)$ random variables with at most $(n-1)$ positive coefficients. The Appendix contains a proof that this is also the case for the generalised heterogeneity statistic $Q_a$ where it is further shown that this statistic is distributed as such a linear combination with exactly $(n-1)$ positive coefficients.

3.4. Property 2: The cumulative distribution function of $Q_a$ is a continuous and strictly decreasing function in $\tau^2$

Biggerstaff and Jackson (2008) implicitly assumed this property when obtaining confidence intervals using the conventional $Q$ statistic.

Proof: The entries of $S$ are continuous in $\tau^2$ and hence, so are its eigenvalues. Hence, from the form of (2), the cumulative distribution function of $Q_a$ is continuous in $\tau^2$ and all that remains to be shown is that it is strictly decreasing in $\tau^2$.

The eigenvalues in (2) are those of $\Sigma^{1/2}B\Sigma^{1/2}$ where $\Sigma$ depends on $\tau^2$ and $B$ does not. Suppose that we also consider a larger value of $\tau^2$, denoting the larger variance of $Y$ as $2M$, where $M$ is a diagonal matrix where all the diagonal entries are greater than one. Then $\lambda_i(M) > 1$ for all $i$. The eigenvalues in (2) for the larger value of $\tau^2$ are those of $\Sigma^{1/2}M^{1/2}B\Sigma^{1/2}M^{1/2}$, which are equal to those of $SM$. Then the second inequality in Equation (5) in the Appendix, with $C = S$ and $D = M$, immediately shows that all the eigenvalues that provide coefficients in (2) are greater when considering the larger $\tau^2$ and so are strictly increasing in $\tau^2$. Hence, the cumulative distribution function of $Q_a$ is strictly decreasing in $\tau^2$. 

© 2013 The Authors. Research Synthesis Methods published by John Wiley & Sons, Ltd. Res. Syn. Meth. 2013, 4 220–229
3.5. Confidence intervals for $\tau^2$ by test inversion

Now that the necessary properties of $Q_a$ have been established, the test inversion procedure described by Casella and Berger (2002), Section 9.2.1, will be used to construct confidence sets with coverage probability $1 - x_1 - x_2$, where $x_1 + x_2$ denotes the significance level associated with the two-tailed test. Our proposed method essentially follows the method proposed by Biggerstaff and Jackson (2008) when using $Q_a$ with the conventional weights $a_i = w_i$.

We accept the null hypothesis $H_0 : \tau^2 = \tau_0^2$ and thus, $\tau_0^2$ lies in the corresponding confidence set, if and only if

$$P(Q_a \geq q_a; \tau^2 = \tau_0^2) \geq x_1$$

(3)

and

$$P(Q_a \leq q_a; \tau^2 = \tau_0^2) \geq x_2$$

(4)

Conceptually, (3) ensures that $\tau_0^2$ is not too small to result in the observed $q_a$ and so provides a lower bound; (4) provides an upper bound. We reject the null hypothesis $H_0 : \tau^2 = \tau_0^2$ using (3) if $q_a$ is greater than the $(1 - x_1)$ quantile of $Q_a$. Similarly, we reject the null hypothesis using (4) if $q_a$ is less than the $x_2$ quantile. Because $x_1 + x_2 < 1$, it is impossible for (3) and (4) to simultaneously reject the null hypothesis. Hence, the significance level of the test is $x_1 + x_2$, as required when producing a confidence set with coverage probability $1 - x_1 - x_2$.

3.6. Obtaining confidence intervals numerically

The cumulative distribution function of $Q_a$, $P(Q_a \leq q_a; \tau^2)$, can be evaluated using the algorithm for a positive linear combination of $\chi^2$ random variables proposed by Farebrother (1984). The CompQuadForm R package implementation of this was used throughout.

