Examining how meta-analytic methods perform in the presence of bias: A simulation study

Background

Meta-analysis is a widely used technique for quantitatively integrating the results from multiple studies. Conclusions drawn from meta-analysis results are increasingly considered by healthcare practitioners and policy makers. However meta-analysis techniques are vulnerable to several forms of bias. In this paper we address three distinct types of bias for which there are well documented problems in meta-analyses of randomised trials [1–6]: within-study biases arising from limitations in design and conduct, publication bias and selective outcome reporting bias.

Limitations in the design and conduct of trials can introduce within-study bias[1]. Limitations include poor randomisation methods and lack of blinding, with the latter being a problem particularly for studies using subjective outcomes. Publication bias occurs when the availability of information about studies relevant to a systematic review is correlated with the results obtained by those studies. This is typically conceptualised as being driven by the statistical significance of results, with publication being either suppressed or delayed when results are non-significant. The problem particularly affects small studies, as these are less likely to reach statistical significance unless the effect observed is of large magnitude. Outcome reporting bias can occur when multiple outcome measures are used in a study, but only a subset of these is reported. Bias arises if the decision of which outcome measures are reported is related to the magnitude (and often the statistical significance) of their results. Both publication and outcome reporting biases are forms of reporting bias and ignoring them in a meta-analysis may lead to an overestimation (or underestimation) of the true effect magnitude in the field of study.

Each of these three problems might cause the estimates from a meta-analysis to be biased, or the coverage from confidence intervals to be altered from the nominal level. Several methods for meta-analysis have been proposed as being less vulnerable to these biases. These include (i) methods that assume multiplicative rather than additive heterogeneity, leading to results that may be more robust to correlations between effect sizes and study sizes; (ii) methods that adjust confidence intervals around summary effects and (iii) methods that specifically model dependencies between effect sizes and study sizes. To our knowledge, these different approaches have not been compared with standard methods in an independent simulation study, or their performance determined in the presence of biases other than publication bias. We aimed to examine the statistical properties of the overall effect estimate and confidence interval for a selection of novel and existing meta-analysis methods across a wide range of conditions in the presence of several types of bias. To do this, we first outline the meta-analysis methods we examine. Then we describe the methods and parameters for a simulation study. We introduce models for introducing each type of bias, and how they were implemented in the simulations. Finally we present and explore the results and discuss their implications.

Methods

Models for meta-analysis

A popular model for meta-analysis is a fixed-effect (FE) model, sometimes referred to as a common-effect model, where study results ($\hat{\mu}_i$ for study *i*) are assumed to be estimating a common underlying effect parameter (θ). The differences between results are due only to within-study variability with variance $\hat{\sigma}_i^2$, which is conventionally assumed known:

$$\hat{\mu}_i \sim \mathrm{N}(\theta, \hat{\sigma}_i^2).$$

The inverse variance method (IV) estimates this common effect by constructing a weighted mean where study weights (w_i for study *i*) are equal to the inverse of the study variances. The variance of this estimate is a combination of the individual study variances. We provide statistical formulae used for estimates of effect and its variance, study weights, and critical values for confidence intervals for the FE method, along with all the other methods we address, in Table 1.

Random-effects models assume there is a distribution of true effect parameters rather than a single common effect. This means that rather than estimating a common effect, they estimate the mean of a distribution of true effects, the variance of which is the between-studies heterogeneity (τ^2):

$$\hat{\mu}_i \sim \mathrm{N}(\theta, \hat{\sigma}_i^2 + \tau^2).$$

This is often conceptualised as representing differences in study design or population which might affect the size of effect being measured. The variance of study results is therefore a combination of between and within-study variability. There are several methods of generating an estimate of the heterogeneity variance ($\hat{\tau}^2$), the most widely used of which is the DerSimonian and Laird (DL) estimate as it is computationally straightforward. A restricted maximum likelihood (REML) method is also generally recommended[12]. Critical values for confidence intervals for traditional fixed and random-effects models are conventionally taken from the standard normal distribution.

Many have highlighted a limitation of the standard random-effects method because a point estimate of τ^2 (for example using the DerSimonian and Laird or REML method) is assumed known so that uncertainty in the estimate is ignored. Knapp and Hartung (KH)[13] aim to account for this uncertainty. They adopt the random-effects model, keeping the same estimate of overall effect but proposing an alternative estimate for the variance of this overall effect estimate, and using a t-distribution for significance testing and calculating confidence intervals.

A well-known problem with a random-effects approach is that, since the weights are calculated based on the sum $\hat{\sigma}_i^2 + \hat{\tau}^2$, the weights given to each study will become relatively

more similar as the amount of between-studies variability increases[7,8]. This decreases the relative weight of studies with small variance (typically large studies) compared with those with large variance, increasing any risk of bias driven by small studies (e.g. publication bias). A multiplicative model (which we refer to as Mult) for heterogeneity has been suggested to avoid this, as an alternative to the standard additive model, such that

$$\hat{\mu}_i \sim N(\theta, \hat{\sigma}_i^2 \phi).$$

This gives the same overall effect estimate as the fixed-effect model, but with the standard errors from each study inflated by $\sqrt{\phi}$. The parameter ϕ can be estimated by running a weighted linear regression of the observed effect sizes against a constant and calculating the mean square error, fixed to be ≥1 [8] (the estimated value of ϕ is also equal to the H^2 statistic [14]).

