The median and the mode as robust meta-analysis methods in the presence of small study effects

Fernando Pires Hartwig[1,2]*, George Davey Smith[2,3], Amand Floriaan Schmidt[4,5], Jack Bowden[2,3]

[1]Postgraduate Program in Epidemiology, Federal University of Pelotas, Pelotas, Brazil.

[2]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom.

[3]Population Health Sciences, University of Bristol, Bristol, United Kingdom.

[4]Institute of Cardiovascular Science, Faculty of Population Health, University College London, London, United Kingdom.

[5]Faculty of Science and Engineering, Groningen Research Institute of Pharmacy, University of Groningen, Groningen, The Netherlands.

*Corresponding author. Postgraduate Program in Epidemiology, Federal University of Pelotas, Pelotas (Brazil) 96020-220. Phone: 55 53 981068670. E-mail: fernandophartwig@gmail.com; fh15144@bristol.ac.uk.

Abstract

Meta-analyses based on systematic literature reviews are commonly used to obtain a quantitative summary of the available evidence on a given topic. Despite its attractive simplicity, and its established position at the summit of the evidence-based medicine hierarchy, the reliability of any meta-analysis is largely constrained by the quality of its constituent studies. One major limitation is small study effects, whose presence can often easily be detected, but not so easily adjusted for. Here, robust methods of estimation based on the median and mode are proposed as tools to increase the reliability of findings in a meta-analysis. By re-examining data from published meta-analyses, and by conducting a detailed simulation study, we show that these two simple methods offer notable robustness to a range of plausible bias mechanisms, without making any explicit modelling assumptions. In conclusion, when performing a meta-analysis with suspected small study effects, we recommend reporting the mean, median and modal pooled estimates as a simple but informative sensitivity analyses.

Keywords: Meta-analysis; Small study effects; Robust estimation.

1. Introduction

Meta-analysis is a statistical technique for obtaining a quantitative summary of the totality of evidence on a given topic, as part of a broader systematic review.[1,2] In its archetypal form, it provides an overall effect estimate for a well-defined intervention that has been assessed across several independent studies. In addition, meta-analyses provide a further opportunity to explore between study heterogeneity, which might highlight novel patient subgroups with contrasting treatment responses.[1,2]

Unfortunately, between-study heterogeneity may also indicate the presence of bias, which wish to understand, but ultimately remove, from the final analysis. For example, it is generally accepted that results in agreement with the prevailing wisdom and achieving conventional levels of significance are more likely to be reported and published than negative or inconclusive findings. Therefore, published studies are more likely to present large effect estimates, corresponding to a biased sample of all studies.[2] Larger studies are affected to a lesser extent than smaller studies, most obviously because an increased sample size raises the likelihood of achieving conventional statistical significance when the true effect to be measured is non-zero. Large studies also require a sizeable financial outlay from funding agencies and often represent the collective effort of many cross-institutional researchers. These factors provide added impetus to place the results in the public domain, regardless of their conclusions. Moreover, small studies are more often early-phase trials (which may use less stringent designs) than larger studies, or may preferentially include high risk patients in order to improve power. This phenomenon, which encompasses many different mechanisms, is referred to generically as “small study effects”.[2]

It is often difficult to identify whether any observed correlation between study size and reported treatment effect is due to “true” between-study differences, selective reporting and publication, or a combination of both.[2] Many different methods to detect and/or correct for small study effects have been proposed. One of the earliest of such methods is the funnel plot (where study-specific point estimates are plotted against their precision), which has been proposed more than 30 years ago. Asymmetry in the funnel plot may be indicative of selective reporting and publication, although “true” between-study differences correlated with precision can exist.[3] This motivated the development of methods that “correct” for asymmetry, such as Egger regression and trim-and-fill.[3-5] However, because these methods make either implicit or explicit assumptions about the selection process, their performance suffers acutely when the true bias mechanism is different. This can easily result in a bias-adjusted estimate that is further from the truth than the original result. This motivates our proposal of two simple methods which are naturally robust to small study effects, whilst making no assumptions about its precise nature.

2. Methods

Before discussing our proposed estimation procedures in detail, we first describe the underlying data generating model they will be evaluated against. This slightly unusual step is necessary for the reader to understand when each method can, in theory, identify the true treatment effect.

2.1. Data generating model

We start by defining a summary data generating model with K studies indexed by j $(j = 1,2, \dots, K)$. Each study reports an estimated mean difference between randomised groups (e.g., one to an experimental intervention and one to standard intervention) denoted by $\hat{\beta}_j$, where:

$$\hat{\beta}_j = \beta + b_j + \sigma_j \varepsilon_j. \qquad (1)$$

Here:

- β is the true effect of the exposure on the outcome;
- b_j denotes the bias/heterogeneity parameter for study j;
- σ_j is the standard error of the mean difference;
- $\varepsilon_j \sim N(0,1,l_j,u_j)$ is a draw from a standard truncated Normal distribution with lower limit l and upper limit u. When $l = -\infty$ and $u = \infty$, then ε_j denotes pure random error due to sampling variation.
- The parameters b_j, σ_j, l_j, and u_j are all allowed to depend on the study size, n_j.