If $P(Q_a \leq q_a; \tau^2 = 0) < x_2$, then no $\tau_0^2$ satisfies (4); $\tau^2$ is nonnegative and the cumulative distribution function of $Q_a$ is decreasing in $\tau^2$. Hence, the strict implementation of (3) and (4) provides a null set for the confidence set for $\tau^2$. This is analogous to the possible null confidence set that the Q profile method proposed by Viechtbauer (2007) may also provide. However, this only occurs when the observed $q_a$ is very small, so instead of providing a null set for the confidence set in such instances, it is preferable to give $[\tau_1^2, \tau_2^2] = \{0, 0\}$. If $\tau^2 > 0$, as typically thought to be the case, then this does not increase the coverage probability of the confidence set. If $\tau^2 = 0$, then the coverage probability of the confidence set increases by $x_2$ and the confidence set is conservative. This convention can also be adopted when using the Q profile method, and this was suggested by Knapp et al. (2006). Alternative interpretations of the null set are possible in application however, such as ‘the data appear to be highly homogenous’ or even ‘the interval estimation fails’.

If instead $P(Q_a \leq q_a; \tau^2 = 0) \geq x_2$, then, because the cumulative distribution function $P(Q_a \leq q_a; \tau^2)$ is continuous and strictly decreasing in $\tau^2$, we can use any simple numerical method to find the value $\tau_0^2$ that satisfies $P(Q_a \leq q_a; \tau^2 = \tau_0^2) = x_2$. Then all $\tau^2$ in the interval $[0, \tau_0^2]$ satisfies Equation (4) and all other values of $\tau^2$ do not.

Next, if $P(Q_a \geq q_a; \tau^2 = 0) \geq x_1$, then all $\tau^2$ in the interval $[0, \infty]$ satisfy (3) and we define $\tau_0^2 = 0$; Biggerstaff and Jackson (2008) refer to this as ‘truncating’ the lower confidence bound to zero. Otherwise, we can use any simple numerical method to find the value of $\tau^2$ that satisfies $P(Q_a \geq q_a; \tau^2 = \tau_0^2) = x_1$ and then all $\tau^2$ in the interval $[\tau_1^2, \infty]$ satisfies (3) and all other values do not. The intersection of the intervals $[\tau_1^2, \infty]$ and $[0, \tau_0^2]$ provide the confidence set for $\tau^2$.

It is easily shown that the resulting confidence set is the interval $[\tau_1^2, \tau_2^2]$, that is $\tau_2^2 \geq \tau_1^2$. First, note that if the null confidence set has been interpreted as $[0, 0]$, then trivially, $\tau_1^2 \geq \tau_2^2$. Otherwise $\tau_1^2 > 0$ and if the lower confidence bound has been truncated to zero then, again trivially, $\tau_2^2 > \tau_1^2$. Finally, if the lower bound is not truncated, when solving $P(Q_a \geq q_a; \tau^2 = \tau_1^2) = x_1$, and noting that $x_1 + x_2 < 1$, we have that

$$P(Q_a \leq q_a; \tau^2 = \tau_1^2) = 1 - x_1 > x_2 = P(Q_a \leq q_a; \tau^2 = \tau_0^2)$$

Because the cumulative distribution function of $Q_a$ is strictly decreasing in $\tau^2$, we must have $\tau_1^2 < \tau_0^2$. Hence, $\tau_2^2 \geq \tau_1^2$.

4. A simulation study

To assess the performance of the proposed method, a simulation study was performed with $n = 5$ studies. This represents the common situation where there is just a handful of studies and the proposed method may be anticipated to be especially valuable. If instead there were just two or three studies, any form of estimation for the random effects model would be extremely challenging, and if there were many more studies, then the sample
size would be sufficient, for example, to make accurate inferences using the profile likelihood in the way proposed by Hardy and Thompson (1996).

Within-study variances of $\sigma^2 = 0.009$, $\sigma^2 = 0.046$, $\sigma^2 = 0.122$, $\sigma^2 = 0.265$ and $\sigma^2 = 0.600$ were used. These values correspond to the 0%, 25%, 50%, 75% and 100% quantiles of the scaled and truncated $\chi^2$ distribution used by Brockwell and Gordon (2007) for generating within-study variances. Five values of $\tau^2 = 0, 0.029, 0.069, 0.266, 1.302$ were used, because these correspond to $I^2 = 0, 0.30, 0.50, 0.75, 0.95$ (Higgins and Thompson, 2002) and so cover a wide range of possibilities.