Henmi and Copas (HC)[9] propose using a fixed-effect model (with the IV method specifically) to estimate the overall effect, noting that such a weighted average will be less influenced by small studies, therefore reducing the impact of publication bias. However, they then advocate using a random-effects model to construct a confidence interval around the fixed-effect estimate, using the DL estimator for τ^2. Rather than using the usual normal distribution, they derive alternative values (described in detail in the original paper). A similar method is the IVHet[15] model, which yields the same overall effect estimate and variance as HC but does not use the alternative values for the confidence interval. For this paper we examine both the HC and IVHet models, assuming a standard normal distribution to compute confidence intervals for the latter.

Finally, noting that studies with lower variance are likely to be less biased in the presence of publication bias, Moreno et al. [10] propose an extension to Egger's test, which they refer to as Egger-Var (abbreviated EV in this paper). The observed effect sizes are regressed against study variances, weighting by the reciprocal of the study variance (with a multiplicative dispersion parameter as in the traditional Egger test). The intercept in this model is then interpreted as the effect size in an infinitely large study.

In summary, the methods we compare in this paper are the standard fixed-effect and random-effects methods, the multiplicative model (Mult), the HC and IVHet approaches to applying a random-effects based confidence interval around a fixed-effect estimate and the EV regression-based method. Although there are many more methods we could have compared, we chose this selection to provide a range of different approaches that have been proposed in the medical meta-analysis literature, are reasonably well known and/or are frequently used.

Simulation

Our process of simulation for each condition was as follows: (a) set parameters, including effect measure and type of bias (if any); (b) draw a study level effect parameter and sample size; (c) simulate individual participant-level data within the study; (d) compute study level effect estimates and standard errors; (e) repeat (b) to (d) until the specified number of studies is obtained; (f) perform meta-analysis; and (g) collect data about meta-analysis results.

Mean difference simulation: base case

We simulated unstandardised mean differences (MDs) as a base case because these should perfectly follow a normal distribution with variances that are independent of sample size. This allows us to study the effect of each method under optimal conditions. The study-level true MD for study i (μ_i) was drawn from $\mu_i \sim \mathrm{N}(\theta, \tau^2)$.

Four scenarios were specified for within-study sample sizes (n_i), setting them (i) all to be equal for every study; (ii) variable across studies to reflect typical meta-analyses in practice; (iii) variable across studies but all small; and (iv) variable across studies but all large. Details are provided below under 'Choice of simulation parameters'. Next, values for each individual j were sampled for the treatment arm x_{ijT} and control arm x_{ijC} with equal numbers in each group:

$$x_{ijT} \sim \mathrm{N}\left(\frac{\mu_i}{2}, \sigma^2\right)$$

$$x_{ijC} \sim \mathrm{N}\left(-\frac{\mu_i}{2}, \sigma^2\right).$$

Here σ^2 is the within-study, between-individual variance which was held to be equal in the two arms. Without loss of generality, we fixed these at 1, which allows interpretation of the MDs as standardized mean differences (SMDs) to allow comparison with log odds ratio values. From these individual values, study estimates of effect and variance were calculated using standard methods for mean differences (see supporting information 1F). This process was repeated for each of the k studies in each meta-analysis and the whole process is represented as a diagram in supporting information 1I.

Log odds ratios simulation: base case

We also examined properties of the meta-analysis methods when applied to odds ratios, which are a common way of presenting the results of medical studies. Estimates of log odds ratios (LORs) are correlated with their estimated variances, which can make it difficult to interpret the results of meta-analyses on this scale. Since our simulated biases would themselves introduce correlations between LORs and variances, we wanted to explore the properties of the meta-analysis methods in this situation.

To simulate results on the LOR scale, true study level LORs (μ_i) were simulated in the same way as MD values. Event probabilities for treatment (π_{iT}) and control (π_{iC}) groups were then calculated using

$$\text{logit}(\pi_{iT}) = \gamma + \mu_i/2$$

$$\text{logit}(\pi_{iC}) = \gamma - \mu_i/2$$

where γ is the logit of the (prespecified) average event frequency across the two groups, assumed for the sake of simplicity to be identical in every study. In practice the event frequency could vary between studies. Cell counts for the treatment and control arm were then obtained using a binomial draw using the number in the arm and probability π_i.

Any studies that had zero event counts in both arms (or event counts equal to the sample sizes in both arms) were excluded from the analysis and were not re-sampled (to approximate a realistic scenario and prevent skewing of study sample sizes). In studies with a zero event count in one arm (or an event count equal to the sample size in one arm), we added 0.5 to all four cell counts. Finally log odds ratios and variances were computed from the cell count data using standard methods (see supporting information 1F).

