The Diamond Ratio: A Visual Indicator of the Extent of Heterogeneity in Meta-Analysis

Maxwell Cairns*[1], Geoff Cumming[1], Robert Calin-Jageman[2] and Luke A. Prendergast[1]

Abstract:

The result of a meta-analysis is conventionally pictured in the forest plot as a diamond, whose length is the 95% confidence interval (CI) for the summary measure of interest. The *Diamond Ratio* (DR) is the ratio of the length of the diamond given by a random effects meta-analysis to that given by a fixed effect meta-analysis. The DR is a simple visual indicator of the amount of change caused by moving from a fixed-effect (FE) to a random-effects (RE) meta-analysis. Increasing values of DR greater than 1.0 indicate increasing heterogeneity relative to the effect variances. We investigate the properties of the DR, and its relationship to four conventional but more complex measures of heterogeneity. We propose for the first time a CI on the DR, and show that it performs well in terms of coverage. We provide example code to calculate the DR and its CI, and to show these in a forest plot. We conclude that the DR is a useful indicator that can assist students and researchers to understand heterogeneity, and to appreciate its extent in particular cases.

Note: A revised version of this pre-print has been published in the British Journal of Mathematical and Statistical Psychology. The published version can be found at the following link: https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/bmsp.12258

[1] La Trobe University

[2] Dominican University

*Corresponding author information: Maxwell Cairns, Department of Mathematics and Statistics, La Trobe University, Melbourne Campus, Melbourne, Australia 3086. Phone: +613 9479 1107, Email: mrcairns994@gmail.com.

We propose confidence intervals and provide discussion on a previously defined, simple visual indicator of the extent of heterogeneity in a meta-analysis, the *Diamond Ratio* (DR). In what follows, we favour the use of length of intervals, to avoid confusion with the width of the *diamond* which is commonly used in meta-analysis (more details on the diamond to follow). Introduced by Higgins and Thompson (2002) as the *R Statistic*, and re-investigated by Cumming and Calin-Jageman (2017), the DR is the ratio of the lengths of the interval estimates (the confidence interval, CI, commonly depicted in a forest plot as a diamond) from the random effects (RE) meta-analysis and fixed effect (FE) meta-analysis. The length of the RE interval estimate is either equal to or greater than the length of the FE interval estimate. Consequently, the minimum for the DR is one, indicating equivalence between the RE and FE intervals. Values greater than one for the DR reflect the increase in length of the RE interval over that of the FE interval.

As a motivating example, Figure 1 shows the forest plot of a meta-analysis that includes, at the bottom, diamonds representing the results of a RE and FE analyses (labelled RE Model and FE Model respectively) of mean differences. The horizontal axis of each diamond is the 95% CI for the summary measure of interest, in this case, the population mean difference (in this article all CIs are 95% unless otherwise stated). The FE model assumes that the true mean differences are identical for all studies whereas the RE model allows them to vary to account for additional uncertainty. The DR is reported to be 1.40. For many graphs, especially those with comparatively long RE CIs, this can be estimated visually by comparing the lengths of the diamonds. This value of 1.40 indicates that the RE interval is 40% longer than the FE interval.

In a meta-analysis where heterogeneity is present, a Prediction Interval (PI) indicates a range of values within which we expect most of the unknown true effects to fall. These true effects, which we can think of as a super-population of effects from which those in the studies in our meta-analysis are drawn, only vary in terms of the heterogeneity variance. Thus, the length of the PI can be used as an indication of the amount of heterogeneity present with a larger length indicating a greater spread of the effects. The length of the PI, which is displayed in Figure 1 as the thin red line under the RE diamond, illustrates our estimate of the extent of variation of the true mean differences around the mean of the super-population. We estimate that an interval of this length captures 95% of the mean differences in the super-population. The interval is displayed centred on the RE mean (centre of the RE diamond), which is our point estimate of that unknown super-population mean. It is important to appreciate, however, that the interval refers to spread in the population and does not allow for uncertainty in our estimate of the super-population mean, which is depicted by the extent of the RE diamond. Informally, the PI indicates an interval length over which true mean differences being

estimated by different studies are likely to range. A short or zero length PI indicates small or zero heterogeneity; a long PI reflects large heterogeneity.

We refer to Figure 1 as an *enhanced* forest plot because it includes both RE and FE diamonds, the value of the DR, with its CI (discussed below), and a red line below the RE diamond to represent the length of the PI.

Heterogeneity and Two Models of Meta-Analysis

Heterogeneity is the extent to which the different studies in a meta-analysis estimate different values of the population parameter. In practice, the studies included in a meta-analysis virtually always vary, so an FE model is rarely justified. An example where an FE model might be justified would be an educational intervention run under identical conditions and for independent, yet homogenous cohorts. The RE model makes demanding assumptions, but is usually more realistic in that it allows for any amount of heterogeneity. A further advantage is that the RE model is still valid if studies are homogeneous and, if heterogeneity variance is estimated to be zero, then the RE and FE analyses are identical. For these reasons, the RE model is often recommended (Cumming, 2012; Schmidt, Oh, & Hayes, 2009).

For our purposes, the FE model provides a baseline, and our interest is in the extent to which heterogeneity causes the RE model to give a longer diamond. If the two models give the same result, DR = 1 and there is little or no heterogeneity. In approximate benchmarks that we introduce later, the DR of 1.55 in Figure 2 indicates moderate heterogeneity within this meta-analysis. Of course, one should not confuse this with measuring population heterogeneity, hence our reason for also depicting the prediction interval in our figure.

Heterogeneity may arise for many reasons and may be a nuisance and merely add error variability to a meta-analysis. However, if we can identify one or more moderating variables, or *moderators*, that can account for at least some of the heterogeneity, then heterogeneity may be valuable. Moderator analysis has the potential to identify variables that vary over studies and contribute to variation in the effects, the main measure in the meta-analysis. No single study may have manipulated a particular variable as an independent variable, but if that variable varies over studies, and is identified as a moderator, then the meta-analysis is giving fresh insight that may guide future research. Assessing the role of moderators can be especially important for using meta-analysis to guide theory evaluation and development.

Heterogeneity and moderator analysis are essential parts of meta-analysis. It is important that researchers and students understand heterogeneity and can appreciate the extent of heterogeneity in a particular meta-analysis. Unfortunately, the four conventional measures of heterogeneity (*Q*, I^2, H^2 and T^2) are a little complex and difficult to explain and interpret. For example, Borenstein, Higgins, Hedges, and Rothstein, (2017) pointed out that measures such as I^2 are often incorrectly thought of as population measures of heterogeneity despite being dependent on study sample sizes. They explained the PI and suggested that it can be a useful estimate of heterogeneity. In their introductory textbook, Cumming and Calin-Jageman (2017) used the DR as the basis for their discussion of heterogeneity. Their freely-available esci software (esci, 2017) accompanying that textbook reports the DR for any meta-analysis it can calculate.

Our aim is to examine properties of the DR, describe how it relates to conventional measures of heterogeneity, propose a CI on the DR, and make recommendations.

Fixed Effect and Random Effects Models

Throughout, we consider meta-analyses conducted using the inverse-variance weights approach as discussed in, e.g., Borenstein, Higgins, and Rothstein (2009); Cumming (2012); and Cumming and Calin-Jageman (2017).

Let μ_i be the true effect and V_{M_i} be the within studies variance associated with the i^{th} study. The FE model can be defined as:

$$M_i = \mu + \varepsilon_i, \quad (1)$$

where ε_i represents sampling error. Similarly, the RE model can be represented as

$$M_i = \mu + \zeta_i + \varepsilon_i, \quad (2)$$

where ζ_i is normally distributed with mean 0 and variance τ^2. This model can be used to meta-analyse data for different outcome measures including standardised mean differences (see Equation 3 below), log odds ratios, mean differences and Fisher transformed correlations, to name a few.