Weights of the form $a_i = 1/(\sigma_i^2 + x)$ were investigated when applying the proposed method, where $x$ took the five values of $\tau^2$ described previously, in conjunction with $p = 0.05, 1$. Provided the analyst specifies the fixed values of $x$ and $p$ to be used prior to examining the data or performing the analysis, the resulting weights are fixed constants because the $\sigma_i^2$ are treated as such in the random effects model. The repeated sampling properties of confidence intervals under the assumptions of the random effects model, where $x$ and $p$ depend on any way on an examination of the data, or any other random variables, are much harder to evaluate but one possibility is explored in Section 4.2. Weights of the form $a_i = 1/(\sigma_i^2 + x)^{p}$ could also be used to make inferences for $\mu$ in the way described by DerSimonian and Kacker (2007), but here, the focus is the proposed method for constructing confidence intervals for $\tau^2$.

These three values of $p$ provide an unweighted analysis, and also weights that are related to the within-study standard errors and the corresponding variances. Hence, these values of $p$ are intuitively appealing values to consider; $x = 0$ and $p = 1$ correspond to the conventional weights $w_i = \sigma_i^{-1}$. If $p = 0$, then an unweighted heterogeneity statistic is used irrespective of $x$; unweighted methods for meta-analysis have previously been proposed (Bonett, 2008; Bonett 2009; Shuster, 2010), and completely arbitrarily $x = 0$ was used in conjunction with $p = 0$. Setting $x = 0$ and $p = 0.5$ means that the reciprocals of the studies’ within-study standard errors are used as weights. If, for example, an a priori value of $\tau^2$ is thought plausible, then setting $x$ equal to this means that the weights are related to the total study variances thought plausible before examining the data, which are intuitively appealing weights to use. In situations where a suitable positive value of $x$ is difficult or impossible to state in advance, then this should be set to zero to use weights that are more akin to the conventional ones.

Confidence intervals with $z_{1} = z_{2} = 0.025$, and hence, 95% coverage probability, were used throughout. Putting an equal probability of 0.025 in each tail follows common practice, but other possibilities are returned to in the discussion.

To compare the proposed procedure to the established Q profile method, the R package metafor was used to apply Viechtbauer’s implementation of this. This method was chosen because it has become popular and, like the proposed method, only requires that the random effects model is assumed for the outcome data used in analysis. For example, neither the proposed method nor the Q profile method requires that the raw data follow a normal distribution.

The results of the simulation study are shown in Table 1, where the mean and the standard deviation of the lengths of the resulting confidence intervals are shown. Here, the proposed method with each set of weights, and the Q profile method, was applied to the same 40,000 simulated datasets for each $\tau^2$; $\mu = 0$ was used when simulating data, but this is immaterial. A different random seed was used for each value of $\tau^2$. The convention where null confidence sets were interpreted as intervals of [0,0] was adopted throughout. The empirical coverage probabilities of the 95% confidence intervals are also shown in Table 1 and are within Monte Carlo error of 0.95 if $\tau^2 > 0$ and 0.975 if $\tau^2 = 0$, as the theory predicts. One conclusion from Table 1 is that the confidence intervals are in general very wide, reflecting the considerable uncertainty in $\tau^2$ for examples with just five studies. The results for the best performing method (shortest average confidence interval) for each value of $\tau^2$ are highlighted in bold in Table 1.

The results in Table 1 show that as $\tau^2$ increases, the average lengths of all confidence intervals, and the variation in these lengths, also increase. In terms of the previously proposed methods, the Q profile method provides shorter confidence intervals when there is considerable heterogeneity than intervals based on Cochran’s conventional heterogeneity statistic (B and J in Table 1). However, this requires $I^2 > 0.5$ ($\tau^2 > 0.069$), and otherwise, Biggerstaff and Jackson’s method is preferable.