Simulation of publication bias

To simulate publication bias, we implemented a selection rule after each study result was generated. If it 'passed' then the simulation would continue, if it did not pass then another study result was drawn using the same sample size, a process that repeated until a study result was selected. This method was chosen to maintain the distribution of study sizes. A one-tailed significance test was chosen to model the situation where one direction of effect is preferred (in this case negative values). We used Hedges and Vevea's model for 'light' publication bias which is a step function based on the study's p-value[16]:

$$\text{probability of publication} = \begin{cases} 1, & p < 0.05 \\ 0.75, & 0.05 < p < 0.2 \\ 0.25, & 0.2 < p \leq 1 \end{cases}.$$

This model for publication bias generates a familiar pattern in the funnel plot, with absent studies in one corner of the plot, but with most large studies retained and remaining unbiased. The model was chosen because it is conceptually simple, widely used, mechanistically clear, involves both effect size and study size (through their combined influence on the p-value), and there is good evidence for publication bias being dependent on p-value. However, the impact of this selection mechanism will depend on the value of the effect parameter, θ. When θ is large and in the preferred direction, almost all studies will be selected, whereas when θ is large and in the opposite direction, a small proportion of studies will be selected and the plot will largely be populated by unlikely results. We illustrate the impact of the bias for $\theta = 0$ using a contoured funnel plot in the supporting information 1E.

Simulation of methodological limitations

Bias could be introduced by many methodological limitations in trials such as inadequate randomisation, lack of allocation concealment, or lack of blinding in clinical trials. We sought to mimic the effect of such limitations based on real-life data. We re-analysed the BRANDO

meta-epidemiological study[1] and observed an association between study sample size and adequate randomisation and allocation concealment, although not with blinding (based on logistic regression of sample size on risk of each methodological flaw). Approximately 20% of studies with a sample size of less than 500 had evidence of adequate allocation concealment, as opposed to 36% with a sample size above 500. For randomisation, 24% of studies below 500 participants were adequate, compared with 33% above that threshold. Such an association would lead to an asymmetrical funnel plot if the flaws are also associated with differences in effect sizes, as was observed to be the case in the BRANDO study[1].

To simulate bias due to methodological limitations, we introduced an impact of two possible methodological flaws (representing randomisation and allocation concealment) each with a fixed magnitude of bias. Based on our findings from the BRANDO data, we assumed that both flaws had a decreasing likelihood with increasing sample size. We assumed that the flaws were otherwise independent, which is a substantial simplifying assumption since evidence from BRANDO suggests that biases do not operate independently[1]. The size of the bias in studies possessing the flaws was modelled as an additive bias of −0.090 per bias on the MD scale or as a multiplicative bias of 0.85 per bias on the odds ratio scale (corresponding to −0.16 on the LOR scale). This bias was applied to the underlying study-level effect parameter (μ_i) as it was drawn.

Assuming that both flaws introduce the same bias, and that they are equally common with a given sample size, we approximated the relationships observed in BRANDO by drawing the number of flaws (c_i) in the ith study from

$$c_i \sim \text{Binom}\left(2, \frac{1}{n_i^{0.06}}\right).$$

The effect of this bias on the funnel plot of results should differ from publication bias in several ways. First, the bias is unaffected by the value of θ, unlike in publication bias. Second, the extent of bias is limited and will not be as extreme, particularly with low sample sizes (see supporting information 1E).

Simulation of outcome reporting bias

To model outcome reporting bias we focussed on bias in selection of the reported result from among multiple outcome measures. Selective *non-reporting* of results is akin to publication bias so would lead to patterns similar to those addressed earlier. We imagined a situation where two correlated outcome measures were collected in every trial, but only the outcome measure with the lower p-value was reported. For MD simulations, we created two correlated variables at the level of individuals in each trial (separately for intervention and control arms) by creating correlated uniform distributions using Gaussian copulas then converting to a normal distribution. Both variables were analysed and the outcome with the lower p-value was used in the meta-analysis. Correlation introduced into individual values

was found to propagate through to correlation in MD with minimal attenuation, so a correlation of 0.6 was used at the individual level, giving an approximate correlation of 0.6 between MD values in the same study. For LOR simulations, correlated binomial outcomes were created. However individual level correlation attenuated more for LOR (this may have been partly due to the granularity of correlated binomial variables) so an individual level correlation of 0.8 was used to give a correlation in LOR of approximately 0.58.

Given that outcome reporting bias was present in every trial, this is likely to generate funnel plots where the entire plot was translated away from the true effect size, but the effect is more pronounced in smaller studies (see supporting information 1E). It would also be present across all values of θ.

Choice of simulation parameters

Study size

We simulated four different study sample size (n_i) conditions – fixed sample size, empirically-based sample sizes, small studies, and large studies. For the fixed condition, a single sample size was used for every study (60 for MD, 100 for LOR) approximating the empirical median value in a large sample of Cochrane Reviews[17]. In the condition with empirically-based sample sizes, the empirical distribution was approximated by drawing from a log normal distribution, $n_i \sim log\mathcal{N}(a, b)$. For MD we set a = 4.2, b = 1.1; and for LOR we set a = 4.7, b = 1.2, all chosen to approximate the empirical distribution. To ensure that standard deviations were calculable, we added 4 to simulated sample sizes for MDs or 2 for LORs. For the small and large studies conditions, we assumed a uniform distribution for square rooted samples sizes: $n\_i \sim \text{Unif}\left(\sqrt{a}, \sqrt{b}\right)^2$ .We set a = 20, b = 100 for small studies and a = 250, b = 1000 for large studies. All simulated sample sizes were rounded up to the next even integer to ensure integer values in each treatment group.

