Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
21(4) 409–426

DOI: 10.1177/0962280210392008
smm.sagepub.com
SAGE

Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: A simulation study

Evangelos Kontopantelis[1] and David Reeves[2]

Abstract

Meta-analysis (MA) is a statistical methodology that combines the results of several independent studies considered by the analyst to be 'combinable'. The simplest approach, the fixed-effects (FE) model, assumes the true effect to be the same in all studies, while the random-effects (RE) family of models allows the true effect to vary across studies. However, all methods are only correct asymptotically, while some RE models assume that the true effects are normally distributed. In practice, MA methods are frequently applied when study numbers are small and the normality of the effect distribution unknown or unlikely. In this article, we discuss the performance of the FE approach and seven frequentist RE MA methods: DerSimonian–Laird, Q-based, maximum likelihood, profile likelihood, Biggerstaff–Tweedie, Sidik–Jonkman and Follmann–Proschan. We covered numerous scenarios by varying the MA sizes (small to moderate), the degree of heterogeneity (zero to very large) and the distribution of the effect sizes (normal, skew-normal and 'extremely' non-normal). Performance was evaluated in terms of coverage (Type I error), power (Type II error) and overall effect estimation (accuracy of point estimates and error intervals).

Keywords

meta-analysis, non-normal, profile likelihood, power, coverage, simulation, DerSimonian–Laird, Biggestaff–Tweedie, Sidik–Jonkman, permutations

[1]National Primary Care Research and Development Centre, University of Manchester, Williamson Building, 5th Floor, Oxford Road, M13 9PL, UK
[2]Health Sciences Primary Care Research Group, University of Manchester, Williamson Building, 5th Floor, Oxford Road, M13 9PL, UK

Corresponding author:
Evangelos Kontopantelis, National Primary Care Research and Development Centre, University of Manchester, Williamson Building, 5th Floor, Oxford Road, M13 9PL, UK
Email: e.kontopantelis@manchester.ac.uk

1 Introduction

Although efforts to pool results from individual studies date as far back as 1904 and the first such attempt that assessed the effect of a therapeutic intervention was published in 1955,[1] it was not until 1976 that the term 'meta-analysis' (MA) appeared. The progenitor of this label, an educational researcher, defined MA as 'the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings'.[2]

A primary concern for meta-analysts is the statistical heterogeneity between the studies included in an analysis, which can be attributed to clinical and/or methodological diversity.[3] Clinical heterogeneity describes variability that arises from different populations, interventions, outcomes and follow-up times, while methodological heterogeneity is related to differences in trial design and quality.[4] Detecting,[5] quantifying,[6,7] investigating causality[8,9] and dealing with heterogeneity[10] is not particularly straightforward, to say the least.

If variation among the estimated effects of the individual studies is not above that expected by chance (homogeneity not rejected) researchers usually select the fixed-effect (FE) model[11] to combine the separate estimates into a single result. However, the numerous underlying variables in medical research studies establish homogeneity as a rare commodity and some degree of variability between studies may be anticipated.[12] When heterogeneity is identified, as is usually the case, it has been emphasised that it should be explored in order to increase both the clinical relevance and the scientific value of the findings.[4,9] Unfortunately, this is not always possible in practice; subgroup analysis or meta-regression using study-level variables can explain and reduce heterogeneity but rarely are there sufficient studies to support such approaches. The best alternative is to analyse the heterogeneous studies using a random-effects (RE) model.[13] This model builds in an estimate of the between-study variation in effect size and typically provides wider confidence intervals for the overall effect than its FE counterpart. Nevertheless, critics argue that it may still be inappropriate to combine heterogeneous trials: an 'average' MA result may not be clinically useful if the effect is context-specific.[4,14]

An additional problem for both RE and FE models is that most commonly available analysis techniques assume true study effects to be normally distributed. This assumption is deemed to be critical[14] but, within a particular analysis, not easily verified or justified[11] – especially when the number of studies is small. Consequently, MA models are open to the criticism of being unrealistic.[15] Departure from normality may affect a model's performance and since normality in real-life data might be the exception rather than the rule,[16] the question of how the available methods perform in this situation is not a trivial concern.

Despite this, very few studies have investigated the impact of non-normal true study effects on the performance of MA techniques. Research on the subject has been mostly restricted to limited comparisons of newly suggested distribution free methods to widely used parametric ones, under specific conditions (usually assuming normality). Follmann and Proschan[17] proposed a non-parametric RE permutation method that was proved to be comparable, in terms of coverage and power, to the DerSimonian–Laird (DL) estimator, in simulations of normally distributed effects. A non-parametric maximum likelihood approach has also been investigated, in the multi-level generalised linear models context and compared with its parametric counterpart.[18] Aitkin,[19] using non-normal study effects from three multi-centre clinical trials, displayed the model's increased robustness against parametric model misspecification. However, an overall comparison of the widely used DL model and more recently developed methods, using non-normally distributed true study effects has not been performed yet.

Whether or not normality is satisfied, another important factor that affects model performance is the number of studies included in the MA. Simulations of normally distributed effects have shown

that the coverage levels of all MA methods are affected when the number of studies is small, for even moderate degrees of heterogeneity. For meta-analyses of 10 studies coverage was found to be below the nominal level by approximately 5% for the DL method and by 10% for maximum likelihood (ML), while the least affected method was profile likelihood (PL).[11] A solution to the coverage deterioration problem of the DL model, for small numbers of studies, was later proposed by Sidik and Jonkman (SJ).[20] This model is a non-iterative variant of the DL model, based on the *t*-distribution, with higher coverage probabilities than the DL method, especially for small numbers of studies, despite not being computationally intensive. However, the power of the methods was not investigated in any of these simulation studies.

A third factor that has been identified in simulation studies as highly relevant to model performance is heterogeneity among the studies being meta-analysed. Coverage for all models has been found to deteriorate when between-study variance τ^2 increases, especially when the number of studies is small – 10 or below.[11] Since all models calculate and use τ^2, an estimate of the true between-study variance, their performance is closely linked to the accuracy of the estimate. Generally, the larger the true between-study variance the more biased the estimate can be (in absolute value), which leads to a drop-off in method performance.[21] In addition, the τ^2 estimation methods are specified to use population-averaged within study variances when in reality study-specific variances are used instead, a convenient yet high-risk choice that can lead to a considerable bias.[14] Some RE methods provide confidence intervals on the estimate of between-study variance, but only the Biggerstaff–Tweedie[22] (BT) method (a variant of the DL method) builds error in the point estimate of τ^2 into the estimation of the overall effect.

2 Methods

Due to the large number of different MA methods investigated, only a brief overview of each is provided here. Readers should refer to the original papers for full descriptions. Consider a group of k studies with effect size estimates Y_i ($i = 1, 2, \ldots, k$), whose mean effect μ we wish to estimate.

2.1 Fixed-effects

The FE model provides the simplest approach for estimating μ and its error term. It can be defined as:

$$Y_i = \theta_i + e_i, \quad e_i \sim \mathrm{N}\left(0, \sigma_i^2\right) \tag{1}$$

where, for study i: Y_i is the effect size estimate, θ_i the true effect size and e_i the random error. The true study effects are all assumed to be equal ($\theta_i = \mu, \quad i = 1, 2, \ldots, k$) and the only deviations from the true effect are the errors e_i, assumed to be independent and normally distributed with mean zero and variance σ_i^2. The FE estimate of the overall effect μ is usually calculated as a weighted average, using the estimates of the within study variances σ_i^2 as precision weights:[11]

$$\hat{\mu}_F = \frac{\sum_{i=1}^{k} \hat{w}_i Y_i}{\sum_{i=1}^{k} \hat{w}_i} \tag{2}$$

where $\hat{w}_i = 1/\hat{\sigma}_1^2$ and $\mathrm{var}(\hat{\mu}_F) = 1/\sum_{i=1}^{k} \hat{w}_i$.

2.2 Random-effects

Often the homogeneity assumption is unlikely and variation between studies in the size of the true effect is assumed. This is known as the RE model, in which the true effects θ_i are typically assumed to be normally distributed. A second error term is then incorporated into (1), to account for across study variability, and the model becomes:[11]

$$\begin{aligned} Y_i &= \theta_i + e_i, \quad e_i \sim \mathrm{N}\left(0, \sigma_i^2\right) \\ \theta_i &= \mu + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}\left(0, \tau^2\right) \end{aligned} \tag{3}$$

In this case, the overall effect estimate is provided by:

$$\hat{\mu}_R = \frac{\sum_{i=1}^{k} \hat{w}_i' \cdot Y_i}{\sum_{i=1}^{k} \hat{w}_i'} \tag{4}$$

where $\hat{w}_i' = \frac{1}{\tau^2 + \hat{\sigma}_i^2}$ and $V(\hat{\mu}_R) = 1/\sum \hat{w}_i'$

The RE model usually provides wider confidence intervals for the estimate of the overall effect.[23] The variance parameter τ^2, of the between studies error term ε_i, is a measure of the between study heterogeneity. Since τ^2 is rarely – if ever – known, it needs to be estimated.