Standardised Mean Difference

As another example outcome measure, to use as the effect size for the models, we consider an estimate of the *Standardised Mean Difference* (SMD). Suppose we have a study containing two independent groups with population means μ_1 and μ_2 respectively and we assume that the population variances from each group are equal ($\sigma_1 = \sigma_2 = \sigma$). Then the standardized mean difference is defined to be

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}. \quad (3)$$

δ can be estimated by Cohen's *d* which is defined as

$$d = \frac{\overline{X_1} - \overline{X_2}}{S_{pooled}}, \quad (4)$$

where $\overline{X_1}$ and $\overline{X_2}$ are the sample means for each group and S_{pooled} is the pooled estimate of the standard deviation given as

$$S_{pooled} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}},$$

where the n_i are the group sample sizes and the S_i^2 are the sample variances. There is a small bias when using *d* to estimate δ, and a multiplicative bias-correcting factor is sometimes used (Hedges L. V., 1981). Throughout, when meta-analysing the SMD we use this bias corrected version in what follows and in all the figures, tables, and appendices. Furthermore, we prefer this bias correction because it is the estimator of δ used in the metafor package (Viechtbauer, 2010) which we use for the analyses displayed in this paper.

Figure 2 shows the forest plot for a second meta-analysis, this time with SMD as the outcome measure. DR = 1.55 suggests moderate heterogeneity relative to the effect variances. Note that the CIs for DR and *T*, while long, are not so very long as in Figure 1, where *k* (the number of studies) is much smaller.

Estimates of Heterogeneity

We start by giving a brief account of the conventional estimates *Q*, I^2, H^2, T^2, and the PI before considering how the DR relates to these. We use notation and formulas from the fuller discussions provided by Borenstein, et al. (2009), Cumming (2012), and Higgins and Thompson (2002).

Cochran's *Q,* being the total weighted sum of squares between studies, is defined as

$$Q = \sum W_i M_i - \frac{\sum (W_i M_i)^2}{\sum W_i}, \quad (5)$$

where W_i is the fixed-effect weight assigned to the i^{th} study. Under the FE model, the expected value of *Q* is its degrees of freedom, *df = (k* – 1). Larger values of *Q* indicate increasing heterogeneity, but values of *Q* are strongly influenced by *k*. It is therefore difficult to interpret *Q* without aligning it with the numbers of studies.

More often used in practice is I^2, which can be defined in a number of ways, depending on the choice of estimator for τ, but is most often defined as

$$I^2 = \frac{Q-df}{Q} \times 100\% \,. \tag{6}$$

Now, I^2 is the estimated percentage of total variance that reflects real differences in the population effects. Increasing I^2 values from zero towards 100% indicate heterogeneity increasing from zero to large. However, Borenstein et al. (2017) and others argued that I^2, as a percentage of variance measure, does not give an easily interpreted absolute indication of heterogeneity since I^2 is dependent on within-study sample sizes. Therefore, it cannot be interpreted as an estimate of heterogeneity in the population and they recommended the PI be used for that purpose.

The third conventional measure we consider is H^2 defined to be

$$H^2 = \frac{Q}{df} \,.$$

CIs for H^2 and I^2 were discussed by Higgins and Thompson (2002).

The fourth measure for heterogeneity is an estimate of τ^2. Statisticians have extensively studied a number of estimators for τ^2, perhaps around 15, including maximum likelihood estimators and iterative methods. A series of articles compared the different methods and evaluated the associated CIs (Viechtbauer, 2007; Veroniki et al., 2016). One of the more common methods is the DerSimonian and Laird (DL) method (DerSimonian & Laird, 1986). The DL estimator is simple to implement, meaning that it is often used as the default, sometimes only, estimator in software packages. The DL estimate of τ^2 given as

$$T^2 = \frac{Q - df}{\sum W_i - \frac{\sum W_i^2}{\sum W_i}}. \tag{7}$$

Given it is an estimator of heterogeneity variance, $T^2 = 0$ suggests little or no heterogeneity, and increasing T^2 indicates increasing heterogeneity. Note that τ^2 cannot be negative, so if T^2 is calculated to be less than 0, we set it equal to 0. Now τ, and therefore also T, are in the units of our effect and so using it to understand levels of heterogeneity can only be done in the context of its units of measurement. A value of T estimates the SD of the super-population of μ_i values and thus can be interpreted as an indication of how dispersed the values of μ_i are likely to be. Accordingly, Cumming (2012) discussed the CI on T, rather than the CI on T^2. Figure 1 and Figure 2 report T and its CI. Note that the error of estimation, and thus the length of that CI, is likely to be large, unless k is large.

More recently, Borenstein et al. (2017) built on these properties of *T* by advocating the PI as a readily interpretable way to indicate the spread of true values being estimated by the separate studies in a meta-analysis. They explained (p. 8) that the length of the PI is approximately 4*T* if within-study variances are small. In other words, we estimate that an interval of this length centred on μ, will capture 95% of those μ_i values. The red line in Figure 1 depicts this length; note that the reported values are consistent with length approximately equal to 4*T*. We don't know μ, so we don't know where along the lower axis in the forest plot the PI should be centred. Our best bet is to centre it at our point estimate of μ, the centre of the RE diamond. This, however, ignores the estimation error in that mean, which is quantified by the length of the RE diamond. Borenstein et al. explained (pp. 16-18) how a usually longer PI interval can be calculated allowing for that estimation error, but they recommended the simple PI of length about 4*T* when the focus is on heterogeneity considered as the spread of the true means in the super-population. That is our focus in this article, so this is the PI reported in Figure 1. Note that when $T = 0$ this simple PI has zero length, corresponding to the estimated zero variability in the super-population.

The Diamond Ratio

For V_{FE} denoting the variance estimate of the FE model summary estimate described in (1), M, an approximate $(1 - \alpha/2) \times 100\%$ CI for μ is $M \pm z_{(1-\alpha/2)}\sqrt{V_{FE}}$ where $z_{(1-\alpha/2)}$ is the $(1 - \alpha/2)$ percentile from the standard normal distribution and the length of this interval (the FE diamond) is $2 \times z_{(1-\alpha/2)}\sqrt{V_{FE}}$. For V_{RE} denoting the variance estimate of the RE model seen in (2), the RE interval and its length are similarly defined. The DR is then simply to be defined as

$$DR = \frac{Length\ of\ the\ RE\ diamond}{Length\ of\ the\ FE\ diamond} = \frac{\sqrt{V_{\text{RE}}}}{\sqrt{V_{\text{FE}}}}. \quad (8)$$

Thus, the DR, being a ratio of two CI lengths, tells us the proportional increase in CI length from the FE model to the RE model and therefore has a different interpretation to other measures. It can also be seen from its definition, that the DR is invariant to the confidence level (95% or some other value) of the CIs. So long as the FE and RE models are calculated under the same distributional assumptions (i.e. $z_{(1-\alpha/2)}$ as seen above or $t_{k-1,\alpha/2}$ etc) the DR will be the same. This invariance is useful since a measure that provides insight into what happens as one moves from one model to another from an inference perspective, should not in itself depend on the level of confidence. We can use the DR with some supplementary analyses to discuss how much heterogeneity is present in the analysis and this is similar to the way that I^2 can be used.