Number of studies

The number of studies per meta-analysis (k) varied across conditions with values of 3, 5, 10 and 30 selected to approximate 50th, 75th, 90th and 99th centiles of the empirical distribution found in Cochrane Reviews[17]. We added two larger numbers of studies, 50 and 100, to represent closer-to-ideal conditions.

Effect size

The values of θ for MD were -0.76, -0.12, 0, 0.12, 0.76. These were chosen to correspond to values of 0.25, 0.8, 1, 1.25 and 4 on the odds ratio scale, representing small and large effects symmetrically distributed around no effect (the logarithms of these values were used for the LOR simulation). The equivalence between MD and LOR scales arises from the conversion between odds ratios and standardised mean differences (SMD) described by Hasselblad and Hedges[18], coupled with our specification that all between-individual variances are equal to 1 (so that our MDs are interpretable as SMDs). This approximation is suitable when variances are equal in both arms and the data are not skewed[19], as is the case in our simulations.

Between study variance

To give τ^2 values for the simulation, we chose I^2 values of 0%, 5%, 20%, and 95%, which roughly approximate the 15th, 50th and >97.5th centile of the empirical distribution of I^2 for published surveys of meta-analyses of odds ratios[20]. We refer to these as no, low, moderate, and high heterogeneity. Average τ^2 values to give these I^2 values were calculated using the median study size which gave τ^2 = 0, 0.008, 0.04, 3.04 for LOR. The same Hasselblad and Hedges method was used to convert these to τ^2 = 0, 0.005, 0.022, 1.676 for the MD simulation (this gave simulations where the median I^2 values approximated 25th, 50th, and 95th centile of the empirical distribution for MD).

Event frequency

For LOR simulations, we varied the event probability ($logit^{-1}(\gamma)$) between 0.1, 0.3 and 0.5 to approximate typical frequencies of events in clinical trials whilst keeping failed results due to very rare events to a minimum.

Repetitions

Combinations of θ (5 conditions: no effect, small and large effect sizes in both direction), τ^2 (4 conditions: none, low, moderate, high), sample size (4 conditions: empirical distribution, small studies, study size fixed at median, large studies) and number of studies (6 conditions) resulted in 480 unique conditions to simulate for MD simulations. The extra three event frequencies resulted in 1,440 conditions for LOR. Each condition was repeated 10,000 times, a number chosen to produce on an estimated Monte Carlo Error of between 0.01 and 0.001 (based on a small pilot study with a fixed-effect estimate of effect for MD, three studies, θ = 0, τ^2 = 0, empirical distribution of study size, using the bootstrap grouping prediction method[21]).

Implementation of meta-analysis methods

Once simulation of study results was complete, meta-analysis was performed using all the methods we have described in every repetition for each condition. This gave 4 estimates of effect (FE using IV, RE using DL, RE using REML, and EV) and 7 confidence intervals (FE, RE using DL, RE using REML, KH, HC, IVHet, and Mult). The results of these were aggregated across the repetitions in a condition to give the mean bias and averaged mean squared error (MSE) for each estimate of effect, and the proportion of confidence intervals for which included the true value (coverage probability) for each method of calculating a confidence interval. Equations for these calculations can be found in supporting information 1F.

Software

All simulation and analysis was performed in R version 3.0.2 [22] using the packages *data.table*[23], *metafor*[24], *doParallel*[25], *foreach*[26], and *copula*[27]. Both simulation of data and meta-analysis were parallelised and performed using facilities at the Advanced Computing Research Centre at the University of Bristol. The package *doRNG*[28] was used to ensure independence and reproducibility of parallel loops. This uses the L'Ecuyer-CMRG random number generator. Seeds were set for each simulation so the R code required to generate

the full data as well as summary data for each condition are available online via supporting information 1G.

Results

The simulations generated a large volume of results. Full results can be found between the paper and the supporting information. In the body of the paper we present results from the MD simulation, and in the supporting information 1D we present corresponding results for the LOR simulation. The MD results are easier to interpret due to the lack of correlation between effect size and variance. However, the LOR simulation trends were broadly very similar. We present results for simulated outcome reporting bias also only in supporting information 1B. These are very similar in pattern to publication bias, though less extreme and with almost no variation with θ. We found no substantial difference with regard to bias, coverage, or MSE when comparing DL and REML heterogeneity estimators (see supporting information 1A). In all figures we display the effect of changing one variable with the others held in the reference condition of: five empirically sized studies; with no effect (θ = 0); and moderate magnitude of heterogeneity, τ^2. Heterogeneity and size of study are not connected as they are not displayed to scale.