Intuition suggests that using $p = 1$ in conjunction with a value of $x$ that is appropriate in the context of the meta-analysis in question will ensure that the weights used will most accurately reflect the true variance structure in the data and hence provide shorter confidence intervals. This intuition is confirmed in Table 1, where the method that provides the shortest confidence intervals is in every case the one that uses the inverse of the true total study variances ($p = 0$ and $x = \tau^2$). Cochran’s heterogeneity statistic can be seen as quite an extreme case, where the weights incorporate no between-study variance whatsoever. In practice, using values of $x$ that are thought a priori close to the true value can be expected to perform better than sticking to the conventional weights. Those more comfortable with the $I^2$ statistic could specify a plausible value and convert this to a value of $x$ to use in the weights. However, this suggestion requires some a priori knowledge about the likely extent of the between-study variation, which the author does not possess for the examples that follow, and so may be difficult to implement in practice.

The results for $p = 1/2$ are interesting because, for example, $p = 1/2$ and $x = 0$ outperforms the Q profile method unless the heterogeneity is very severe and only increases the length of the intervals slightly compared to…
Table 1. Results from the simulation study. 40,000 simulated datasets were produced for each value of $t^2$. The average lengths of the 95% confidence intervals are shown with their standard deviations in parentheses. The results for the procedure that provides, on average, the shortest intervals is highlighted in bold font for each value of $t^2$. 'B and J' indicates the conventional weights, and so the procedure suggested by Biggerstaff and Jackson (2008) is used. The empirical coverage probabilities of the 95% confidence intervals are also shown.

| Method            | $t^2 = 0$ |          | $t^2 = 0.029$ |          | $t^2 = 0.069$ |          | $t^2 = 0.206$ |          | $t^2 = 1.302$ |          |
|-------------------|-----------|----------|---------------|----------|---------------|----------|---------------|----------|---------------|----------|
|                   | Length    | Coverage | Length        | Coverage | Length        | Coverage | Length        | Coverage | Length        | Coverage |
| Q profile         | 1.420 (1.547) | 0.974 | 1.677 (1.670) | 0.950 | 2.011 (1.854) | 0.948 | 3.142 (2.562) | 0.950 | 11.910 (8.606) | 0.949 |
| $p = 1$, $x = 0$ (B and J) | **0.804 (0.672)** | 0.974 | 1.207 (1.008) | 0.949 | 1.759 (1.524) | 0.948 | 3.652 (3.414) | 0.950 | 18.975 (19.114) | 0.950 |
| $p = 1$, $x = 0.029$ | 0.856 (0.733) | 0.976 | **1.175 (0.938)** | 0.950 | 1.613 (1.269) | 0.948 | 3.105 (2.525) | 0.950 | 15.126 (13.264) | 0.949 |
| $p = 1$, $x = 0.069$ | 0.924 (0.813) | 0.976 | 1.209 (0.976) | 0.951 | **1.601 (1.239)** | 0.948 | 2.935 (2.248) | 0.950 | 13.624 (11.050) | 0.949 |
| $p = 1$, $x = 0.206$ | 1.073 (0.993) | 0.976 | 1.328 (1.126) | 0.950 | 1.679 (1.333) | 0.950 | **2.872 (2.129)** | 0.950 | 12.341 (9.204) | 0.948 |
| $p = 1$, $x = 1.302$ | 1.363 (1.372) | 0.975 | 1.605 (1.494) | 0.951 | 1.930 (1.674) | 0.950 | 3.054 (2.372) | 0.951 | **11.865 (8.467)** | 0.949 |
| $p = 1/2$, $x = 0$ | 0.976 (0.891) | 0.976 | 1.255 (1.039) | 0.951 | 1.636 (1.283) | 0.950 | 2.938 (2.244) | 0.949 | 13.334 (10.789) | 0.949 |
| $p = 1/2$, $x = 0.029$ | 1.074 (1.008) | 0.976 | 1.331 (1.137) | 0.950 | 1.682 (1.342) | 0.950 | 2.884 (2.144) | 0.950 | 12.407 (9.336) | 0.948 |
| $p = 1/2$, $x = 0.069$ | 1.143 (1.092) | 0.976 | 1.393 (1.216) | 0.951 | 1.734 (1.409) | 0.950 | 2.900 (2.161) | 0.950 | 12.113 (8.882) | 0.949 |
| $p = 1/2$, $x = 0.206$ | 1.261 (1.240) | 0.975 | 1.505 (1.361) | 0.951 | 1.835 (1.545) | 0.950 | 2.970 (2.255) | 0.950 | 11.897 (8.538) | 0.949 |
| $p = 1/2$, $x = 1.302$ | 1.439 (1.478) | 0.976 | 1.681 (1.600) | 0.951 | 2.004 (1.780) | 0.951 | 3.125 (2.476) | 0.951 | 11.899 (8.499) | 0.949 |
| $p = 0$ (unweighted) | 1.523 (1.595) | 0.976 | 1.766 (1.719) | 0.951 | 2.087 (1.900) | 0.951 | 3.209 (2.600) | 0.951 | 11.975 (8.593) | 0.949 |
Biggerstaff and Jackson’s method when the heterogeneity is mild. Weighting studies by the reciprocal of their within-study standard errors in this way, rather than by their variances as convention dictates, appears to provide a sensible and viable option when there is little a priory knowledge about the extent of heterogeneity, but some is anticipated. This weights the studies more equally than the usual weights and so better reflects the true variance structure when heterogeneity is present. Hence, it makes intuitive sense to consider this alternative set of weights under these circumstances. This proposal is compared with the established alternatives using some real datasets in Section 5.