When discussing H^2, Higgins and Thompson (2002) noted that

$$I^2 = \frac{H^2 - 1}{H^2}\,.$$

They note this is true for their definition of I^2 (which we do not discuss here), though this is also true for the definition seen in (4). They also made the interesting observation that, when all studies have the same within-study variances, $\sqrt{H^2}$ = *DR*. In other words, when all the precisions are the same $DR^2 = 1/(1 - I^2)$. This relationship holds exactly only in the special case of uniform precision, but it gives general guidance for our suggestion of approximate benchmarks for the DR. Because the 25%, 50% and 75% benchmarks (Higgins J. P., Thompson, Deeks, & Altman, 2003) that are often used to indicate low moderate and high heterogeneity using I^2, and also based on our experience with DR, we suggest approximate benchmarks for the DR of

- DR < 1.2 little difference in the length of the FE and RE diamonds,
- DR 1.2 to 1.6 low to moderate difference,
- DR 1.6 to 2.0 moderate to large difference and
- DR > 2.0 to indicate very large differences between the diamond lengths.

Percentage of total variance is a type of measure familiar to many researchers, for example as the percentage of variance explained in multiple regression. Considering variability in a forest plot, with experience one can eyeball the amount of variability of estimates; this is I^2, a common measure of heterogeneity. For example, a number of CIs that don't overlap suggests some heterogeneity (Borenstein, Higgins, Hedges, & Rothstein, 2017). This amounts to assessing total variability against likely variability in the homogeneous case. The DR is based on an analogous comparison of RE and FE variability, and is certainly easier to eyeball, given the two diamonds. Note, again, that the DR being based on variability contrasts with the PI, which gives information about likely spread of true mean differences, in units of the effect.

Confidence Intervals for the DR

We consider several types of CIs for the DR. The first is a simple Wald-type CI for the log DR (i.e. $\log \widehat{DR} \pm z_{(1-\alpha/2)} \times SE$) based on an approximate standard error, SE. For this approach, we construct the CI for the log DR to improve statistical properties such as convergence to normality, and then back-exponentiate to return to the DR scale. The second, also constructed firstly for the log DR, is a bias-corrected version of the first. The third type uses a substitution of the intervals calculated for τ^2. We will consider two variations of each of those three types of CIs, for a total of six ways to construct such an interval.

First, we state a theorem we need for the first two types of CI. Let $W_i^* = 1/(V_{M_i} + T^2)$ be the weight assigned to the i^{th} study of an RE meta-analysis, where T^2 is an estimate of τ^2 (Borenstein et al., 2009, Chapter 12).

Theorem 1: *The approximate variance and bias, b, for the* $\log(\widehat{DR})$ *estimator are*

$$\mathrm{Var}[\log \widehat{DR}] = \mathrm{Var}(T^2)\left[\frac{1}{2}\frac{1}{\sum_i^k W_i^*}\sum_{i=1}^{k}(W_i^*)^2\right]^2,$$

$$b = \frac{1}{2}\mathrm{Var}(T^2)\left[\frac{1}{2}\left[\frac{1}{\left(\sum_i^k W_i^*\right)^2}\right] - \frac{1}{\sum_i^k W_i^*}\sum_{i=1}^{k}(W_i^*)^3\right].$$

The proof of Theorem 1 can be found in Appendix A.

Our first CI for the DR is a (1 – α) × 100% Wald-type CI of the form

$$\exp\left\{\log\left(\widehat{DR}\right) \pm z_{1-\alpha/2}\sqrt{\mathrm{Var}\left[\log\left(\widehat{DR}\right)\right]}\right\}, \tag{9}$$

where the variance component can be found in Theorem 1.

The second CI seeks to reduce the bias in the estimation of $\log\left(\widehat{DR}\right)$ and is defined to be

$$\exp\left\{\left(\log\left(\widehat{DR}\right) - b\right) \pm z_{1-\alpha/2}\sqrt{\mathrm{Var}\left[\log\left(\widehat{DR}\right)\right]}\right\} \tag{10}$$

where *b* can also be found in Theorem 1.

We refer to the first CI as a Wald-type (WT) CI and to the second as a bias-corrected WT (bWT). Given that there are several known estimators for τ^2, we will consider two of the more commonly used estimators: the DL estimator (7) and a restricted maximum likelihood estimator (REML). The REML estimator is the default choice in the R (R Core Team, 2020) package metafor (Viechtbauer, 2010) and has been shown to have good properties (Viechtbauer, 2007). It is important to note, however, that the REML estimator is not so readily available in other packages, so consideration of the DL estimator is also warranted. Specifically, our intervals under consideration for types 1 and 2 are

- WT-DL: Wald-type CI with the DL estimator.
- WT-REML: Wald-type CI with the REML estimator.
- bWT-DL: Wald-type with bias correction CI with the DL estimator.
- bWT-REML: Wald-type with bias correction CI with the REML estimator.

Our third type of CI is to use a substitution of the CI found for τ^2. As *DR* is a monotonic increasing function of τ^2 we can use the bounds of the CI for τ^2 to calculate the *DR* CI. This method was mentioned in Higgins and Thompson (2002) and is carried out as follows. Recall that $\widehat{DR}$ (8) depends on the variance estimators for the FE and RE models and the RE variance depends on T^2. Consequently, to make things clearer we let $\widehat{DR}(T^2)$ denote the estimated DR, emphasising dependence on T^2 and a confidence interval for DR is then $\left[\widehat{DR}(L), \widehat{DR}(U)\right]$ where L and U are the lower and upper bounds for the interval for τ^2. For our CIs for τ^2 we have used the restricted maximum likelihood (REML) estimator and a Q-profiling approach (Viechtbauer, 2007; Knapp, Biggerstaff, & Hartung, 2006). Note that the Q-statistic in the Q-Profiling method is not the same statistic as (5), but a general method using RE weights.

We denote these intervals to be considered as

- Sub-REML: Substitution CI with the τ^2 interval based on the REML estimator.
- Sub-Q: Substitution CI with the Q-profile τ^2 interval estimator.

Data availability

All data and code are available online. We also provide R scripts, so a researcher can run their own simulations as well as being able to replicate the results in this paper. We provide additional examples and plots to aid researchers in using the DR for their own research. We also include a wrapper function adding the DR and CIs to an rma object from metafor. For non-R users, the esci module (esci in jamovi, 2021) for the jamovi software (jamovi (Version 1.6), 2021) includes the DR and bWT-DL CIs.

Evaluation of CIs on the DR

In this section we assess our CIs by using simulation to estimate coverage probability and compare with the nominal 95%. We used real data sets available on the metafor package in R to guide choices of numbers of studies, sample sizes and parameter values.

Simulations

We report two separate simulation studies. The first is simulated using data from real data sets that we introduce below. In this study we chose values of τ = 0.2, 0.4, 0.6 and 0.8. The second study is a series of simulations with sample sizes randomly sampled from between 10 and 50 inclusive. We considered τ equal to 0.2, 0.4 and 0.6. We simulated the SMD estimates assuming that the within-study data were normally distributed. In total 100,000 simulation runs were conducted for each value of τ, for each CI method. We sampled our observed statistics using a non-central *t* distribution (for details see the simulation code) with a varying number of studies. In our first simulation study

we chose our true values to be the values estimated from the entire data set while in the second study we estimated our true values using inverse variance weights assuming a normal distribution (see our example code for more details).

The first data set used in the first study is Hospital Stay of Stroke Patients (Normand, 1999) data set, which contains nine studies with study total sample sizes of: 311, 63, 146, 36, 21, 109, 67, 293 and 112, split approximately evenly between two arms. The second is the Teacher Expectations on Pupil IQ data set (Raudenbush, 1984), which contains 19 studies with total sample sizes between 33 and 746 with varying degrees of imbalance between the arms. The third is the Writing to Learn Interventions (Bangert-Drowns, Hurley, & Wilkinson, 2004) data set, which has 48 studies with sample sizes from 16 to 542 inclusive—in our simulations we split total sample size evenly between the two arms. All confidence intervals and coverages are available on our OSF page.