No bias

Figure 1 shows results for all estimators in the base case in which none of the three types of bias were simulated. As might be expected, all estimators performed well when there is no underlying bias. They all quickly converged with increasing numbers of studies, were not affected by the value of θ, and improved marginally with increasing study size. The exception to this is the EV estimator, which appeared to improve in conditions in which studies had a greater range of sample sizes (i.e. it was substantially worse in the fixed sample size condition). This is expected since, as a regression-based method, it benefits from values over a greater range and closer points to the origin. This pattern for the EV estimator continues throughout all simulations. The IV method had the smallest bias, though the DL method was equivalent by either 10 studies or in moderate/high heterogeneity. The EV method achieved this low magnitude of bias by 30 studies, or 50 in high heterogeneity conditions.

Results for MSE can be found in the supporting information 1B. They mirror the trends found for bias. However they demonstrate that the EV is strictly inferior to other methods, and that the DL outperformed FE in high heterogeneity conditions, but also in low and moderate heterogeneity conditions with empirical sample sizes and 10 or more studies.

Coverage probability was also unaffected by the value of θ, and typically improved with increasing number of studies (though not for FE and multiplicative methods – except in no heterogeneity conditions). Larger sample size typically improved coverage, but figure 2 demonstrates an interaction where with fewer than 10 studies and low/moderate heterogeneity small and fixed sample sizes had better coverage than large and empirical sample sizes. However, the multiplicative method performed better with fixed sample sizes

than other conditions throughout. Increasing heterogeneity reduced coverage for all intervals with low numbers of studies except the KH interval, which performed near nominal level across almost all simulations. The KH interval coverage fell in conditions with the combination of moderate heterogeneity, small numbers of studies, and empirical distribution of sample sizes but in these scenarios it still outperformed the other intervals. The HC interval performed second best and approached the KH interval after 30 or more studies (and less with high heterogeneity).

Publication bias

Introducing publication bias makes the mean bias dependent on θ, where small absolute values were the most susceptible shown by figure 3. Increasing heterogeneity also worsens bias, and this effect is larger with small absolute θ values (which can be seen in the loop plots in supporting information 2). Increasing number of studies shows a small improvement in bias which is greatest with an empirical distribution of sample sizes. Large sample sizes give the lowest bias, followed by empirical sample sizes, whilst small and fixed sample sizes are broadly similar for IV and DL estimators. The EV method gives lower bias than both IV and DL except in fixed sample sizes. The IV estimator outperforms DL in most conditions, but by the largest margin in empirical study size and high heterogeneity conditions. However despite improvement in bias the EV estimator only started to outperform IV and DL in MSE with a combination of 10 or more studies, low heterogeneity, and θ values close to zero.

Coverage of each of the intervals was largely determined by the value of θ, with low absolute values giving poor coverage, but increasing numbers of studies exaggerated this effect (see figure 4). For example, using our reference condition the coverage of the HC interval (which performed best after 10 studies) was 82.8% with 3 studies, falling to 77.3% by 10, 60.3% by 30, 41.4% by 50 and only 12.3% by 100 studies. The IVHet method performed similarly at 50 studies with 37.3% but the FE was 13.1%, DL was 17.4%, KH was 20.9%, and Mult was 21.7%. When θ was 0 or -0.12, intervals had coverage less than 90% in almost all cases, falling to 60% or lower with 30 studies or more. Larger absolute values of θ were closer to nominal coverage, except with high heterogeneity. A θ of 0.12 fell somewhere between these, where there was reasonable coverage with small numbers of studies and low heterogeneity but it fell precipitously with higher numbers of studies (especially with small or fixed sample sizes) or rising heterogeneity. Large sample size improved coverage except where θ was zero but increasing heterogeneity typically worsened coverage. From these results there is minimal evidence to suggest IVHet or Mult outperform other methods in the presence of publication bias. As when there is no bias the KH interval performs the best with small numbers of studies, and the HC with larger numbers or high heterogeneity though coverage was globally poor in many conditions.

Methodological limitations

In the presence of methodological bias, the value of θ had little effect as shown in figure 5. Larger sample sizes improved performance, as did increasing numbers of studies. Bias was similar with all heterogeneity values except high. The EV estimate outperformed IV and DL in bias values, but also displayed its previously noted instability with low numbers of studies (especially in combination with high heterogeneity). Again the MSE showed that the EV did not outperform standard methods, except for conditions with a combination of more than 50 studies, empirical study sizes and less than maximal heterogeneity.

Coverage was largely unaffected by θ, but typically fell with increasing number of studies as confidence intervals tightened around a biased estimate. Increasing heterogeneity improved coverage with 10 studies or more (see figure 6) but with fewer the effect varied by method, with KH and HC typically improving in more heterogeneous scenarios, but FE and Mult deteriorating. Larger sample sizes improved coverage with 10 studies or more, with the ongoing exception of multiplicative performing best under fixed sample sizes and FE performing worse with heterogeneity and increasing sample size. With less than 10 studies empirical and large sample sizes only led to better performance than small and fixed for KH and HC intervals. Again the KH performed best with 3 or 5 studies, often greater than 90% coverage. With more studies or high heterogeneity the HC tended to outperform KH, but all intervals showed poor coverage (though better than under publication bias).