Parametric RE meta-analysis methods differ principally in the way that τ^2 is estimated. The most widely used RE method is DL[13], which uses a moments based estimator of τ^2 based on the expected value of Cochran's Q-statistic:[24]

$$Q_{\hat{w}} = \sum_{i=1}^{k} \hat{w}_i (Y_i - \hat{\mu})^2 \tag{5}$$

$Q_{\hat{w}}$ follows a χ^2_{k-1} distribution under the null hypothesis of homogeneity. It should be noted that the DL method makes no assumptions regarding the distribution of the effects.

Biggerstaff–Tweedie[22] proposed a variant of the DL method which takes into account variation in the point estimate of τ^2, by weighting the studies according to their relative contribution to the overall distribution of the estimated τ^2. Sidik and Jonkman[20] suggested a simpler DL variant which uses the t-distribution in constructing confidence intervals for the overall effect mean, and offers improved coverage.

In a further variation, a test for heterogeneity based on Cochran's $Q_{\hat{w}}$ is initially applied.[6,25]. If homogeneity is not rejected the FE model is used, otherwise a RE model is applied, by tradition the DL model. In accordance with Brockwell and Gordon,[11] we call this method Q-based (Q).

Alternatives to the DL 'family' of methods above also exist. These include the simple Maximum Likelihood and the Profile Likelihood, both proposed by Hardy and Thompson.[26] The simple maximum likelihood estimates $\hat{\mu}_{ml}$ and $\hat{\tau}^2_{ml}$ can be obtained from the log-likelihood function in (6).

$$\log L\left(\mu, \tau^2\right) = -\frac{1}{2}\sum_{i=1}^{k} \log\left(2\pi\left(\hat{\sigma}_i^2 + \tau^2\right)\right) - \frac{1}{2}\sum_{i=1}^{k} \frac{(y_i - \mu)^2}{\hat{\sigma}_i^2 + \tau^2}, \quad \mu \in \mathbb{R} \quad \tau^2 \geq 0 \tag{6}$$

The PL method calculates asymmetric confidence intervals for the estimates by taking into account the fact that both parameters need to be estimated simultaneously. The Likelihood

methods are iterative and can be computationally intensive, especially PL which involves a two-level maximisation process.[11]

A final non-parametric RE method involves mass permutations of the effects' signs to create a distribution for the overall effect estimates $\hat{\mu}_{pe}$ under the DL random-effects model and the hypothesis that μ is zero.[17] Rejection (or not) of the hypothesis is decided on the position of the observed mean estimate within the distribution.

This article compares the performance of all eight methods outlined above: fixed-effects model (FE) model and DerSimonian & Laird (DL), Biggerstaff & Tweedie (BT), Sidik and Jonkman (SJ), Q-based (Q), maximum-likelihood (ML), profile-likelihood (PL), permutations method on DL (PE) random-effects models.

3 Simulations

Consider a single MA involving k studies. Using the effect size and variability estimates of each study, we applied the MA models and estimated an overall effect size and its confidence interval for each method. This process was repeated for 10 000 meta-analyses and the performance of the models was assessed using three measures that will be described in more detail later: coverage, power and confidence interval performance. The simulations, including all data generation, were programmed in STATA v9.2 for Windows,[27] the only exception being a component of BT. BT requires a Gauss–Laguerre quadrature function for computing integrals, which we coded in MATLAB v7.6 for Windows.[28] The source code is available from the first author.

3.1 Data generation methods

Following the approach taken by Brockwell and Gordon,[11] for each study i within a simulated MA we randomly assigned an effect size estimate Y_i and within-study variance estimate $\hat{\sigma}_i^2$. Although Brockwell and Gordon take Y_i to be a log-odds ratio, it is useful to observe that this is not a necessary assumption, since all subsequent theory and methods apply equally to all effect metrics. The $\hat{\sigma}_i^2$ was assumed to be a realisation from a χ_1^2 distribution, divided by 4 and restricted to the (0.009, 0.6) interval.[11] Values that fell outside this range were redrawn to avoid obtaining a large percentage of extremes. This resulted in a mean within-study variance estimate to 0.173.

The Y_i value for each study i was generated using:

$$Y_i = \theta_i + e_i \tag{7}$$

We assigned e_i on the assumption that $e_i \sim \mathrm{N}(0, \hat{\sigma}_i^2)$, as per the standard model. This was the one assumption of normality that we did not vary in our simulations, as many common effect metrics, including the log-odds ratio, standardised mean difference and risk difference, have an approximate normal distribution.

The true study effects θ_i were generated to have a mean of $\mu = 0.5$ and a variance τ^2. We ran simulations under a variety of values for τ^2, using realistic values based on an analysis by Engels et al.[29] of heterogeneity of effects (odds-ratios) for 125 meta-analyses. Q-test values for each MA were obtained from the authors and converted into H^2 values. H^2 is the heterogeneity measure least affected by the number of studies in the analysis[25] and it can be defined as:

$$H^2 = \frac{Q_{\hat{w}}}{k-1} \tag{8}$$

H^2 has a value of 1 in the case of homogeneity and heterogeneity is assumed to be present when $H^2 > 1$.

The observed H^2 values were compared with those obtained from simulations using various τ^2 values. The very large majority of empirical H^2 values were returned by values of $\tau^2 \in [0.01, 0.1]$. We adopted values for τ^2 of 0.01, 0.03, 0.07 and 0.1 for the main study, with the last three values providing continuity with Brockwell and Gordon's[11] research. Relative to these τ^2 values and this particular distribution of within-study variances, we calculated the expectation of H^2 in each case to be 1.18, 1.54, 2.25 and 2.78, respectively.[6] We also investigated very high heterogeneity ($\tau^2 = 0.5$, $E[H^2] = 9.92$) and the results are available from the authors (not presented in this article).

The use of a χ^2_1 distribution to generate within-study variances results in a predominance of smaller variances and essentially simulates meta-analyses dominated by larger studies. However, meta-analyses encountered in practice vary widely in this respect. Therefore, to represent alternative within-study variance distributions we also generated variances based on (a) a normal distribution and (b) a χ^2_1 'tail' distribution (whereby the χ^2_1 distribution is rotated around its mean), to represent meta-analyses dominated by smaller studies. In both the cases, we again restricted the distributions to the interval (0.009, 0.6) and rescaled to the same final mean of 0.173. For the normal, we used $N(9.1724, 0.036)$ to minimise truncation. Varying the within-study variance distribution, at the same level of heterogeneity, had a very small effect on the results and therefore, we only present results for the χ^2_1 distribution.

We next simulated various forms of underlying distribution in the effects θ_i: unimodal (both normal and skew-normal), bimodal, uniform, beta, 'double-spike' and zero between-study variance.

For unimodal distributions, each true effect θ_i was drawn from a unimodal distribution of positive skewness *sk* and kurtosis *ku*: $\theta_i \sim \mathrm{M}^{ku}_{sk}(\mu, \tau^2)$. Ramberg's[30] method was selected to simulate this distribution as this method can simulate distributions with wider ranges of skewness and kurtosis values than the alternatives.[31,32] We investigated 25 different unimodal distributions based on different combinations of (positive) skew and kurtosis, ranging from skew = 0 and kurtosis = 3 for the normal distribution to the extreme case where skew = 2 and kurtosis = 15. Such asymmetry may not be uncommon in practice: Micceri[16] examined 440 ability and psychometric scale score distributions, many of which were as asymmetric as the most extreme of our generated distributions. The method has been implemented in a STATA module and is available for download from the Statistical Software Components (SSC) archive under the name *sknor*.

Bimodal distributions were simulated by combining two unimodal distributions. A bimodally distributed variable θ_i was generated using:

$$\theta_i \sim \begin{cases} \mathrm{M}^{ku_1}_{sk_1}(\mu_1, \sigma_1^2), & p \\ \mathrm{M}^{ku_2}_{sk_2}(\mu_2, \sigma_2^2), & (1-p) \end{cases} \tag{9}$$

Each θ_i was randomly drawn from one of the two distributions, according to probability p. The desired conditions for the bimodal – mean $\mu = 0.5$ and variance τ^2 fixed to one of the four values previously designated – can be achieved using many different pairs of unimodal distributions. To reduce the options, we restricted examination to pairs of normal distributions of equal probability ($p = 0.5$). Three pairs were simulated: $\sigma_1^2 = \sigma_2^2 = 0.1$; $\sigma_1^2 = \sigma_2^2 = 0.5$; and $\sigma_1^2 = \sigma_2^2 = 0.9$. Results are presented for the first scenario only, which represents the greatest degree of bimodality. The method is available for download from the SSC archive under the name *skbim*.