Results

Recall that we define WT and bWT intervals in (9) and (10) respectively and DL and REML refer to the methods of estimating τ^2. The coverage results from the first simulations, shown in Table 1, indicate that the three REML based CIs (WT-REML, bWT-REML and Sub-REML) did not perform as well as the other methods in most cases, with coverage being very over-conservative. Considering WT-DL and bWT-DL, the table shows that bWT-DL performed better overall with WT-DL being more conservative. Finally, Sub-Q performs well in most circumstances. Typically, it was the bWT-DL and Sub-Q CIs with closest to nominal coverage. Given also that bWT-DL is easy to implement in any package and that Sub-Q is possible in languages such as R, we focus further attention on these two CIs.

Figure 3 provides coverage probabilities for the bWT-DL, bWT-REML, bWT-HE (discussed briefly below) and Sub-Q CIs based on simulated data from the second set of simulations. For these settings, for small τ in Plot A, coverage for the bWT-DL interval changes from slightly liberal (i.e., approximately 0.92 at $k = 3$) to conservative with increasing k. For moderate to large τ (Plot C), coverage can be conservative for small k, and declines as k increases. In this case the bWT-REML intervals provide coverage closer to the nominal as k increases. This could be as a result of the negative bias of the DL estimator which has been shown to exist when τ is large (see, Veroniki, et al., 2016 and associated papers). We note that in Plot C the coverage of the bWT-DL intervals continues to decline as k increases. We also show intervals using an unbiased estimator of τ^2, the Hedges-Olkin (HE) estimator (Hedges & Olkin, 1985). These intervals show strong coverage for all values of τ^2. The Sub-Q method stays very close to the nominal .95 in all situations depicted. Considering our full set of simulation results, use of the Sub-Q method appears optimal. However, this method is not

available in all packages and across all platforms and is difficult to calculate. On the other hand, bWT-DL can be easily calculated using any software, and even by hand. Consequently, many researchers may prefer to use this method given that the simulation results showed that interval coverage is typically close to nominal, except when *k* is very small or τ is suspected to be large. Thus, we recommend in general that the CIs should be calculated using the Sub-Q method when reporting the DR. The bias-corrected Wald-type intervals may be preferred by some researchers due to their simplicity, although we recommend using either the REML or HE estimators of τ^2 when there is a large amount of heterogeneity present since the DL estimator of τ^2 results in intervals for the DR with low coverage. We briefly discuss lengths for bWT-DL and Sub-Q methods in Appendix B.

An example using real data

To illustrate use of the DR in practice we consider as an example the Hospital Stay of Stroke Patients (Normand, 1999) data set, with SMD (3) as the outcome measure. We use code available at our OSF page to carry out the meta-analysis and generate the enhanced forest plot shown in Figure 4.

Figure 4 reports DR = 4.21, which indicates a very large amount of heterogeneity relative to the error variability that gives the short FE diamond. The 95% CI of [2.81, 8.99] for the DR, calculated using the bWT-DL approach, is long, indicating very imprecise estimation, even though the non-inclusion of 1.0 indicates statistically significant heterogeneity. The PI length of 2.94 also indicates a very large amount of heterogeneity. There could be many causes for this heterogeneity, but the researcher may choose to try to account for at least some of it by using, for example, subgroup analyses or a moderator analysis using meta-regression. Ideally, the details of a moderator analysis should be specified before conducting the meta-analysis; otherwise it needs to be described as exploratory.

Conclusions: Interpretation of the DR and PI

A main advantage of the DR is that it can be eyeballed from a simple forest plot that reports both FE and RE results. Looking, for example, at Figures 1, 2, or 4, we can easily compare the lengths of the CIs for FE and RE analyses. Our discussion of the DR in the context of the traditional estimates of heterogeneity leads to our conclusion that the DR is a useful and understandable indicator of the extent of heterogeneity in a meta-analysis. Like I^2, the DR depends on the variances of the effect estimators and so is a measure of study-specific heterogeneity and not a population (global) measure. In other words, the DR is based on lengths of interval estimates (CIs), which should help readers appreciate that it is an estimate of heterogeneity calculated from the data, and not itself a population quantity. Additionally, the DRs interpretation as a measure of what happens when we move from the fixed-effect model to the random effects model, with DR = 1 denoting identical

inference, may make this a useful complement to other estimates of heterogeneity. Higgins and Thompson (2002) also comment on interpretation and cite Kish (1965, Page 258) when noting that the the DR (or R in their paper) may be interpreted similarly to the design effect in cluster sampling.

Our main original contribution in this article is to provide a good approximate CI on the DR. We recommend that researchers should use the DR as a measure of study-specific heterogeneity together with the PI as a further estimate of population heterogeneity. It is essential when considering the DR to also consider its CI. In Figure 1, e.g., the DR is 1.40 [1.00, 3.09]. The great length of the CI indicates that in this meta-analysis with only k = 10 studies we have an extremely poor estimate of the extent of heterogeneity, which could plausibly be anywhere from zero to large. The value of T is .07 [0, .21]. Recalling that PI length is approximately 4T, the CI on PI length is approximately [0, .84], which indicates that PI length could plausibly be anywhere from zero to more than twice the length of the red line in Figure 1. The CIs on DR and on PI length are giving us the same message of great uncertainty about the extent of heterogeneity, with the former based on the trade-off in moving from the FE model to the RE model. By contrast, in other examples we saw evidence for large heterogeneity, e.g., Figure 4 even for just k = 9 studies.

Following consideration of the link between I^2 and DR, both being study-specific measures of heterogeneity, we suggested approximate DR benchmarks to indicate low, moderate and high differences between the FE and RE diamond lengths. We take study-specific heterogeneity to mean heterogeneity relative to the settings of the meta-analysis being considered. We emphasise that interpretation of any DR value should be based primarily on informed judgment in the particular research context, rather than on our or any other benchmarks.

In conclusion, the DR is an easily understood visual indicator of the move from the FE interval to the RE interval in a meta-analysis. A researcher may then use this to investigate the amount of heterogeneity for the specific meta-analysis of interest. The CI for the DR indicates the extent of uncertainty—often large—in the estimate of heterogeneity. Understanding heterogeneity is central to getting the most out of meta-analysis. We suggest that the DR, and its CI, can be valuable for students as they learn about meta-analysis, and for researchers as they interpret and communicate their meta-analyses.

Declarations of Conflict of Interests

The author(s) declare that there were no conflicts of interest with respect to the authorship or the publication of this article.

References

Bangert-Drowns, R. L., Hurley, M. M., & Wilkinson, B. (2004). The effects of school-based writing-to-learn interventions on academic achievement: A meta-analysis. *Review of educational research, 74*, 29-58. doi:10.3102/00346543074001029

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Intorduction to Meta-Analysis.* Chichester: John Wiley and Sons.

Borenstein, M., Higgins, J. P., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. *Research Synthesis Methods, 8*, 5-18. doi:10.1002/jrsm.1230

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis.* New York: Routledge.

Cumming, G., & Calin-Jageman, R. (2017). *Introduction to the New Statistics: Estimation, Open Science, and Beyond.* New York: Routledge.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled clinical trials, 7*, 177-188. doi:10.1016/0197-2456(86)90046-2

esci [Computer software]. (2012). Retrieved from Introduction to the New Statistics: thenewstatistics.com/itns/esci/esci-for-utns/

esci [Computer software]. (2017). Retrieved from Introduction to the New Statistics: thenewstatistics.com/itns/esci

esci in jamovi [Computer software]. (2021). Retrieved from Introduction to the new Statistics: https://thenewstatistics.com/itns/esci/jesci/

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *journal of Educational Statistics, 6*, 107-128. doi:10.3102/10769986006002107

Hedges, L. V., & Olkin, I. (1985). *Statistical Methods for Meta-Analysis.* Academic Press.