Discussion

Our extensive Monte Carlo study examined the properties of seven meta-analysis methods in simulated meta-analyses into which we introduced various types of bias. In **Error! Reference source not found.** we summarize the apparent strengths and limitations of each method on the basis of our findings. Examining bias results it is clear that the regression-based Egger-Var method confers an advantage over standard methods only in certain circumstances. Whilst it seems strictly inferior to standard methods in simulations with no bias, in simulations with publication bias or methodological limitations it has a lower bias than fixed or random-effects methods in most conditions. However this improved bias does not lead to an improved MSE except in optimal conditions for the method (large numbers of studies with a wide range of sample sizes and without heterogeneity). It is possible that the MSE for the Egger-Var with 3 studies (in particular) is being inflated by outlier values which might not be reported by a thoughtful meta-analyst. This limits its use as a primary estimate of effect but may marginally improve its value as part of a sensitivity analysis. Though this was not examined in the main paper, the Egger-Var method also appears to be more vulnerable to the correlation between effect size and variance than standard methods in log odds ratio simulations. In some situations this could be ameliorated by transforming the outcome variable (e.g. using an arcsine transformation). Additionally Egger-Var is very vulnerable to deteriorating when there are low event rates.

Regarding confidence intervals, it is notable that coverage probability when any bias mechanism was simulated was broadly poor. Whilst some methods performed better than others, in any conditions that were vulnerable to bias, coverage routinely fell below 90% and in the most severely biased conditions coverage fell below 50%.

The multiplicative method performed poorly relative to other proposed methods, and often became substantially worse when studies had empirically-based sample sizes and high heterogeneity. It only offered advantage over FE coverage in the presence of bias, and over DL coverage with the combination of low heterogeneity, 30 or more studies, and publication or outcome reporting bias.

As might be expected by their equivalent variance construction, the HC and IVHet intervals followed similar trends. However the HC interval outperformed the IVHet where there was high heterogeneity. It also tended to outperform IVHet in conditions with 30 or more studies, where the HC often performed the best of any interval. With 10 studies or fewer, the interval with closest to nominal coverage was typically the KH interval, especially with high heterogeneity. This was true in MD and LOR simulations for both methodological and publication biases.

There are several important limitations to our investigation. We examined only a small selection from a large array of potential methods for detecting or correcting for bias. We chose methods that covered a variety of different approaches and that we have observed informally to be considered attractive to medical meta-analysts. As a simulation study, the results may not be representative of `real world' data from any specific field, though we chose the conditions to match the known empirical distributions of biases in meta-analyses within the Cochrane Database of Systematic Reviews as closely as possible. In addition, the models we used to introduce biases are necessarily a simplistic representation of how these flaws work in reality and different models would likely affect the efficacy of each method. We also simulated only mean differences and log odds ratios, and patterns could be different for other types of data and other effect measures. Finally there are other potential causes of bias which, if they create different patterns of funnel plot asymmetry, could show different patterns of bias from these methods.

However, despite these limitations, we believe our paper adds value by comparing a number of statistical methods in a very broad range of empirically derived scenarios to inform recommendations about future use. We have also described new methods for modelling the effect of outcome reporting bias and bias resulting from methodological limitations in trials which represent, as far as we are aware, the first attempts to model these important sources of bias in meta-analysis simulations.

Conclusions

Within the limitations of simulation studies, our results confirm that meta-analysis methods remain susceptible to multiple types of bias, with no method giving satisfactory results. They

also highlight the risk that reporting biases and bias from methodological limitations in constituent studies pose to the results of meta-analysis and their use in decision making or forming guidance.

Acknowledging that no method corrects bias adequately, these results suggest that the use of the Knapp-Hartung adjustment for confidence intervals on estimate of effect give the closest to nominal coverage with approximately 10 studies or fewer, whilst with more than that the Henmi-Copas interval is preferable. With regards to point estimation of effect, the IV fixed-effect method appeared to outperform random-effects methods marginally in both bias and MSE, except when heterogeneity was high, though differences were often very small. The Egger-Var method did successfully correct for bias but performed poorly in conditions with a small range in sample sizes, small numbers of studies, high heterogeneity, or low event rates. Given this it seems to have a use in conditions that are optimal for the method, although the range of such conditions is limited.

As new methods are proposed in an effort combat bias we would also suggest that authors consider how they would function in the presence of multiple types of bias, rather than just publication bias.

Highlights

It is well established that several forms of bias can affect meta-analysis results. The best known is publication bias, but selective reporting of outcomes (outcome reporting bias), and flaws in the conduct of constituent studies can also introduce bias. Several statistical methods have been proposed to reduce the effect of bias, publication bias in particular.

Our results show that publication bias, outcome reporting bias, and flaws from studies can introduce slightly different patterns of bias and that no existing method is robust to all these problems, though some perform slightly better. Additionally we found that statistical performance under bias could conceivably be very poor.

Though the simulation was based on characteristics of medical studies, the results should be at least partially transferrable to all fields where meta-analysis is performed. Particularly the knowledge both that bias has the potential to be very problematic and that existing methods cannot correct for this.