Finally, we simulated three extreme distributions: uniform, 'U shaped' beta and 'double spike'. All three were adjusted to have mean $\mu = 0.5$ and the desired variance τ^2. 'Double spike' was designed to simulate a situation where study effects can take one of two 'Fixed Effect' values, a plausible scenario when two alternate experimental designs exist for testing an intervention. It can be thought of as a special case of the bimodal when the within study variances of the unimodal distributions are zero ($\sigma_1^2 = \sigma_2^2 = 0$).

Using the methods described in this section, we created datasets of 10 000 size k meta-analyses. The number of studies k varies across datasets from 2 to 35. Each study's effect size and variance were randomly selected according to the methods above, avoiding systematic repetition either within or across meta-analyses. Since the number of simulated datasets was very large (16 014 in total), we only present results for a representative selection. Full results pertaining to all the distributional assumptions are available from the authors.

3.2 Measures of performance

Performance of the MA methods on the simulated datasets is quantified using three measures.

3.2.1 Coverage probability

For each simulated MA case, we calculated confidence intervals for the overall effect estimate $\hat{\mu}$, for all the methods. The coverage probability is the proportion of confidence intervals that contain μ in a sample of 10 000 meta-analyses. This probability should be close to 0.95 for 95% confidence intervals, but it can only be explored through simulations since the distributions of the confidence intervals of the methods are unknown. The arbitrarily selected value of μ does not bear any impact on the process of calculating the coverage probability, for any of the methods.

3.2.2 Power probability

A power comparison of the various models is important if we are to evaluate their performance in every respect.[33] Overton[34] compared the FE to the RE model using both Type I error rates and power probabilities. He concluded that in many cases the RE model over-performed in terms of the Type I error but provided extremely wide confidence intervals, understating the information actually gained from the MA and inflating Type II error rates. Follmann and Proschan[17] performed a power analysis for the permutation method but only relative to a t-test. We proceeded with an overall power comparison of the methods, on an absolute basis for μ.

Although for most of the methods a power probability could be computed for each MA case, we decided to calculate power in a manner similar to coverage for consistency. In each set of 10 000 meta-analyses, power was denoted as the proportion of confidence intervals that did not include zero. This approach was applied to all methods.

We computed power for a range of non-zero mean true effect sizes. To produce results directly comparable across the different distribution shapes and degrees of between-study variance, we examined power relative to centile values, rather than absolute values. For example, we calculated the power to detect a true mean effect equivalent to the 25th percentile of the population distribution of effect sizes. We produced results for a range of centiles and found the pattern of results to be essentially the same in all cases, therefore we report only on the 25th centile here, as this gave a good spread of power relative to sample size.

3.2.3 Overall effect estimation (point and error interval estimation)

The ability of each method to return an accurate point estimate of the true overall effect ($\mu = 0.5$) under each set of conditions was examined by computing mean overall effect estimates for each sample of 10 000 meta-analyses.

We assessed performance with respect to estimation of the error interval (95% confidence interval) around the point estimates by computing the median, across each 10 000 sample, of:

$$P_{CI} = \frac{upperCI(\hat{\mu}) - lowerCI(\hat{\mu})}{3.92\sqrt{\text{var}}} \tag{10}$$

where $\text{var} = (\sum_{i=1}^{k} \frac{1}{\tau^2 + \hat{\sigma}_i^2})^{-1}$ and $upperCI(\hat{\mu})$, $lowerCI(\hat{\mu})$ are the upper and lower bounds of the 95% confidence interval for the estimated effect.

Thus, performance is measured as a percentage of over or under-estimation in the derived confidence interval compared to the interval based on the true between-study variance. This approach differs from that of other authors, who typically have assessed variance performance in terms of the accuracy with which $\hat{\tau}^2$ is estimated.[21] We had several reasons for preferring to focus on the estimated confidence interval around the overall effect. The between-variance estimate for the PL and ML methods is the same but PL 'amends' its computed confidence intervals to take the uncertainty of $\hat{\tau}^2$ into account. Also, between-variance estimates for PL and PE methods can be misleading since the methods provide asymmetric confidence intervals. Finally, the confidence interval has more direct bearing on how the results of any particular meta-analytic study are interpreted.

4 Results

Necessarily, we cannot present detailed results for all of the combinations of method, effect distribution, study numbers and heterogeneity (H^2) that we investigated. In tables and figures, we have therefore omitted certain results.

We investigated models using three different within-study variance distributions, but we only report results for the χ_1^2 distribution since differences were small. We examined a range of skewed study effect distributions but restrict our reporting to 'moderate' non-normality (skew = 1, kurtosis = 4) and severe non-normality (skew = 2, kurtosis = 9): results for the latter were similar to those for distributions with even higher levels of skew and/or kurtosis. We do not report on the extreme beta distribution, as results were all very close to those for the bimodal distribution. We also omit results for $H^2 = 2025$ from the tables, since the main findings are adequately represented by the results for the remaining H^2 values.

The tables report mean results for meta-analyses of size 2–5 studies, 6–15, 16–25 and 26+ studies. Inspection of results indicated that this division captured all the main differences.

4.1 Coverage

4.1.1 Zero between-study variance

When between-study variance was zero, the FE and Q models gave 95% coverage at all study numbers. Most other models gave slightly inflated coverage, though rarely exceeding 96%. PL was the poorest performing model, returning an average coverage of 98% for small meta-analyses ($k \leq 5$) and 97% for larger numbers of studies (Table 1 and Figure 1).

Table 1. Coverage performance by degree of heterogeneity, between-study effect distribution, and MA size, assuming χ^2_1-based within-study variances