4.1. Additional simulation studies

Further simulation studies were performed, again using 40,000 simulated datasets for each scenario, where different random seeds were used for each combination of n and \( \tau^2 \). Sample sizes of \( n = 10, n = 20 \) and \( n = 40 \), and the same values of \( \tau^2 \) as in Table 1, were used in these additional simulation studies. Within-study variances were obtained in the same manner as before, as equally spaced quantiles, from 0% to 100%, from the distribution suggested by Brockwell and Gordon (2007). The results from these simulation studies reinforce the previous conclusions. In each case, Biggerstaff and Jackson’s method outperformed (shorter confidence intervals) the Q profile method when the heterogeneity was mild, but when the heterogeneity was large, these roles were reversed. In every instance, \( p = 1 \) in conjunction with \( x \) equal to the true between-study variance performed very well as anticipated. Perhaps most importantly, the choice of \( p = 1/2 \) and \( x = 0 \) seemed a reasonable compromise between Biggerstaff and Jackson’s method and the Q profile method, exactly as it did for \( n = 5 \). The results from these additional simulation studies are available in the supplementary materials that accompany the paper.

4.2. Weighting by the reciprocal of the estimated total study variances

Weights of the form \( a_i = 1/(\sigma_i^2 + \tau^2) \), so that the weights are the reciprocals of the true total study variances, were found to perform well in the simulation study. However, because \( \tau^2 \) is unknown, it is not entirely fair to compare the use of these weights with the established methods. It is however tempting to use weights equal to the estimated total study variances so that \( a_i = 1/(\sigma_i^2 + \hat{\tau}^2) \). These weights are easily computed and straightforward to use in application but their use invalidates the theory that ensures that exact confidence intervals are obtained, because the weights \( a_i \) are now random variables.

Despite this theoretical objection, results were also obtained, using the simulated datasets used to produce Table 1 and weights of the form \( a_i = 1/(\sigma_i^2 + \hat{\tau}^2) \), where \( \hat{\tau}^2 \) is the usual estimate originally proposed by DerSimonian and Laird (1986). The average length of the resulting confidence intervals (with standard deviations in parentheses, as in Table 1) were 0.872 (0.862), 1.182 (1.050), 1.564 (1.306), 2.751 (2.136) and 11.541 (8.396) for \( \tau^2 = 0, 0.029, 0.069, 0.206 \) and 1.302, respectively. Furthermore, the empirical coverage probabilities of the nominal 95% confidence intervals were for these same \( \tau^2 \) were 0.978, 0.953, 0.953, 0.954 and 0.948. Because the Monte Carlo standard error associated with estimating a probability of 0.95 with a sample size of 40,000 is \( \sqrt{0.95 \times 0.05/40000} \approx 0.001 \), these results provide evidence that this procedure fails to provide the nominal coverage probability exactly but also that this departure from the nominal probability level is too small to be of any practical concern. This method also appears to perform well compared with the alternatives, in that it provides relatively short 95% confidence intervals, and in particular, it appears to provide shorter confidence intervals than using the weights proposed for general use in Section 4.1 (\( p = 1/2 \) and \( x = 0 \); seventh row of Table 1). These conclusions are supported by the results using \( n = 10, 20 \) and 40; all the results for these larger sample sizes are available in the supplementary materials that accompany this paper.