Bibliography

1. Savović J, Jones HE, Altman DG, et al. Influence of Reported Study Design Characteristics on Intervention Effect Estimates From Randomized, Controlled Trials. *Ann Intern Med*. 2012;157(6):429. doi:10.7326/0003-4819-157-6-201209180-00537

2. Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *BMJ*. 1997;315(7109):640-645. doi:10.1136/bmj.315.7109.640

3. Williamson PR, Gamble C, Altman DG, Hutton JL. Outcome selection bias in meta-analysis. *Stat Methods Med Res*. 2005;14(5):515-524. doi:10.1191/0962280205sm415oa

4. Dwan K, Altman DG, Arnaiz JA, et al. Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias. *PLOS ONE*. 2008;3(8):e3081. doi:10.1371/journal.pone.0003081

5. Mathieu S, Boutron I, Moher D, Altman DG, Ravaud P. Comparison of Registered and Published Primary Outcomes in Randomized Controlled Trials. *JAMA*. 2009;302(9):977-984. doi:10.1001/jama.2009.1242

6. Kicinski M, Springate DA, Kontopantelis E. Publication bias in meta-analyses from the Cochrane Database of Systematic Reviews. *Stat Med*. 2015;34(20):2781-2793. doi:10.1002/sim.6525

7. Stanley TD, Doucouliagos H. Neither fixed nor random: weighted least squares meta-analysis. *Stat Med*. 2015;34(13):2116-2127. doi:10.1002/sim.6481

8. Mawdsley D, Higgins JPT, Sutton AJ, Abrams KR. Accounting for heterogeneity in meta-analysis using a multiplicative model—an empirical study. *Res Synth Methods*. 2017;8(1):43-52. doi:10.1002/jrsm.1216

9. Henmi M, Copas JB. Confidence intervals for random effects meta-analysis and robustness to publication bias. *Stat Med*. 2010;29(29):2969-2983. doi:10.1002/sim.4029

10. Moreno SG, Sutton AJ, Ades A, et al. Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Med Res Methodol*. 2009;9:2. doi:10.1186/1471-2288-9-2

11. Moreno SG, Sutton AJ, Turner EH, et al. Novel methods to deal with publication biases: secondary analysis of antidepressant trials in the FDA trial registry database and related journal publications. *BMJ*. 2009;339:b2981. doi:10.1136/bmj.b2981

12. Langan D, Higgins JPT, Jackson D, et al. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Res Synth Methods*. 2019;10(1):83-98. doi:https://doi.org/10.1002/jrsm.1316

13. Hartung J, Knapp G. On tests of the overall treatment effect in meta-analysis with normally distributed responses. *Stat Med*. 2001;20(12):1771-1782. doi:10.1002/sim.791

14. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med*. 2002;21(11):1539-1558. doi:10.1002/sim.1186

15. Doi SAR, Barendregt JJ, Khan S, Thalib L, Williams GM. Advances in the meta-analysis of heterogeneous clinical trials I: The inverse variance heterogeneity model. *Contemp Clin Trials*. 2015;45, Part A:130-138. doi:10.1016/j.cct.2015.05.009

16. Hedges LV, Vevea JL. Estimating Effect Size under Publication Bias: Small Sample Properties and Robustness of a Random Effects Selection Model. *J Educ Behav Stat*. 1996;21(4):299-332. doi:10.2307/1165338

17. Davey J, Turner RM, Clarke MJ, Higgins JP. Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: a cross-sectional, descriptive analysis. *BMC Med Res Methodol*. 2011;11:160. doi:10.1186/1471-2288-11-160

18. Hasselblad V, Hedges LV. Meta-analysis of screening and diagnostic tests. *Psychol Bull*. 1995;117(1):167-178.

19. Anzures-Cabrera J, Sarpatwari A, Higgins JP. Expressing findings from meta-analyses of continuous outcomes in terms of risks. *Stat Med*. 2011;30(25):2967-2985. doi:10.1002/sim.4298

20. Rhodes KM, Turner RM, Higgins JPT. Empirical evidence about inconsistency among studies in a pair-wise meta-analysis. *Res Synth Methods*. 2016;7(4):346-370. doi:10.1002/jrsm.1193

21. Koehler E, Brown E, Haneuse SJ-PA. On the Assessment of Monte Carlo Error in Simulation-Based Statistical Analyses. *Am Stat*. 2009;63(2):155-162. doi:10.1198/tast.2009.0030

22. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2017. https://www.R-project.org/

23. Dowle M, Srinivasan A, Gorecki J, Short T, Lianoglou S, Antonyan E. *Data.Table: Extension of "Data.Frame."*; 2017. https://cran.r-project.org/web/packages/data.table/index.html

24. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw*. 2010;36(3):1-48.