		Number of studies															
		2–5								6–15							
H^2	θ_i distribution (skew, kurtosis)	FE	DL	BT	SJ	Q	ML	PL	PE	FE	DL	BT	SJ	Q	ML	PL	PE
1	None	0.95	0.96	0.96	0.95	0.95	0.95	0.98	–	0.95	0.96	0.96	0.95	0.95	0.96	0.97	0.95
1.18	Normal (0,3)	0.91	0.94	0.94	0.94	0.92	0.92	0.97	–	0.90	0.94	0.94	0.94	0.91	0.93	0.95	0.95
1.18	Skew-normal (1,4)	0.91	0.94	0.94	0.94	0.92	0.93	0.97	–	0.90	0.94	0.94	0.94	0.91	0.93	0.96	0.95
1.18	Skew-normal (2,9)	0.91	0.95	0.94	0.94	0.92	0.93	0.97	–	0.90	0.94	0.95	0.94	0.92	0.93	0.96	0.95
1.18	Uniform	0.91	0.94	0.94	0.94	0.92	0.92	0.97	–	0.90	0.94	0.94	0.93	0.91	0.92	0.95	0.95
1.18	Bimodal	0.91	0.94	0.94	0.94	0.92	0.92	0.97	–	0.90	0.94	0.94	0.94	0.91	0.92	0.95	0.95
1.18	D-spike	0.91	0.94	0.94	0.94	0.92	0.92	0.97	–	0.90	0.94	0.94	0.94	0.91	0.92	0.95	0.95
1.54	Normal (0,3)	0.85	0.91	0.91	0.94	0.87	0.88	0.95	–	0.82	0.92	0.92	0.93	0.87	0.90	0.94	0.95
1.54	Skew-normal (1,4)	0.85	0.91	0.91	0.93	0.88	0.89	0.95	–	0.82	0.92	0.92	0.93	0.87	0.90	0.94	0.95
1.54	Skew-normal (2,9)	0.86	0.92	0.92	0.94	0.89	0.89	0.96	–	0.83	0.92	0.92	0.93	0.88	0.91	0.94	0.95
1.54	Uniform	0.84	0.90	0.90	0.93	0.86	0.87	0.95	–	0.81	0.91	0.91	0.92	0.86	0.89	0.94	0.95
1.54	Bimodal	0.84	0.90	0.90	0.93	0.86	0.87	0.95	–	0.81	0.91	0.91	0.92	0.85	0.88	0.93	0.95
1.54	D-spike	0.84	0.90	0.90	0.93	0.86	0.87	0.95	–	0.81	0.90	0.90	0.91	0.85	0.88	0.92	0.95
2.78	Normal (0,3)	0.71	0.87	0.85	0.92	0.80	0.81	0.91	–	0.64	0.90	0.90	0.92	0.85	0.87	0.92	0.95
2.78	Skew-normal (1,4)	0.71	0.86	0.85	0.92	0.80	0.81	0.91	–	0.64	0.90	0.89	0.92	0.85	0.87	0.92	0.95
2.78	Skew-normal (2,9)	0.74	0.88	0.87	0.93	0.82	0.83	0.92	–	0.67	0.89	0.89	0.92	0.84	0.86	0.91	0.95
2.78	Uniform	0.68	0.85	0.83	0.92	0.78	0.79	0.90	–	0.62	0.89	0.89	0.92	0.84	0.87	0.92	0.95
2.78	Bimodal	0.66	0.83	0.81	0.91	0.75	0.76	0.89	–	0.61	0.88	0.88	0.91	0.83	0.85	0.91	0.95
2.78	D-spike	0.65	0.81	0.79	0.90	0.73	0.74	0.88	–	0.60	0.87	0.87	0.91	0.81	0.84	0.90	0.95
		16–25								26+							
1	None	0.95	0.96	0.96	0.95	0.95	0.96	0.97	0.95	0.95	0.96	0.96	0.95	0.95	0.96	0.97	0.95
1.18	Normal (0,3)	0.90	0.94	0.94	0.94	0.91	0.93	0.95	0.95	0.90	0.94	0.94	0.94	0.91	0.93	0.95	0.95
1.18	Skew-normal (1,4)	0.90	0.94	0.94	0.94	0.91	0.93	0.95	0.95	0.90	0.94	0.94	0.94	0.91	0.93	0.95	0.95
1.18	Skew-normal (2,9)	0.90	0.94	0.94	0.94	0.92	0.93	0.95	0.95	0.90	0.94	0.94	0.94	0.92	0.94	0.95	0.95
1.18	Uniform	0.90	0.94	0.94	0.94	0.91	0.93	0.95	0.95	0.90	0.94	0.94	0.94	0.91	0.93	0.95	0.95
1.18	Bimodal	0.90	0.94	0.94	0.94	0.91	0.93	0.95	0.95	0.90	0.94	0.94	0.94	0.91	0.93	0.95	0.95
1.18	D-spike	0.90	0.94	0.94	0.94	0.91	0.93	0.95	0.95	0.90	0.94	0.94	0.94	0.91	0.93	0.95	0.95
1.54	Normal (0,3)	0.81	0.92	0.92	0.93	0.88	0.91	0.94	0.95	0.80	0.93	0.93	0.93	0.89	0.92	0.94	0.95
1.54	Skew-normal (1,4)	0.81	0.92	0.92	0.93	0.88	0.91	0.94	0.95	0.81	0.93	0.93	0.93	0.89	0.92	0.94	0.95
1.54	Skew-normal (2,9)	0.82	0.93	0.93	0.93	0.89	0.91	0.94	0.95	0.81	0.93	0.93	0.93	0.90	0.92	0.94	0.94
1.54	Uniform	0.80	0.92	0.92	0.93	0.87	0.91	0.94	0.95	0.80	0.93	0.93	0.93	0.88	0.92	0.94	0.95
1.54	Bimodal	0.80	0.92	0.92	0.93	0.87	0.91	0.93	0.95	0.80	0.92	0.92	0.93	0.88	0.92	0.94	0.95
1.54	D-spike	0.80	0.91	0.92	0.92	0.86	0.90	0.93	0.95	0.80	0.92	0.92	0.93	0.88	0.92	0.94	0.95
2.78	Normal (0,3)	0.62	0.92	0.92	0.94	0.90	0.91	0.93	0.95	0.61	0.93	0.93	0.94	0.92	0.93	0.94	0.95
2.78	Skew-normal (1,4)	0.62	0.92	0.91	0.93	0.89	0.90	0.93	0.94	0.61	0.93	0.93	0.94	0.92	0.92	0.94	0.94
2.78	Skew-normal (2,9)	0.63	0.91	0.91	0.92	0.88	0.89	0.92	0.94	0.62	0.92	0.92	0.93	0.90	0.91	0.93	0.94
2.78	Uniform	0.61	0.92	0.92	0.94	0.90	0.91	0.93	0.95	0.60	0.93	0.93	0.94	0.92	0.92	0.94	0.95
2.78	Bimodal	0.60	0.92	0.91	0.93	0.90	0.91	0.93	0.95	0.59	0.93	0.93	0.94	0.92	0.92	0.94	0.95
2.78	D-spike	0.60	0.92	0.91	0.93	0.90	0.91	0.93	0.95	0.59	0.93	0.92	0.94	0.92	0.92	0.94	0.95

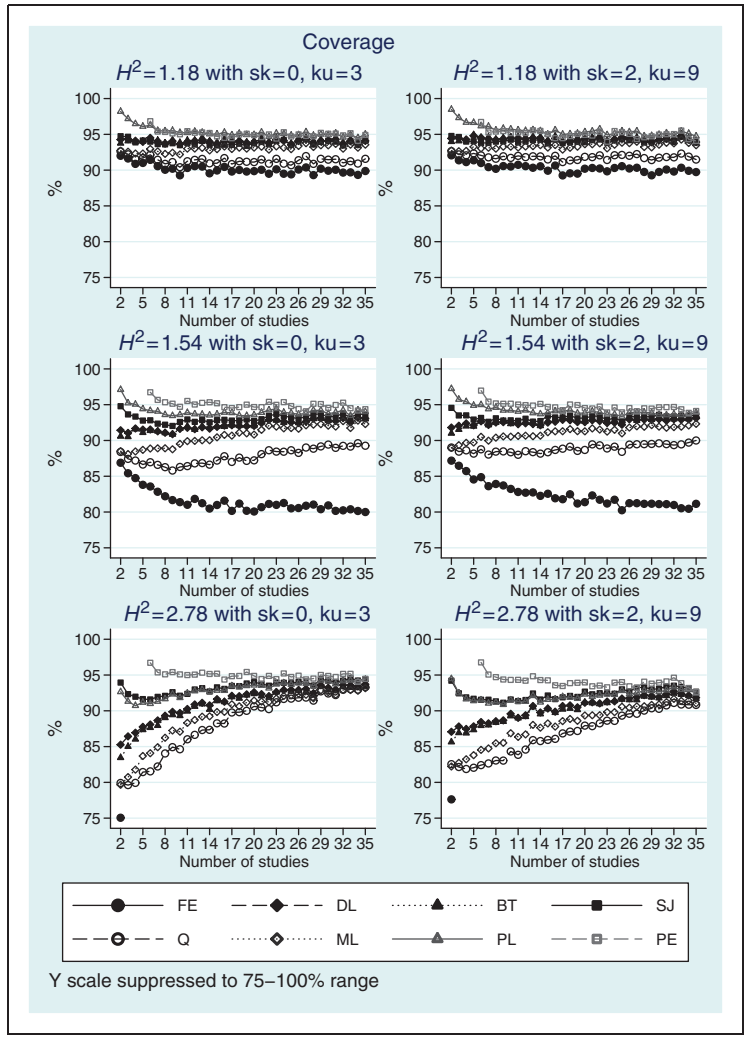

Figure 1. Coverage at $H^2 = 1.18$, 1.54 and 2.78 for eight MA methods, under assumptions of χ^2_1 based within-study variances and normal and highly skewed between-study variance distributions.

4.1.2 Non-zero between-study variance

The introduction of between-study variation changed the results considerably.

Normally distributed study effects. With a small amount of between-study heterogeneity ($H^2 = 1.18$), FE and Q gave the lowest coverage ($p \leq 0.92$) at all study sizes. ML performed slightly better with coverage up to 0.93 for large meta-analyses, while coverage for DL, BT and SJ was consistently 0.94. PL and PE were the best performers at most study sizes, returning a coverage of 0.95 for all but the lowest numbers of studies.

When a larger amount of between study variation was introduced, FE coverage performance became very poor ($p \leq 0.85$), increasingly so as heterogeneity or study numbers increased. Q and ML provided coverage that rarely exceeded 0.91, while coverage using DL was only slightly better.

BT performed at a similar level to DL. PE was the best performer, yielding 95% coverage at all study numbers and levels of H^2. PL and SJ performed at a level somewhere between DL and PE.

Skew-normal effects. At all levels of skew in the effects distribution, the coverage of PE remained close to 95%. The impact on other methods was also mostly minor, with changes of at most 1% or 2% relative to the findings using normally distributed effects, for all degrees of heterogeneity and size of MA. As before, PE consistently gave the best coverage, followed by PL/SJ and then DL/BT.

'Extreme' non-normal effects. Results for the extreme distributions closely mirrored those for the normal and skewed distributions. PE consistently provided accurate coverage while the remaining methods lost coverage – to varying degrees – as heterogeneity increased. Remarkably, the pattern of results with each method was very much the same regardless of whether the extreme distribution was uniform, bimodal, 'double-spiked' or 'U-shaped' (not shown in table). Among the 'extreme' distributions, coverage was best with the uniform distribution and poorest with the 'double-spike', although the difference rarely exceeded 2%. Our expectation was that FE would perform much better with the 'double-spike' than with other non-normal distributions, but this turned out not to be the case.

Use of within-study variances based on the χ^2_1-tail and normal distributions made very little difference to the results: at all levels of H^2 and study numbers, coverage values only occasionally deviated by more than 1% from those obtained using χ^2_1 distributed variances.