Further investigation is required before this procedure can be safely recommended for general use, because the random weights invalidate the theory, but this simulation study suggests that using the reciprocal of the estimated total study variances as weights, which are the conventional weights for making inferences about \( \mu \) (DerSimonian and Laird, 1986; Jackson et al., 2010) may also prove to be a good option for obtaining confidence intervals for \( \tau^2 \).

5. Examples

Four examples were analysed by Biggerstaff and Jackson (2008), and these will also be used here. See Biggerstaff and Jackson for full details, but briefly, the examples use meta-analytic data from studies that examine: (i) aspirin and heart attack; (ii) diuretic and preeclampsia; (iii) glycerol for acute stroke; and (iv) sclerotherapy and cirrhosis. Three methods for obtaining exact confidence intervals will be applied to each example. First, the established Q profile method (first row of Table 1) and the method proposed by Biggerstaff and Jackson using Cochran’s Q statistic (second row of Table 1) will be used. Finally, generalised Cochran heterogeneity statistics will be used, where the weights are the reciprocals of the within-study standard deviations (seventh row of Table 1). This follows the suggestion in Section 4, where some heterogeneity is anticipated but it is uncertain how much.
Now that we have real data, however, the exactness of the confidence intervals is brought into question, because the random effects model only provides an approximation for data such as these.

The results are shown in Table 2. In three of the four examples, the proposed method (weighting by the reciprocal of the within-study standard errors) provides shorter confidence intervals than the method used by Biggerstaff and Jackson. This reflects the fact that statistical heterogeneity is present in all four datasets (Biggerstaff and Jackson, 2008). However, in three of the four examples, the Q profile method provides the shortest confidence interval. This observation may appear to contradict the finding from the simulation study that this method is generally outperformed by the proposed method. However, this is belied by the fact that the Q profile method provides a very much longer confidence interval for the Diuretic data, which gives an indication of the longer confidence intervals that this method has been found to provide.

The confidence intervals are generally in good agreement for all four examples and are wide, as anticipated, given the difficulty in accurately estimating the between-study variance in examples where there are few studies. The reasons for the more specific differences between the results using the three methods are hard to explain however, which is perhaps inevitable given the considerable heterogeneity present in these examples and the imprecision of estimates of \( \tau^2 \) in meta-analyses such as these.

### 6. Discussion

Generalised Cochran heterogeneity statistics provide a convenient method for computing exact confidence intervals for the between-study variance parameter in a random effects meta-analysis. They incorporate an existing method as a special case, and by choosing more appropriate weights than is conventional, shorter confidence intervals may be obtained. The only potential numerical difficulty is the use of Farebrother’s algorithm, but the R implementation is convenient and fast. R code is available in the supplementary materials that accompany the paper, which produces confidence intervals in a few seconds.

In the simulation study, \( \tau_1 = \tau_2 = 0.025 \) was used, so that the convention of using equal probabilities in both tails has been adopted, but shorter confidence intervals with the same coverage might be possible by using \( \tau_2 \neq \tau_1 \) and it is left as an open question as to whether this should be considered more often in practice. However, those who do not consider \( \tau^2 = 0 \) plausible may prefer to calculate one-sided confidence intervals that reflect this and set \( \tau_1 = 0 \). Furthermore, due to the very wide confidence intervals that were obtained in the simulation study and for the examples, coverage intervals with less coverage, 90% or even 80%, might be deemed preferable to obtain tighter confidence intervals.