25. Calaway R, Analytics R, Weston S, Tenenbaum D. *DoParallel: Foreach Parallel Adaptor for the "parallel" Package*.; 2015. https://cran.r-project.org/web/packages/doParallel/index.html

26. Calaway R, Analytics R, Weston S. *Foreach: Provides Foreach Looping Construct for R*.; 2015. https://cran.r-project.org/web/packages/foreach/index.html

27. Hofert M, Kojadinovic I, Maechler M, Yan and J. *Copula: Multivariate Dependence with Copulas*.; 2017. https://cran.r-project.org/web/packages/copula/index.html

28. Gaujoux R. *DoRNG: Generic Reproducible Parallel Backend for "foreach" Loops*.; 2017. https://cran.r-project.org/web/packages/doRNG/index.html

Table 1: Summary of methods

Method	Estimate ($\widehat{\theta}$) Weight	Variance of $\widehat{\theta}$	Quantile for C.I.
FE (IV)	$\theta_{IV} = \frac{\sum w_i Y_i}{\sum w_i}$ $w_i = \frac{1}{\widehat{\sigma}_i^2}$	$Var(\theta_{IV}) = \frac{1}{\sum w_i}$	$z_{1-\alpha/2}$
DL	$\theta_{DL} = \frac{\sum w_i' Y_i}{\sum w_i'}$ $w_i' = \frac{1}{\widehat{\sigma}_i^2 + \widehat{\tau}^2}$	$Var(\theta_{DL}) = \frac{1}{\sum w_i'}$	$z_{1-\alpha/2}$
KH	θ_{DL} w_i'	$Var(\theta_{KH}) =$ $\frac{1}{(k-1)\sum w_i'} \max\left\{1, \sum w_i'(\theta_i - \widehat{\theta})^2\right\}$	$t_{k-1,1-\alpha/2}$
HC	θ_{IV} w_i	$Var(\theta_{HC}) = \frac{\tau^2 \sum w_i^2 + \sum w_i}{(\sum w_i)^2}$	$u_{\alpha/2}$, $u = \frac{\widehat{\theta}_F - \theta}{\sqrt{\widehat{V}}}$ solved numerically
IVHet	θ_{IV} w_i	$Var(\theta_{HC})$	$z_{1-\alpha/2}$
Mult	θ_{IV} $w_i'' = \frac{1}{\widehat{\sigma}_i^2 \phi}$ $\phi = \frac{\sum w_i (\hat{\mu}_i - \theta_{IV})^2}{k-1}$	$Var(\theta_{Mult}) \frac{1}{\sum w_i''}$	$z_{1-\alpha/2}$
Egger-Var	$\widehat{\theta} = \alpha$ from WLS $y_i = \alpha + \beta \sigma_i^2 + \epsilon_i$ Weights $\frac{1}{\sigma_i^2}$ $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2 \phi)$ ϕ estimated		

Table 1 Statistical methods for meta-analysis. WLS is weighted least squares regression, z refers to the normal distribution and t to the t distribution. Processes for estimating the HC confidence intervals, KH variance, and Egger-Var estimate of effect are summarised here, but presented in greater detail in the original papers.

Method	Strengths	Limitations
FE (IV)	Computationally simple, widely implemented and used. The estimate of effect is less vulnerable to small-study bias than DL.	Poor estimate of effect under high heterogeneity or any bias. Inaccurate (and often too narrow) confidence interval with any heterogeneity or with any bias.
DL	Widely implemented and used. Confidence interval outperforms FE interval in the presence of any heterogeneity.	Estimate of effect and confidence interval vulnerable to any bias.
KH	Best coverage probability of examined methods with small numbers of studies under most conditions with any bias.	Coverage falls quickly in bias conditions with increasing numbers of studies.
HC	Best coverage probability with large numbers of studies (e.g. 25 or more) under most conditions with any bias, and some with moderate number of studies (e.g. 10). Good performance with high heterogeneity.	Poor coverage with fewer than 10 studies. Coverage probability suffers under bias conditions (though outperforms other intervals).
IVHet	Better coverage than standard intervals in bias conditions with large number of studies (e.g. 25 or more), often outperforms HC with fewer than 10 studies. Simpler to compute than HC.	Never the best choice of interval– for small numbers of studies KH is better, for larger numbers HC is better. Deteriorates with increasing heterogeneity.
Mult	Outperforms FE coverage in presence of heterogeneity or bias, and DL coverage with publication or outcome reporting bias and 30 or more studies.	Performs worse with increasing heterogeneity or an empirical sample size. Despite situational improvement over standard methods it performs poorly overall.
Egger-Var	Lower mean bias than other estimates in almost all bias simulations for mean differences and most bias conditions for odds ratios with a small true effect size. Improved mean squared error when there is also the combination of no heterogeneity, large number of studies, and a wide range of study sizes.	Worse than FE or DL with no bias. Worse mean squared error than FE or DL in most bias conditions despite improved mean bias. Unstable with few studies (e.g. 5 or fewer), particularly with a small range of sample sizes or high heterogeneity. For odds ratios, had a substantial bias towards zero at large effect sizes, exaggerated by low event rates and high heterogeneity.

Table 2: Summary of the strengths and limitations of the seven compared methods (FE = fixed effect; IV = inverse variance; DL = DerSimonian-Laird; KH = Knapp-Hartung; HC = Henmi-Copas; IVHet = Doi et al's inverse variance heterogeneity method; Mult = multiplicative heterogeneity model; Egger-Var = Moreno et al's variant of the Egger regression approach)