4.2 Power

Under the assumption of χ^2_1 distributed within-study variances, FE demonstrated the highest level of power across the board. However, this result was a function of the fact that this method had such poor coverage: the considerably inflated risk of false positives with this method (under most conditions) necessarily resulted in an 'apparent' smaller number of false negatives. To a much smaller degree, this factor also affected the other methods, for example the Q and ML methods also tended to have higher power but lower coverage. Power for BT was similar, though in some cases a little lower, than that for DL. Purely in terms of power, ML consistently outperformed DL and BT, which in turn outperformed PL, SJ and PE. PE had the lowest power across all scenarios (Table 2 and Figure 2).

Skewness in the effects distribution acted to reduce the power of all methods when study numbers were small ($k \leq 5$), but acted to increase power when both study numbers and heterogeneity were high, and on the whole power was better with the extreme distributions than with skew.

Power when using χ^2_1-tail and normal distributions for the within-study variances was lower than for the χ^2_1 distribution, although the power performance of the methods relative to one another remained much the same.

4.3 Overall effect estimation – point estimates

Using a χ^2_1 distribution for the within-study variances and a 'true' overall effect size in the simulations of 0.5, all methods returned a mean overall effect correct to within 0.01 (i.e. 0.49–0.51), for all types of effects distribution and numbers of studies. Median effect sizes were also correct to the same degree, except when skew was present in the distribution of effects and H^2 was 2.25 or above, but even in this case the median never dipped below 0.47.

The degrees of bias on the mean and median estimates of effect were of a similar small order using alternative distributions for the within-study variances.

Table 2. Power (at the 25th percentile) by degree of heterogeneity, between-study effect distribution and MA size, assuming χ^2_1 based within-study variances

		Number of studies															
		2–5								6–15							
H^2	θ_i distribution (skew, kurtosis)	FE	DL	BT	SJ	Q	ML	PL	PE	FE	DL	BT	SJ	Q	ML	PL	PE
1	None	0.36	0.29	0.29	0.20	0.35	0.34	0.20	–	0.75	0.67	0.67	0.62	0.73	0.72	0.63	0.51
1.18	Normal (0,3)	0.39	0.29	0.30	0.19	0.36	0.35	0.21	–	0.75	0.65	0.64	0.58	0.71	0.69	0.60	0.48
1.18	Skew-normal (1,4)	0.38	0.28	0.28	0.18	0.35	0.34	0.20	–	0.76	0.65	0.64	0.59	0.72	0.69	0.60	0.49
1.18	Skew-normal (2,9)	0.37	0.28	0.28	0.18	0.35	0.34	0.20	–	0.75	0.65	0.64	0.59	0.71	0.69	0.60	0.49
1.18	Uniform	0.38	0.28	0.29	0.19	0.35	0.34	0.21	–	0.75	0.65	0.64	0.58	0.71	0.69	0.60	0.48
1.18	Bimodal	0.38	0.29	0.29	0.19	0.36	0.35	0.21	–	0.75	0.65	0.64	0.58	0.71	0.69	0.60	0.48
1.18	D-spike	0.38	0.29	0.29	0.18	0.35	0.35	0.21	–	0.75	0.64	0.64	0.58	0.71	0.69	0.60	0.48
1.54	Normal (0,3)	0.45	0.32	0.33	0.19	0.40	0.39	0.24	–	0.79	0.65	0.64	0.57	0.71	0.69	0.60	0.48
1.54	Skew-normal (1,4)	0.43	0.30	0.31	0.18	0.38	0.37	0.23	–	0.79	0.65	0.64	0.57	0.71	0.70	0.60	0.49
1.54	Skew-normal (2,9)	0.41	0.28	0.29	0.17	0.35	0.35	0.21	–	0.77	0.63	0.62	0.55	0.69	0.68	0.58	0.48
1.54	Uniform	0.44	0.31	0.32	0.18	0.38	0.38	0.23	–	0.78	0.64	0.63	0.57	0.70	0.69	0.60	0.48
1.54	Bimodal	0.46	0.33	0.33	0.19	0.40	0.39	0.25	–	0.80	0.67	0.66	0.59	0.73	0.71	0.62	0.50
1.54	D-spike	0.45	0.32	0.33	0.19	0.39	0.39	0.24	–	0.80	0.67	0.65	0.59	0.72	0.71	0.62	0.50
2.78	Normal (0,3)	0.55	0.35	0.36	0.19	0.43	0.43	0.28	–	0.83	0.64	0.61	0.55	0.67	0.68	0.59	0.48
2.78	Skew-normal (1,4)	0.52	0.32	0.32	0.16	0.39	0.40	0.25	–	0.83	0.63	0.60	0.54	0.67	0.68	0.59	0.49
2.78	Skew-normal (2,9)	0.49	0.30	0.30	0.15	0.37	0.38	0.22	–	0.82	0.63	0.60	0.53	0.67	0.68	0.59	0.50
2.78	Uniform	0.56	0.36	0.37	0.19	0.43	0.44	0.29	–	0.85	0.66	0.63	0.57	0.69	0.70	0.62	0.50
2.78	Bimodal	0.56	0.36	0.36	0.19	0.42	0.43	0.29	–	0.86	0.67	0.64	0.58	0.70	0.71	0.63	0.52
2.78	D-spike	0.55	0.34	0.34	0.18	0.40	0.42	0.28	–	0.85	0.67	0.63	0.57	0.69	0.70	0.62	0.51
		16–25								26+							
1	None	0.96	0.92	0.92	0.91	0.95	0.94	0.91	0.88	0.99	0.99	0.99	0.98	0.99	0.99	0.98	0.98
1.18	Normal (0,3)	0.95	0.90	0.90	0.88	0.93	0.92	0.88	0.86	0.99	0.98	0.98	0.97	0.98	0.98	0.97	0.97
1.18	Skew-normal (1,4)	0.95	0.90	0.91	0.88	0.93	0.92	0.89	0.86	0.99	0.98	0.98	0.97	0.99	0.98	0.98	0.97
1.18	Skew-normal (2,9)	0.95	0.91	0.91	0.89	0.93	0.93	0.89	0.87	0.99	0.98	0.98	0.98	0.99	0.99	0.98	0.97
1.18	Uniform	0.95	0.90	0.90	0.88	0.93	0.92	0.88	0.86	0.99	0.98	0.98	0.97	0.99	0.98	0.97	0.97
1.18	Bimodal	0.95	0.90	0.90	0.88	0.93	0.92	0.88	0.86	0.99	0.98	0.98	0.97	0.98	0.98	0.97	0.97
1.18	D-spike	0.95	0.90	0.90	0.88	0.93	0.92	0.88	0.86	0.99	0.98	0.98	0.97	0.98	0.98	0.97	0.97
1.54	Normal (0,3)	0.95	0.90	0.89	0.87	0.91	0.91	0.88	0.85	0.99	0.97	0.97	0.97	0.98	0.98	0.97	0.96
1.54	Skew-normal (1,4)	0.96	0.91	0.91	0.88	0.92	0.92	0.89	0.86	0.99	0.98	0.98	0.97	0.98	0.98	0.97	0.97
1.54	Skew-normal (2,9)	0.95	0.89	0.89	0.86	0.91	0.91	0.87	0.85	0.99	0.98	0.98	0.97	0.98	0.98	0.97	0.97
1.54	Uniform	0.95	0.89	0.89	0.87	0.91	0.91	0.88	0.85	0.99	0.98	0.98	0.97	0.98	0.98	0.97	0.97
1.54	Bimodal	0.96	0.91	0.91	0.89	0.93	0.92	0.90	0.87	0.99	0.98	0.98	0.98	0.99	0.99	0.98	0.98
1.54	D-spike	0.96	0.91	0.91	0.89	0.93	0.93	0.90	0.87	0.99	0.98	0.98	0.98	0.99	0.99	0.98	0.98
2.78	Normal (0,3)	0.96	0.87	0.86	0.84	0.88	0.88	0.86	0.83	0.99	0.96	0.96	0.96	0.96	0.97	0.96	0.96
2.78	Skew-normal (1,4)	0.97	0.90	0.89	0.87	0.90	0.91	0.88	0.86	0.99	0.98	0.98	0.97	0.98	0.98	0.97	0.97
2.78	Skew-normal (2,9)	0.97	0.90	0.89	0.87	0.91	0.92	0.89	0.87	1.00	0.98	0.98	0.97	0.98	0.98	0.98	0.97
2.78	Uniform	0.97	0.91	0.90	0.88	0.91	0.92	0.89	0.87	0.99	0.98	0.98	0.97	0.98	0.98	0.98	0.97
2.78	Bimodal	0.98	0.92	0.91	0.90	0.92	0.93	0.91	0.89	1.00	0.99	0.99	0.98	0.99	0.99	0.98	0.98
2.78	D-spike	0.98	0.92	0.91	0.90	0.92	0.93	0.91	0.89	1.00	0.99	0.99	0.98	0.99	0.99	0.98	0.98