A wide variety of methods for estimating \( \tau^2 \) are available (Sidik and Jonkman, 2007) and the confidence intervals developed here are built upon just one of these, the method proposed by DerSimonian and Kacker. Methods might also be developed that build upon the alternatives however and this could form the subject of future work. Methods for constructing confidence intervals for \( I^2 \) follow naturally by taking the typical within-study variance proposed by Higgins and Thompson (2002) as fixed, so that \( I^2 \) may be interpreted as a monotonic function of \( \tau^2 \). Generalised Cochran heterogeneity statistics could also be used to test the null hypothesis that \( \tau^2 = 0 \), in a similar way to the conventional one, but this possibility is not examined here. This is because using them is more complicated and no gain in power is anticipated, but this too could provide an avenue for further work.

The Q profile method remains a viable alternative but this appears to be inefficient compared with the alternatives considered here unless the heterogeneity is very considerable, in which case, all confidence intervals become very wide and so very little can be inferred about the extent of the heterogeneity. In any case, many would hesitate to combine very disparate results in a random effects meta-analysis. If some heterogeneity is thought plausible but is difficult to determine how much there might be \( a priori \), then using generalised Cochran heterogeneity statistics with weights equal to the reciprocal of the within-study standard errors would appear to be a sensible option. The possibility of using weights equal to the reciprocal of the estimated total study variances warrants further investigation.

### Table 2. Results for the four examples using three exact methods for obtaining confidence intervals for the between-study variance. For each method, the 95% confidence interval is tabulated using \( \tau_1 = \tau_2 = 0.025 \), and the width of the interval is given in square brackets.

| Dataset     | n | Q profile               | B and J                  | Proposed method         |
|-------------|---|-------------------------|--------------------------|-------------------------|
| Aspirin     | 6 | (0.0171) [0.171]        | (0.0339) [0.339]         | (0.0235) [0.235]        |
| Diuretic    | 9 | (0.072, 2.203) [2.130]  | (0.047, 1.431) [1.384]   | (0.074, 1.678) [1.603]  |
| Glycerol    | 9 | (0.0947) [0.947]        | (0.1124) [1.124]         | (0.1012) [1.012]        |
| Sclerotherapy | 19 | (0.047, 0.816) [0.769] | (0.071, 1.023) [0.953]  | (0.0867) [0.867]        |

© 2013 The Authors. Research Synthesis Methods published by John Wiley & Sons, Ltd. Res. Syn. Meth. 2013, 4 220–229
Appendix:

If C and D are square matrices of the same size, then $\lambda_i(CD) = \lambda_i(DC)$ (Zhang 2010; page 57). Because the rows of B sum to zero, this matrix has an eigenvalue of zero and hence so does S. This can most easily be seen by observing that $\lambda_i(S) = \lambda_i(2B)$, and the premultiplication of B by $\Sigma$ retains the linear dependency amongst the rows of B and hence the eigenvalue of zero. Thus, $\lambda_i(S) = 0$, and the sum in Equation (2) only extends to $(n-1)$.

It can also be seen that $\lambda_i(S) > 0$ for $i < n$ in Equation (2). This is a consequence of the observation that $\lambda_i(A^{-1/2}BA^{-1/2}) = \lambda_i(BA^{-1}) = 1$ for $i < n$; $A^{-1/2}BA^{-1/2} = S$, where $a_i^2 = a_i^{-1}$ and $\tau^2 = 0$. It is a standard result that $Q_a \sim \chi^2(n-1)$ when these standard weights are used and $\tau^2 = 0$ (Biggerstaff and Jackson, 2008). Because (2) shows that $Q_a$ is a linear combination of $\chi^2(1)$ random variables, an inspection of the form of the moment generating function of $\chi^2$ confirms that $\lambda_i(BA^{-1}) = 1$ for $i < n$. If C and D are positive semi-definite Hermitian matrices, then

$$\lambda_i(C)\lambda_i(D) \geq \lambda_i(C)\lambda_i(D)$$

(5)

where $\lambda_i(D)$ and $\lambda_i(C)$ are the largest and smallest eigenvalues of D respectively (Zhang, 2010; page 274). Then, the first inequality in (5) with $C = B$ and $D = A^{-1}$, and noting that $\lambda_i(A^{-1}) > 0$ for all $i$, shows that $\lambda_i(B) > 0$ if $i < n$. Finally, the second inequality in (5) with $C = B$ and $D = \Sigma$, and noting that $\lambda_i(\Sigma) > 0$ for all $i$, shows that $\lambda_i(B) > 0$ for $i < n$.