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *Bmj, 327*, 557-560. doi:10.1136/bmj.327.7414.557

Higgins, J., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in medicine, 21*, 1539–1558. doi:10.1002/sim.1186

jamovi (Version 1.6) [Computer software]. (2021). Retrieved from jamovi project: https://www.jamovi.org

Kish, L. (1965). Survey Sampling. New York: Wiley.

Knapp, G., Biggerstaff, B. J., & Hartung, J. (2006). Assessing the amount of heterogeneity in random-effects meta-analysis. *Biometrical Journal: Journal of Mathematical Methods in Biosciences, 48*, 271-285. doi:10.1002/bimj.200510175

Normand, S. T. (1999). Meta-analysis: Formulating, evaluating, combining, and reporting. *Statistics in Medicine, 18*, 321-359. doi:10.1002/(SICI)1097-0258(19990215)18:3<321::AID-SIM28>3.0.CO;2-P

R Core Team. (2020). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Raudenbush, S. W. (1984). Magnitude of teacher expectancy effects on pupil IQ as a function of the credibility of expectancy induction: A synthesis of findings from 18 experiments. *Journal of Educational Psychology, 76*, 85-97. doi:10.1037/0022-0663.76.1.85

Schmidt, F. L., Oh, I.-S., & Hayes, T. L. (2009). Fixed-versus random-effects models in meta-analysis: Model properties and an empirical comparison of differencesin results. *British journal of mathematical and statistical psychology, 62*, 97–128. doi:10.1348/000711007X255327

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., . . . Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research synthesis methods, 7*, 55-79. doi:10.1002/jrsm.1164

Viechtbauer, W. (2007). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in medicine, 26*, 37-52. doi:10.1002/sim.2514

Viechtbauer, W. (2010). Conducting meta-analyses in 'R' with the 'metafor' package. *Journal of Statistical Software*, 1-48. doi:10.18637/jss.v036.i03

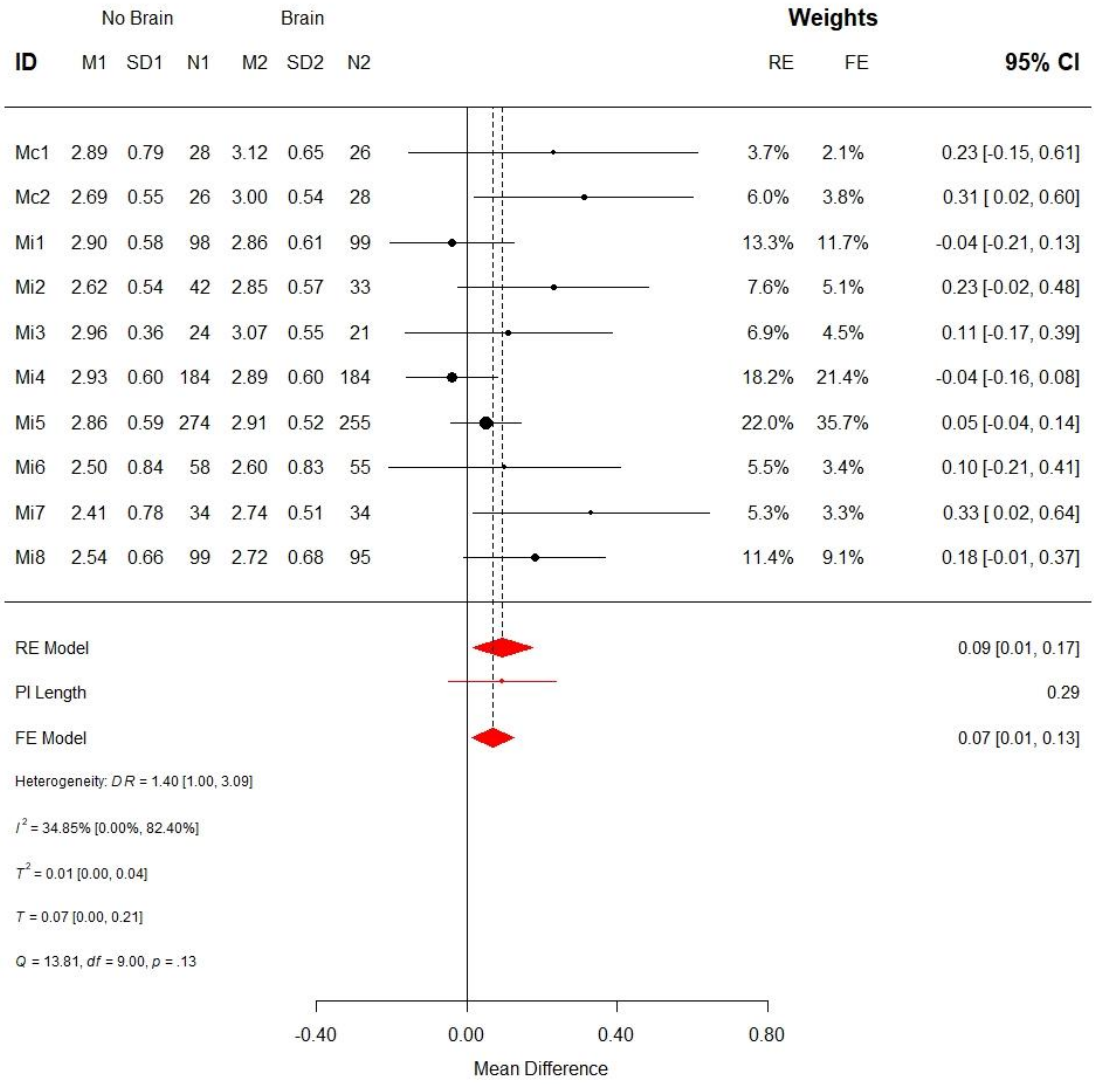

Fig. 1 Forest plot summarising a meta-analysis performed on data in Figure 9.2 of Cumming and Calin-Jageman (2017). The plot is enhanced by inclusion of both RE and FE diamonds, the value of the DR, with its CI (calculated using the bWT-DL method defined in the text), and a red line below the RE diamond that represents the length of the associated prediction interval for the size of effect in the super-population (PI). The length of the PI is 0.29, as reported in the figure. Eyeballing the ratio of the lengths of the two diamonds in the figure agrees with the reported value of DR = 1.40.