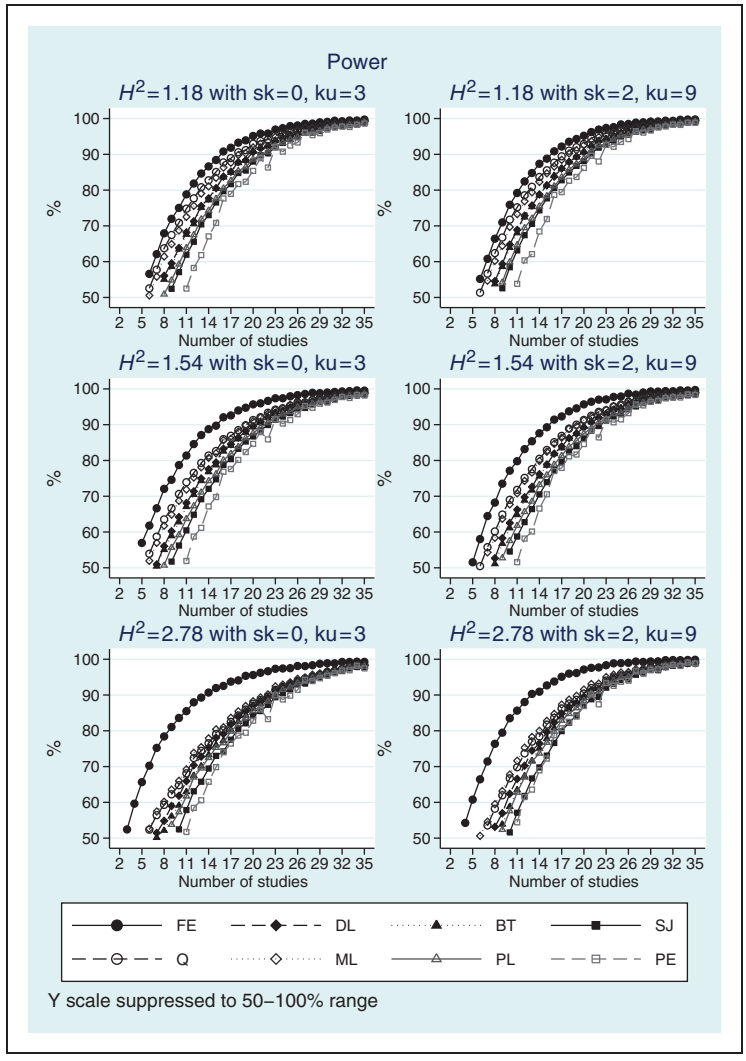

Figure 2. Power at $H^2 = 1.18$, 1.54 and 2.78 for eight MA methods, under assumptions of χ^2_1 based within-study variances and normal and highly skewed between-study variance distributions.

4.4 Overall effect estimation – error intervals

In the case of χ^2_1 based within-study variances – for all types of effects distribution bar 'double-spike' – all the methods except PL, SJ and PE produced confidence intervals around the overall effect estimate that were biased downwards. Underestimation of the confidence interval was particularly severe using FE and, in most cases, Q. ML also performed very badly under conditions of small to moderate study numbers ($k \leq 15$) and moderate to high heterogeneity ($H^2 \geq 1.54$), producing confidence intervals as narrow as 70% of the correct interval. DL and BT produced intervals that were usually within 5% of the actual interval, except in cases of very small study numbers or very high skew. In contrast, PL, SJ and PE tended to overestimate the confidence interval. The overestimation using PE was very severe for $k \leq 15$ – usually two to three times the width of the

Table 3. Confidence Interval performance by degree of heterogeneity, between-study effect distribution and MA size, assuming χ^2_1 based within-study variances

		Number of studies															
		2–0035								6–15							
H^2	θ_i distribution (skew, kurtosis)	FE	DL	BT	SJ	Q	ML	PL	PE	FE	DL	BT	SJ	Q	ML	PL	PE
1	None	1.00	1.00	1.00	2.25	1.00	1.00	1.28	–	1.00	1.00	1.00	1.12	1.00	1.00	1.12	2.62
1.18	Normal (0,3)	0.90	0.96	0.95	2.07	0.91	0.92	1.16	–	0.86	0.95	0.95	1.08	0.87	0.90	1.05	2.52
1.18	Skew-normal (1,4)	0.91	0.96	0.95	2.10	0.91	0.92	1.16	–	0.86	0.95	0.95	1.08	0.87	0.90	1.04	2.56
1.18	Skew-normal (2,9)	0.90	0.96	0.95	2.11	0.91	0.92	1.16	–	0.86	0.95	0.94	1.07	0.87	0.90	1.04	2.54
1.18	Uniform	0.91	0.96	0.95	2.10	0.91	0.92	1.16	–	0.86	0.95	0.95	1.08	0.87	0.90	1.05	2.53
1.18	Bimodal	0.90	0.96	0.95	2.09	0.91	0.92	1.16	–	0.86	0.95	0.95	1.08	0.87	0.91	1.05	2.54
1.18	D-spike	1.00	1.00	1.00	2.36	1.00	1.00	1.31	–	1.00	1.07	1.06	1.25	1.00	1.00	1.22	2.94
1.54	Normal (0,3)	0.78	0.91	0.89	2.03	0.81	0.84	1.07	–	0.72	0.94	0.93	1.08	0.77	0.86	1.02	2.43
1.54	Skew-normal (1,4)	0.78	0.91	0.89	2.05	0.81	0.84	1.06	–	0.72	0.94	0.93	1.08	0.77	0.85	1.01	2.46
1.54	Skew-normal (2,9)	0.78	0.90	0.89	1.99	0.81	0.83	1.05	–	0.72	0.92	0.91	1.05	0.76	0.83	0.99	2.43
1.54	Uniform	0.79	0.91	0.90	2.03	0.82	0.84	1.07	–	0.72	0.95	0.94	1.09	0.77	0.86	1.03	2.49
1.54	Bimodal	0.78	0.91	0.90	2.04	0.82	0.84	1.07	–	0.72	0.95	0.94	1.09	0.77	0.87	1.03	2.55
1.54	D-spike	1.00	1.06	1.02	2.62	1.00	1.00	1.39	–	1.00	1.30	1.28	1.49	1.00	1.14	1.41	3.48
2.78	Normal (0,3)	0.59	0.85	0.82	2.05	0.69	0.73	0.98	–	0.53	0.94	0.96	1.09	0.88	0.87	1.01	2.16
2.78	Skew-normal (1,4)	0.59	0.84	0.81	2.00	0.69	0.72	0.96	–	0.53	0.93	0.94	1.08	0.86	0.86	1.00	2.22
2.78	Skew-normal (2,9)	0.59	0.81	0.78	1.89	0.68	0.70	0.93	–	0.53	0.88	0.89	1.03	0.80	0.81	0.95	2.20
2.78	Uniform	0.59	0.86	0.84	2.10	0.70	0.74	0.99	–	0.53	0.96	0.98	1.10	0.91	0.89	1.03	2.29
2.78	Bimodal	0.59	0.87	0.84	2.08	0.70	0.74	0.99	–	0.53	0.97	1.00	1.11	0.93	0.90	1.03	2.45
2.78	D-spike	1.00	1.36	1.27	3.36	1.00	1.07	1.65	–	1.00	1.80	1.83	2.06	1.72	1.67	1.92	4.65
		16–25								26+							
1	None	1.00	1.00	1.00	1.05	1.00	1.00	1.08	1.43	1.00	1.00	1.00	1.03	1.00	1.00	1.08	1.24
1.18	Normal (0,3)	0.85	0.97	0.96	1.03	0.86	0.91	1.03	1.40	0.85	0.98	0.98	1.02	0.86	0.93	1.02	1.23
1.18	Skew-normal (1,4)	0.85	0.97	0.96	1.03	0.86	0.91	1.03	1.41	0.85	0.98	0.98	1.02	0.86	0.93	1.02	1.23
1.18	Skew-normal (2,9)	0.85	0.96	0.96	1.02	0.86	0.91	1.02	1.40	0.85	0.97	0.97	1.01	0.86	0.92	1.01	1.22
1.18	Uniform	0.85	0.97	0.97	1.03	0.86	0.91	1.03	1.41	0.85	0.98	0.98	1.02	0.86	0.94	1.03	1.23
1.18	Bimodal	0.85	0.97	0.97	1.03	0.86	0.92	1.04	1.42	0.85	0.98	0.98	1.02	0.86	0.94	1.03	1.23
1.18	D-spike	1.00	1.13	1.13	1.20	1.00	1.05	1.21	1.66	1.00	1.15	1.15	1.20	1.00	1.10	1.21	1.45
1.54	Normal (0,3)	0.71	0.97	0.98	1.03	0.77	0.92	1.02	1.37	0.71	0.98	0.98	1.02	0.86	0.95	1.01	1.20
1.54	Skew-normal (1,4)	0.71	0.97	0.97	1.03	0.77	0.91	1.01	1.37	0.71	0.98	0.98	1.02	0.85	0.94	1.01	1.20
1.54	Skew-normal (2,9)	0.71	0.95	0.95	1.01	0.76	0.89	0.99	1.36	0.71	0.96	0.96	1.00	0.78	0.92	0.99	1.19
1.54	Uniform	0.71	0.98	0.98	1.03	0.77	0.93	1.03	1.39	0.71	0.99	0.99	1.02	0.89	0.96	1.02	1.21
1.54	Bimodal	0.71	0.98	0.99	1.04	0.78	0.94	1.03	1.40	0.71	0.99	0.99	1.02	0.89	0.96	1.02	1.22
1.54	D-spike	1.00	1.37	1.37	1.45	1.00	1.31	1.44	1.98	1.00	1.39	1.39	1.44	1.23	1.35	1.43	1.72
2.78	Normal (0,3)	0.53	0.97	0.98	1.03	0.97	0.94	1.01	1.26	0.52	0.98	0.98	1.02	0.98	0.96	1.01	1.12
2.78	Skew-normal (1,4)	0.53	0.96	0.97	1.03	0.96	0.93	1.00	1.28	0.52	0.97	0.97	1.02	0.97	0.96	1.00	1.14
2.78	Skew-normal (2,9)	0.53	0.92	0.93	1.00	0.92	0.90	0.97	1.27	0.52	0.94	0.95	0.99	0.94	0.93	0.97	1.13
2.78	Uniform	0.53	0.98	1.00	1.04	0.98	0.95	1.02	1.29	0.52	0.99	1.00	1.03	0.99	0.97	1.01	1.14
2.78	Bimodal	0.53	0.99	1.00	1.05	0.99	0.96	1.02	1.32	0.52	0.99	1.00	1.03	0.99	0.97	1.01	1.15
2.78	D-spike	1.00	1.87	1.90	1.98	1.87	1.82	1.94	2.52	1.00	1.89	1.90	1.96	1.89	1.85	1.94	2.20