Hence, $Q_a$ is a linear combination of $\chi^2(1)$ random variables with exactly $(n-1)$ positive coefficients as stated.

Acknowledgement

Brad Biggerstaff is thanked for his comments on a previous version of the paper. DJ is employed by the UK Medical Research Council [Unit Programme number U105260558].

References

Biggerstaff BJ, Jackson, D. 2008. The exact distribution of Cochran’s heterogeneity statistic in one-way random effects meta-analysis. Stat Med 27: 6093–6110.

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

Bonett, DG. 2008. Meta-analytic interval estimation for bivariate correlations. Psychol Meth 13: 173–181.

Bonett DG. 2009. Meta-analytic interval estimation for standardized and unstandardized mean differences. Psychol Meth 14: 225–238.

Bowden J, Tierney, JF, Copas AJ, Burdett S. 2011. Quantifying, displaying and accounting for heterogeneity in the meta-analysis of RCTs using standard and generalised Q statistics BMC Med Res Methodol 11: 41.

Brockwell SE, Gordon IR. 2007. A simple method for inference on an overall effect in meta-analysis. Stat Med 26: 4531–4543.

Casella, G, Berger RL. 2002. Statistical Inference (2nd edn). Duxbury, North Scituate, MA.

DerSimonian R, Kacker R. 2007. Random-effects model for meta-analysis of clinical trials: an update. Contemp Clin Trials 28: 105–114.

DerSimonian R, Laird N. 1986. Meta-analysis in clinical trials. Control Clin Trials 7: 177–188.

Farebrother, RW. 1984. Algorithm AS 204: the distribution of a positive linear combination of $\chi^2$ random variables. J Appl Stat 33: 332–339.

Guolo A. 2012. Higher-order likelihood inference in meta-analysis and meta-regression. Stat Med 31: 313–327.

Hardy RJ, Thompson SG. 1996. A likelihood approach to meta-analysis with random effects. Stat Med 15: 619–629.

Higgins JPT, Thompson SG. 2002. Quantifying heterogeneity in a meta-analysis. Stat Med 21: 1539–1558.

Higgins JPT, Thompson, SG, Spiegelhalter DJ. 2009. A re-evaluation of random-effects meta-analysis. J Roy Stat Soc A 15: 137–159.

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

Knapp G, Biggerstaff BJ, Hartung J. 2006. Assessing the amount of heterogeneity in random-effects meta-analysis. Biom J 48: 271–285.

Lambert PC, Sutton AJ, Burton PR, Abrams KR, Jones DR. 2005. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. Stat Med 24: 2401–2428.

Malzahn U, Böhnung D, Holling H. 2000. Nonparametric estimation of heterogeneity variance for the standardised difference used in meta-analysis. Biometrika 87: 619–632.

Noma H. 2011. Confidence intervals for a random-effects meta-analysis based on Bartlett-type corrections. Stat Med 30: 3304–3312.
Pullenayegum EM. 2011. An informed reference prior for between-study heterogeneity in meta-analyses of binary outcomes. *Stat Med* 30: 3082–3094.

Shuster JJ. 2010. Empirical vs natural weighting in random effects meta-analysis. *Stat Med* 29: 1259–1265.

Sidik K, Jonkman JN. 2007. A comparison of heterogeneity variance estimators in combining results of studies. *Stat Med* 26: 1964–1981.

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

Thompson SG. 1994. Why sources of heterogeneity in meta-analysis should be investigated. *Br Med J* 309: 1351–1355.

Viechtbauer W. 2007. Confidence intervals for the amount of heterogeneity in meta-analysis. *Stat Med* 26: 37–52.

Zhang F. 2010. Matrix Theory (2nd edn). Springer, New York.