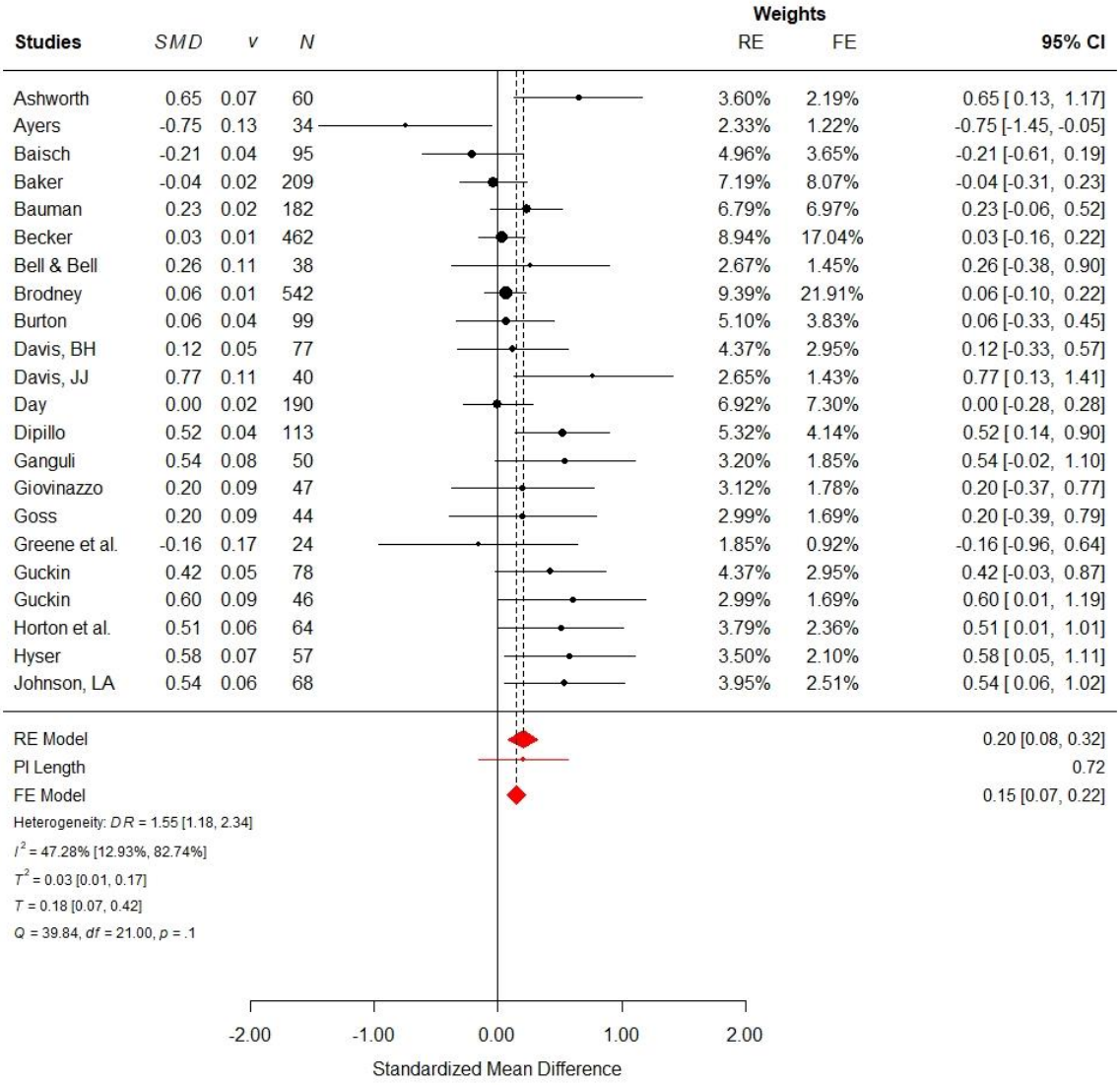

Fig. 2 An enhanced forest plot similar to that in Figure 1, but using SMD as the outcome measure. There appears to be moderate heterogeneity as measured by the move from the FE interval to the RE interval, which gives DR = 1.55. The CI here is calculated using the bWT-DL method described in the text. In this example we use a subset of 22 studies that is available in the R metafor package from the Effectiveness of Writing-to-Learn Interventions (Bangert-Drowns, Hurley, & Wilkinson, 2004) data set. Note for this example, the data set contained the SMD values and associated variances needed for the meta-analysis. These are labelled *SMD* and *v* in the forest plot respectively.

Table 1 Comparison of coverages for the six CI methods, for three data sets and a range of τ values

Data set	Method (CI)	τ 0	0.2	0.4	0.6	0.8
Hospital Stay of Stroke Patients	WT-REML	.87	.96	.90	.92	.93
	bWT-REML	.78	.90	.95	.93	.94
	Sub-REML	1.00	.70	.81	.83	.84
	WT-DL	.99	1.00	.94	.93	.92
	bWT-DL	.92	.97	.97	.96	.96
	Sub-Q	.98	.95	.95	.96	.96
Teacher Expectations on Pupil IQ	WT-REML	.83	.84	.92	.94	.94
	bWT-REML	.77	.92	.93	.94	.95
	Sub-REML	.99	.80	.88	.89	.88
	WT-DL	.98	.97	.94	.94	.93
	bWT-DL	.93	.97	.96	.96	.96
	Sub-Q	.97	.95	.95	.96	.96
Writing-to-Learn Interventions	WT-REML	.74	.81	.92	.94	.94
	bWT-REML	.70	.81	.92	.94	.95
	Sub-REML	.89	.82	.91	.92	.91
	WT-DL	.96	.96	.94	.93	.91
	bWT-DL	.92	.98	.96	.96	.95
	Sub-Q	.98	.95	.95	.96	.96

Note: Nominal coverage is .95. Abbreviated names for the six CI methods are explained above. The simulations, each comprising 10,000 simulated data sets, used sample sizes from the three named datasets, which are from the metafor package in R.

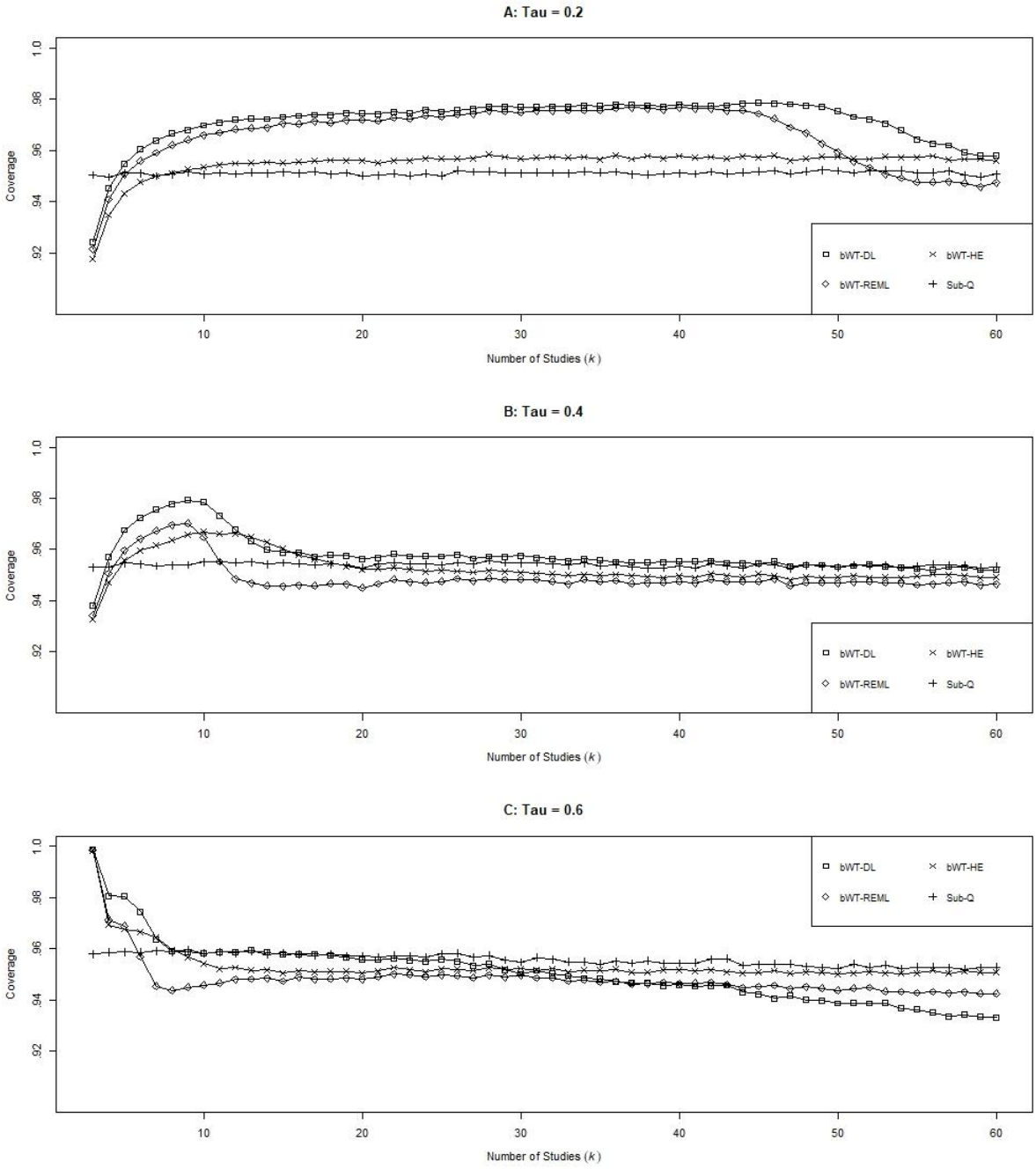

Fig. 3 Comparison of coverages for the bWT-DL, bWT-REML, bWT-HE (bWT intervals using the Hedges-Olkin estimator for τ^2) and Sub-Q methods as the number of studies increases. Nominal coverage was .95 and each coverage curve was calculated from 100,000 trials. Sample sizes for each arm of a study were randomly sampled from between 10 and 50 inclusive, and the effect measure was SMD.