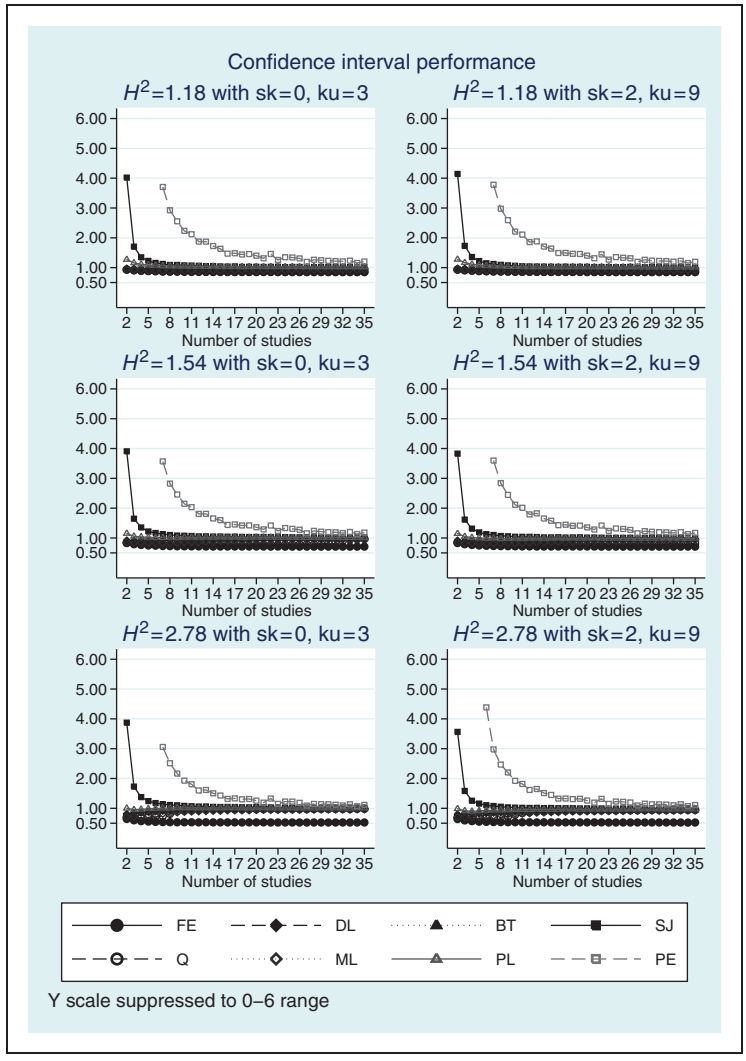

Figure 3. Confidence interval performance at $H^2 = 1.18$, 1.54 and 2.78 for eight MA methods, under assumptions of χ^2_1 based within-study variances and normal and highly skewed between-study variance distributions.

correct interval – and still quite severe even for the largest meta-analyses. SJ also overestimated the confidence interval by a factor of two or more when $k \leq 15$, though the method's performance improved considerably as the number of studies increased. PL produced intervals nearly always correct to within 5% for all but the smallest meta-analyses. However, PL overestimated the confidence intervals by a considerable margin, 16% or more, in small studies with low heterogeneity. The one exception to all the above was the 'double-spike' effects distribution, for which only FE consistently produced accurate confidence interval estimates (Table 3 and Figure 3).

Results were very much the same when we used alternative distributions for the within-study variances, though with generally smaller biases on the confidence interval estimates.

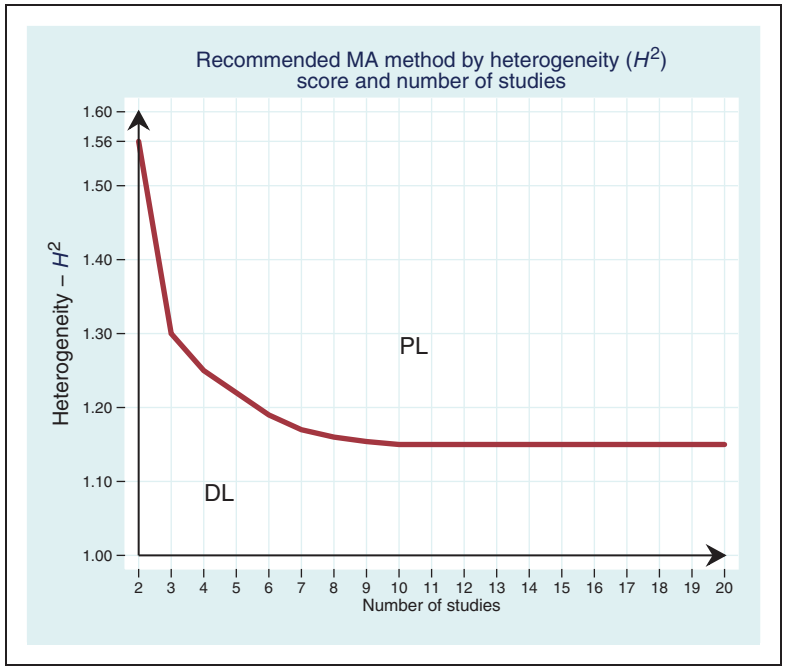

Figure 4. Recommended MA method by degree of heterogeneity and MA size. Notes: PL, Profile likelihood; DL, DerSimonian–Laird.

5 Conclusions

The present investigation has compared eight different frequentist MA methods in terms of coverage, power and effect size estimation (point estimates and error intervals). We found some differences in performance between methods, but we also found that for any particular method, results were highly consistent across different underlying effect size distributions, ranging from normal, through extreme skewness and kurtosis, to uniform, bimodal and 'U-shaped' distributions. Only the 'double-spike' distribution deviated from this pattern, and only in respect to the confidence intervals returned by this method. This is an important finding that can give researchers confidence that, whichever method they adopt, results are highly robust against even very severe violations of the assumption of normally distributed effect sizes.

We also found that results were robust against different assumptions about the distribution of the study-specific variances around the effect estimates. At any given level of heterogeneity, coverage was quite unaffected and error interval estimation only marginally altered, by the shape of the distribution (either high positive, zero or high negative skew), although high negative skew did reduce the power of all methods. In addition, while our simulations made assumptions about distribution shapes, these were not tied to any specific effect metric, such as the log-odds ratio, risk difference or standardised mean difference. Thus, our results apply regardless of the particular effect metric in use.

Selection of an optimum method for any application depends upon balancing the factors of number of studies, coverage, power and effect estimation. If it is reasonable to assume that the effect size does not vary between studies, the FE, Q and ML methods all provide accurate coverage coupled with good power and confidence interval estimation. In this situation, the other methods

generally provide overly high coverage, and correspondingly lower power, particularly with small samples.

However, for most MA applications, zero between-study variance is likely to be the exception rather than the norm, and the presence of even a small amount of between-study variance alters the picture considerably. FE, Q and ML quickly lose coverage as heterogeneity increases. DL rapidly goes from providing a coverage that is slightly high, to one that is overly low. These methods still return accurate point estimates of the overall effect, but the error intervals are often considerably underestimated, although less so for DL than for other methods. In the presence of heterogeneity, PL provides the most accurate coverage, though confidence intervals tend to be overly wide. PE provides exact coverage, regardless of study numbers, but the method is let down by highly inflated confidence intervals. It is also of interest to note that BT and SJ, despite having been expressly developed to address the limitations of other methods in the estimation of between-study variance, were frequently outperformed by those methods.

The situation of very small study numbers – five or fewer – needs special consideration. In this situation, results are much more sensitive to both choice of method and degree of heterogeneity. While the inaccuracy associated with the use of Q or ML – or even FE under a false assumption of a common effect size – may be tolerably small for a large MA, for a small analysis it is far less so. The only truly acceptable methods in this situation are DL and PL. Furthermore, DL only outperforms PL when heterogeneity is low and PL itself struggles to provide coverage much better than 90% when heterogeneity is high.

In the presence of between-study variation, if priority is given to maintaining an accurate Type I error rate then PE is the best method. Usually, however, researchers are as much – if not more – concerned with obtaining an accurate estimate of overall effect size and the associated error interval, as they are with simply testing for statistical significance. In this case, in most situations PL provides the best combination of effect (point and error) estimation and coverage. However, there are some exceptions to this, particularly when heterogeneity is low, where DL performs better. On the basis of our detailed results, Figure 4 presents our recommendations for which method provides the best overall mix of coverage and error estimation under each combination of heterogeneity and study numbers. For meta-analyses with 10 or more studies PL would be our preferred method when $H^2 \geq 1.15$, and DL when $H^2 < 1.15$; but the threshold H^2 value increases as the number of studies in the analysis declines below 10.

For completeness, it is also worth considering how easy it is to successfully implement each model. FE, DL, SJ and Q are simple non-iterative methods, easily programmed and able to provide an overall effect estimate in all cases. PE is also non-iterative but a rather complex and computationally intensive method and cannot be applied to meta-analyses of fewer than six studies. ML and PL are iterative, not overly complex, but computationally intensive and not always successful in providing an overall effect estimate. In our simulations, convergence was unsuccessful for up to 3% of the cases for ML and 1% for PL but results may depend on the specified algorithm parameters. BT is the most complex method since it involves Gauss–Laguerre quadrature integration, and the method failed for up to 12% of the simulated meta-analyses (the effect is greater for high heterogeneity and low study numbers). This is due to the fact that the method depends on the estimate of the variance of Cochran's Q which, according to our simulations, is not always positive.

Most statistical software packages offer FE and DL as analysis options, but not PL or other methods. Since even specialist MA software packages like *RevMan*[35] lack other alternatives and the results of this study clearly demonstrate that DL is often not the most optimum RE method, we have created *MetaEasy*,[36] a Microsoft Excel add-in that implements all the MA methods described in this article.

Most methods have also been implemented in a STATA module, available for download from the SSC archive under the name *metaan*.[37]

Acknowledgements

We thank the two anonymous reviewers for their comments, Dr Eric A. Engels for sharing his data with us and our colleagues in the National Primary Care Research and Development Centre whose computers we repeatedly hijacked for extra computational power.

References

1. Egger M and Smith GD. Meta-analysis. Potentials and promise. *BMJ* 1997; **315**(7119): 1371–1374.
2. Glass GV. Primary, secondary, and meta-analysis of research. *Educ Res* 1976; **5**(10): 3–8.
3. Higgins JPT and Green S (eds) *Cochrane handbook for systematic reviews of interventions 4.2.5 [updated May 2005]*. Chichester, UK: John Wiley & Sons, Ltd, 2005.
4. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. *BMJ* 1994; **309**(6965): 1351–1355.
5. Hardy RJ and Thompson SG. Detecting and describing heterogeneity in meta-analysis. *Stat Med* 1998; **17**(8): 841–856.
6. Higgins JP and Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med* 2002; **21**(11): 1539–1558.
7. Takkouche B, Cadarso-Suarez C and Spiegelman D. Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis. *Am J Epidemiol* 1999; **150**(2): 206–215.
8. Glasziou PP and Sanders SL. Investigating causes of heterogeneity in systematic reviews. *Stat Med* 2002; **21**(11): 1503–1511.
9. Song F, Sheldon TA, Sutton AJ, Abrams KR and Jones DR. Methods for exploring heterogeneity in meta-analysis. *Eval Health Prof* 2001; **24**(2): 126–151.
10. Petitti DB. Approaches to heterogeneity in meta-analysis. *Stat Med* 2001; **20**(23): 3625–3633.
11. Brockwell SE and Gordon IR. A comparison of statistical methods for meta-analysis. *Stat Med* 2001; **20**(6): 825–840.
12. Thompson SG and Pocock SJ. Can meta-analyses be trusted? *Lancet* 1991; **338**(8775): 1127–1130.
13. DerSimonian R and Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986; **7**(3): 177–188.
14. Bohning D, Malzahn U, Dietz E, Schlattmann P, Viwatwongkasem C and Biggeri A. Some general points in estimating heterogeneity variance with the DerSimonian-Laird estimator. *Biostatistics* 2002; **3**(4): 445–457.
15. Thompson SG. Controversies in meta-analysis: the case of the trials of serum cholesterol reduction. *Stat Methods Med Res* 1993; **2**(2): 173–192.
16. Micceri T. The unicorn, the normal curve, and other improbable creatures. *Psychol Bull* 1989; **105**(1): 156–166.
17. Follmann DA and Proschan MA. Valid inference in random effects meta-analysis. *Biometrics* 1999; **55**(3): 732–737.
18. Laird N. Nonparametric maximum likelihood estimation of a mixing distribution. *J Am Stat Assoc* 1978; **73**: 805–811.
19. Aitkin M. Meta-analysis by random effect modelling in generalized linear models. *Stat Med* 1999; **18**(17–18): 2343–2351.
20. Sidik K and Jonkman JN. A simple confidence interval for meta-analysis. *Stat Med* 2002; **21**(21): 3153–3159.
21. Sidik K and Jonkman JN. A comparison of heterogeneity variance estimators in combining results of studies. *Stat Med* 2007; **26**(9): 1964–1981.
22. Biggerstaff BJ and Tweedie RL. Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Stat Med* 1997; **16**(7): 753–768.
23. Poole C and Greenland S. Random-effects meta-analyses are not always conservative. *Am J Epidemiol* 1999; **150**(5): 469–475.
24. Cochran WG. Problems arising in the analysis of a series of similar experiments. *J R Stat Soc* 1937; **Suppl 4**(1): 102–118.
25. Mittlbock M and Heinzl H. A simulation study comparing properties of heterogeneity measures in meta-analyses. *Stat Med* 2006; **25**(24): 4321–4333.
26. Hardy RJ and Thompson SG. A likelihood approach to meta-analysis with random effects. *Stat Med* 1996; **15**(6): 619–629.
27. StataCorp LP. Stata Statistical software. [9.2]. 2005. College Station, Texas, USA, StataCorp LP. Ref type: Computer Program.
28. The MathWorks Inc. MATLAB. [7.6.0.324]. 2008. Natick, Massachusetts, USA, The MathWorks Inc. Ref type: Computer Program.
29. Engels EA, Schmid CH, Terrin N, Olkin I and Lau J. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. *Stat Med* 2000; **19**(13): 1707–1728.
30. Ramberg JS, Dudewicz EJ, Tadikamalla PR and Mykytka EF. A probability distribution and its uses in fitting data. *Technometrics* 1979; **21**(2): 201–214.
31. Tadikamalla PR. On simulating non-normal distributions. *Psychometrika* 1980; **45**(2): 273–279.
32. Reinartz WJ, Echambadi R and Chin WW. Generating non-normal data for simulation of structural equation models using Mattson's method. *Multivariate Behav Res* 2002; **37**(2): 227–244.
33. Hedges LV. The power of statistical tests in meta-analysis. *Psychol Methods* 2001; **6**(3): 203–217.
34. Overton RC. A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. *Psychol Methods* 1998; **3**(3): 354–379.
35. The Cochrane Collaboration. Review Manager for Windows. [4.2.10]. 2007. Ref type: Computer Program.
36. Kontopantelis E and Reeves D. MetaEasy: a meta-analysis add-in for Microsoft Excel. *J Stat Softw* 2009; **30**(7): 1–25.
37. Kontopantelis E and Reeves D. metaan: random effects meta-analysis. *Stata J* 2010; **10**(3): 395–407.