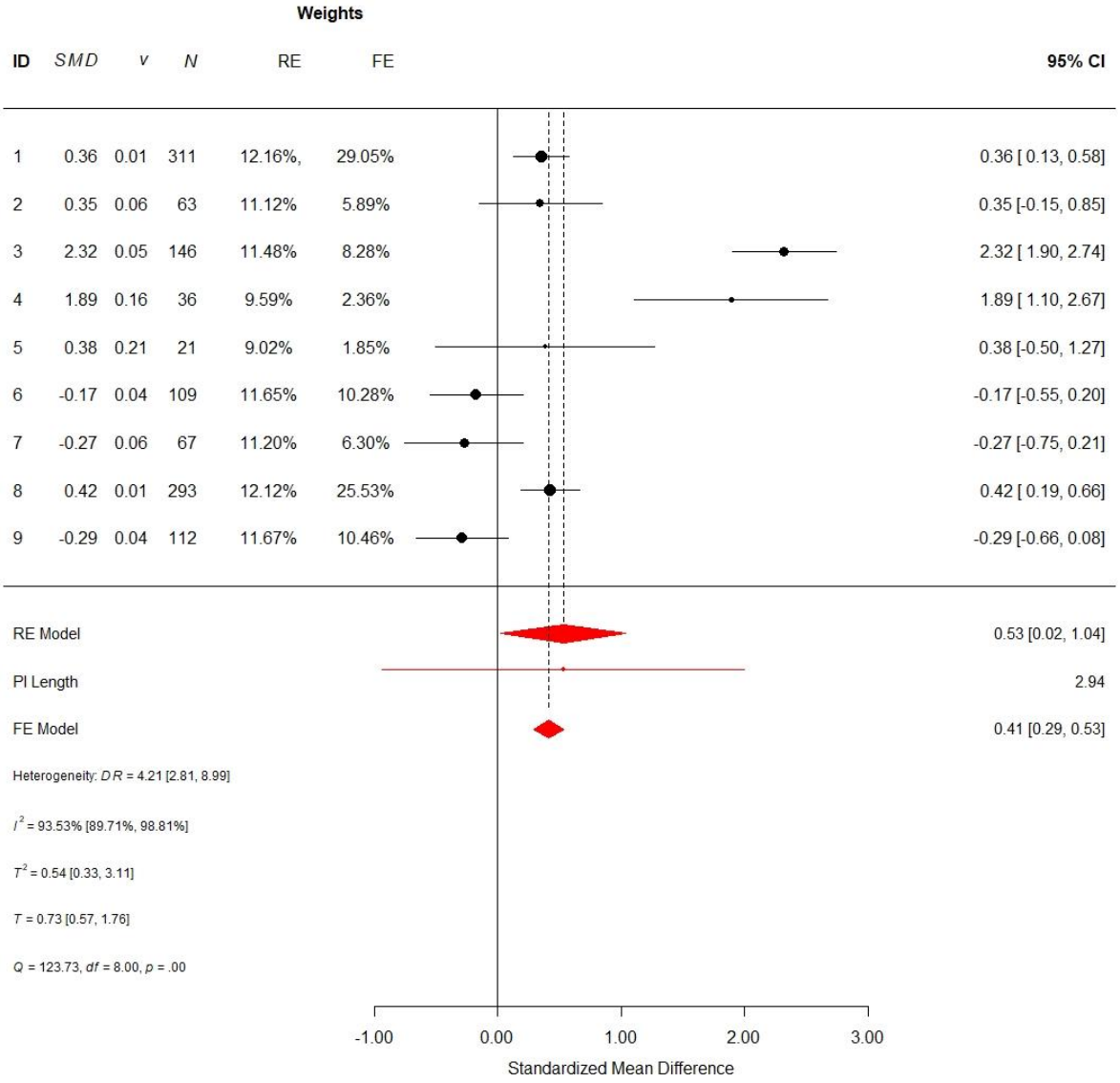

Fig. 4 Enhanced forest plot of data from Hospital Stay of Stroke Patients data (Normand, 1999). The very substantial DR of 4.21, being a ratio, reflects both the small extent of error variability indicated by the short FE diamond and the relatively much larger heterogeneity between studies indicated by the long RE diamond. In other words, DR = 4.21 indicates a large change in moving from the FE model to RE model. The CI on DR is calculated using the bWT-DL method.

Appendix A

Proof of Theorem 1

From the definition of the DR in Equation 8, we can see that

$$\log(DR) = \frac{1}{2}\log\left(\sum_i^k W_i\right) - \frac{1}{2}\log\left(\sum_i^k W_i^*\right), \quad \text{(A1)}$$

where W_i and W_i^* are defined above. Then, the Taylor approximation of the log(DR) is

$$\log(DR) \approx \log\left(\widehat{DR}\right) + (T^2 - \tau^2) \cdot \frac{d\log(DR)}{d\tau^2}\bigg|_{\tau=T} + \frac{1}{2}(T^2 - \tau^2)^2 \cdot \frac{d^2\log(DR)}{d(\tau^2)^2}\bigg|_{\tau=T}. \quad \text{(A2)}$$

From *Equation A1* we have

$$\frac{d\log(DR)}{d\tau^2} = -\frac{1}{2}\frac{1}{\sum_i^k W_i^*}\frac{d\sum_i^k W_i^*}{d\tau^2}$$

$$= \frac{1}{2}\frac{1}{\sum_i^k W_i^*}\sum_{i=1}^k (W_i^*)^2,$$

and

$$\frac{d^2\log(DR)}{d(\tau^2)^2} = \frac{d\left(\frac{1}{2}\left[\frac{1}{\sum_i^k W_i^*}\right]\sum_{i=1}^k (W_i^*)^2\right)}{d\tau^2}$$

$$= \frac{1}{2}\left[\frac{1}{\left(\sum_i^k W_i^*\right)^2}\right] - \frac{1}{\sum_i^k W_i^*}\sum_{i=1}^k (W_i^*)^3.$$

Now we wish to approximate the expected value of $\log(DR)$. Substituting into *Equation A2*

$$\mathrm{E}[\log(DR)] \approx \mathrm{E}\left[\log\left(\widehat{DR}\right) + (T^2 - \tau^2) + \frac{1}{2}(T^2 - \tau^2)^2 \cdot \frac{d^2\log(DR)}{d(\tau^2)^2}\bigg|_{\tau=T}\right]$$

$$= \mathrm{E}\left[\log\left(\widehat{DR}\right)\right] + \mathrm{E}\left[(T^2 - \tau^2) \cdot \frac{d\log(DR)}{d\tau^2}\bigg|_{\tau=T}\right] + \mathrm{E}\left[\frac{1}{2}(T^2 - \tau^2)^2 \cdot \frac{d^2\log(DR)}{d(\tau^2)^2}\bigg|_{\tau=T}\right]$$

$$= \mathrm{E}\left[\log\left(\widehat{DR}\right)\right] + \mathrm{E}\left[\frac{1}{2}(T^2 - \tau^2)^2 \cdot \frac{d^2\log(DR)}{d(\tau^2)^2}\bigg|_{\tau=T}\right], \text{ as } \mathrm{E}[(T^2 - \tau^2)] = 0.$$

Therefore, the bias is

$$b = \mathrm{E}\left[\frac{1}{2}(T^2 - \tau^2)^2 \cdot \frac{d^2 \log(DR)}{d(\tau^2)^2}\bigg|_{\tau=T}\right]$$

$$= \frac{1}{2}\mathrm{E}[(T^2 - \tau^2)^2]\mathrm{E}\left[\frac{d^2 \log(DR)}{d(\tau^2)^2}\bigg|_{\tau=T}\right]$$

$$= \frac{1}{2}\mathrm{Var}(T^2)\left[\frac{d^2 \log(DR)}{d(\tau^2)^2}\bigg|_{\tau=T}\right]$$

$$= \frac{1}{2}\mathrm{Var}(T^2)\left[\frac{1}{2}\left[\frac{1}{\left(\sum_i^k W_i^*\right)^2}\right] - \frac{1}{\sum_i^k W_i^*}\sum_{i=1}^{k}(W_i^*)^3\right].$$

From *Equation A2* we can see that

$$\mathrm{Var}[\log(DR)] \approx \mathrm{Var}\left[\log\left(\widehat{DR}\right) + (T^2 - \tau^2) \cdot \frac{d \log(DR)}{d\tau^2}\bigg|_{\tau=T}\right]$$

$$= \mathrm{Var}\left[(T^2 - \tau^2) \cdot \frac{d \log(DR)}{d\tau^2}\bigg|_{\tau=T}\right]$$

$$= \mathrm{Var}(T^2)\left[\frac{d \log(DR)}{d\tau^2}\bigg|_{\tau=T}\right]^2$$

$$= \mathrm{Var}(T^2)\left[\frac{1}{2}\frac{1}{\sum_i^k W_i^*}\sum_{i=1}^{k}(W_i^*)^2\right]^2.$$

This completes the proof of Theorem 1.

Appendix B

Further Exploration of Coverages

Further explorations of our proposed methods are considered in Figure B1. We show contour plots varying over sample sizes (equal for each arm) and numbers of studies, *k*. SMD was considered and set equal to zero in the population ($\delta = 0$). 10,000 simulation runs were considered, for each setting, and we looked at three choices of τ, 0.2, 0.4, 0.6, for bWT-DL and Sub-Q intervals. The bWT-DL interval coverage moved from being slightly conservative for small to moderate sample sizes and number of studies (depending on the value of τ), to close to nominal for larger values. As τ increases, close to nominal coverage is achieved for small sample sizes and *k*. This behaviour matches the results seen in Figure 3. The Sub-Q interval provides stable and close-to-nominal coverage for all choices. Results from these simulations show that the lengths of the two confidence intervals are reasonable. When $\tau = 0.2$, we saw an average length of between 5 and 7 when $k = 3$ decreasing to between 0.4 and 0.7 as *k* increases for the Sub-Q intervals depending on sample size ($N \in [10, 70]$). The bWT-DL intervals showed similar behaviour except the average lengths moved from between 4-6 down to 0.4-0.7. Increasing τ to 0.6 we saw similar behaviour for both intervals except the spread of values became much wider. For Sub-Q intervals the average length went from between 8-18 to 0.7-1.5 and for the bWT-DL intervals the average lengths went from between 5-12 to 0.6-1.5. It is important to note that as the number of studies increase, the length initially decreases quite quickly: bWT-DL lies between 2.4 and 5.4 while Sub-Q rests between 2.8 and 6.3 for $k = 6$ studies.

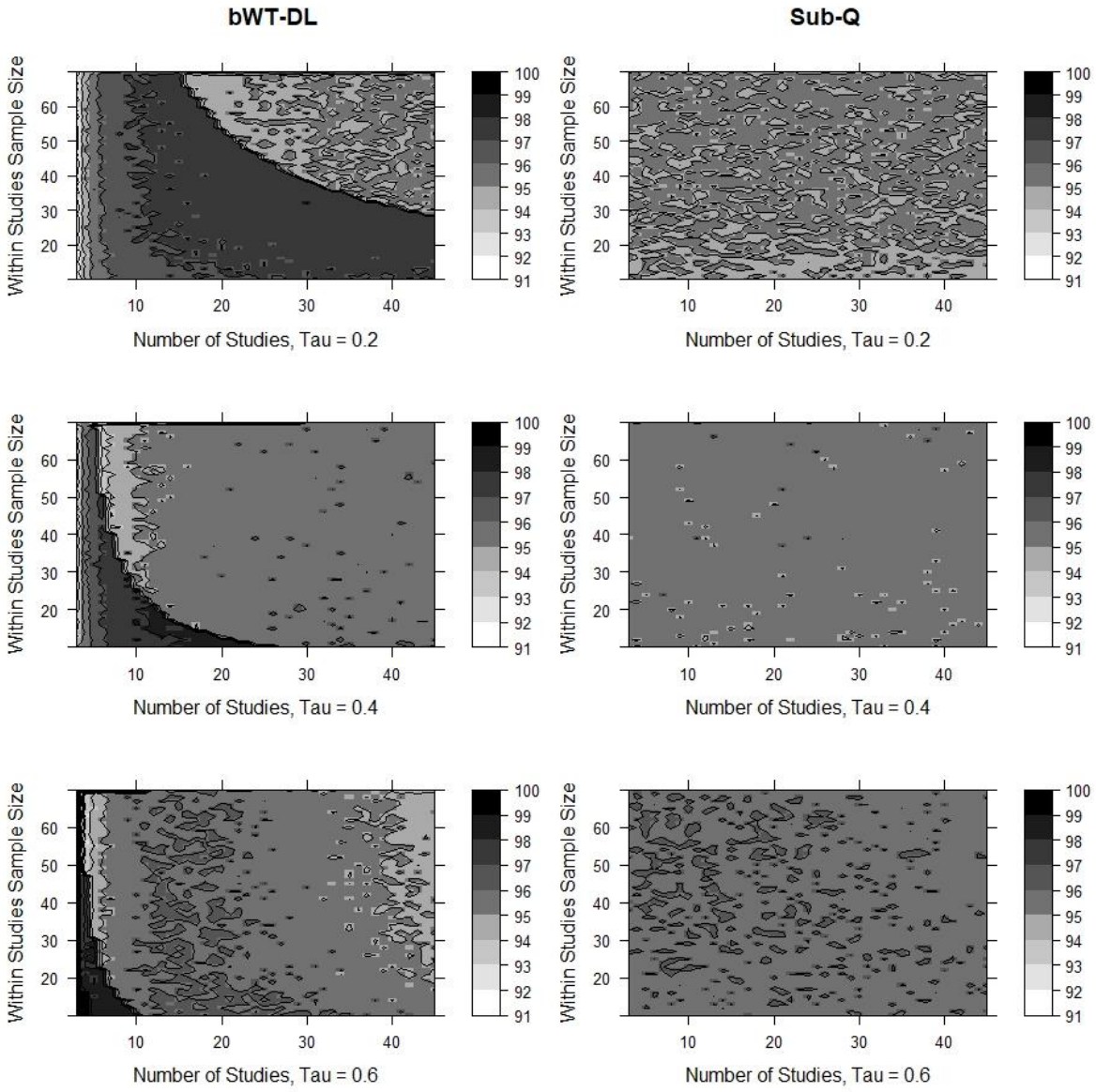

Fig. B1 Contour plots of coverages, with a nominal .95, of the CI on DR from meta-analyses using SMD as effect measure. The bWT-DL and Sub-Q methods were used. There were 10,000 trials simulated for each setting. The population effect size (δ) was 0 in every case.