We will use b_j and ε_j to variously induce heterogeneity, effect modification, and small study bias in the data, as described below.

2.2. General principles of our bias model

We will explore two types of small study bias: bias due to systematic differences between small and large studies due to study quality (type (a)), and bias due to the specific environment of selective reporting and publication in operation at the time when study j was conducted (type (b)).

For type (a), we imagine that bias is a fundamental property of each study, in that the true treatment effect for study j is $\beta + b_j$, where b_j is a simple, non-increasing function of study size (n_j). That is:

$b_j \leq b_k$, whenever $n_k \leq n_j$.

Without loss of generality we assume that the bias is always positive, so that $b_j \geq 0$. We will investigate cases where the bias disappears only asymptotically as a study size grows infinitely large, and cases where the bias disappears beyond a threshold study size, N. That is:

$b_j \rightarrow 0$ as $n_j \rightarrow \infty$, or $b_j = 0$ if $n_j \geq N$ for some (large) N.

Type (b) bias is not a fundamental component of the study itself, but instead the result of selective reporting and publication (i.e., dissemination bias). We induce this through the random error component of model (1), ε_j, in the following manner.

Again, we assume that bias is always positive, so that $E\left[\varepsilon_j | n_j\right] \geq 0$. This corresponds to a situation where the selection process favours the publication of studies that reported positive effect estimates. We achieve this by defining the lower limit of ε_j's truncated normal distribution, l_j, as a non-increasing function of n_j. That is:

$l_j \leq l_k$, and therefore $E\left[\varepsilon_j | n_j\right] \leq E[\varepsilon_k | n_k]$, whenever $n_k \leq n_j$.

Similarly to the type (a) bias model, we will explore cases where:

$l_j \rightarrow -\infty$ and $E\left[\varepsilon_j | n_j\right] \rightarrow 0$ as $n_j \rightarrow \infty$, or

$l_j = -\infty$ and $E\left[\varepsilon_j | n_j\right] = 0$ if $n_j \geq N$ for some large N.

A general expression for the expected value of study j's effect estimate $\hat{\beta}_j$, based on n_j participants and in the presence of type (a) and type (b) bias, is therefore:

$$E\left[\hat{\beta}_j | n_j\right] = \beta + b_j + \sigma_j E\left[\varepsilon_j | n_j\right] \qquad (2)$$

An important distinction between type (a) and type (b) bias is their respective effect on the variance of the study-specific estimates. Type (a) bias will generally increase their variability, leading to over-dispersion, or heterogeneity. Type (b) bias, by contrast, can have the opposite effect of reducing their variability, because of the truncation in the distribution of ε_j. That is, in the presence of this bias, $\text{Var}\left[\varepsilon_j | n_j\right]$ will generally be less than 1, and $\text{Var}\left[\varepsilon_j | n_j\right] \geq \text{Var}[\varepsilon_k | n_k]$ whenever $n_k \leq n_j$.This phenomenon leads to under-dispersion across the set of study-specific estimates constituting the meta-analysis.

2.3. Robust central tendency statistics in meta-analysis

We now introduce three estimators for the overall effect β: the standard approach plus two novel approaches, and discuss their ability to return consistent estimates for data generated under model (1). For the purposes of clarity only, we will momentarily assume that b_j is the sole source of bias in equation (1) – i.e., that $E[\varepsilon_j | n_j] = 0$.

2.3.1. The weighted mean

A standard fixed effect meta-analysis would estimate the effect size parameter β as an inverse-variance weighted average (or pooled mean) of the individual study estimates. That is:

$$\hat{\beta}_{FE} = \frac{\sum_{j=1}^{K} \hat{\beta}_j \sigma_j^{-2}}{\sum_{j=1}^{K} \sigma_j^{-2}} \qquad (2).$$

However, if even a single study contributes a biased estimate to the meta-analysis (e.g., via a non-zero b_j), then the pooled mean will also generally be biased. That is, using the notation of formula (1):

$E[\hat{\beta}_{FE}] \neq \beta$ in general, whenever $b_j > 0$ for some study j in $1, \ldots, K$.

For this reason, in the language of robust statistics, the mean is said to have a 0% “breakdown” level.

2.3.2. The weighted median

The weighted median[6] estimate is defined as the 50th percentile of the inverse-variance weighted empirical distribution of the study specific estimates, which can be calculated as follows. Assume that the $\hat{\beta}_j$’s are sorted in ascending order, so that $\hat{\beta}_K \geq \hat{\beta}_{K-1} \ldots \geq \hat{\beta}_1$. Let the standardised inverse-variance weights for study j be defined as $w_j = \frac{\sigma_j^{-2}}{\sum_{j=1}^{K} \sigma_j^{-2}}$ and sort them in the same order as the $\hat{\beta}_j$’s. Finally, let $s_j = \sum_{g=1}^{j} w_g$ denote the sum of standardised weights up to and including the jth study. This means that $\hat{\beta}_j$ is the $q_j = 100\left(s_j - \frac{w_j}{2}\right)$th percentile of the weighted empirical distribution of $\hat{\beta}_j$’s.

The weighted median estimate is the 50% percentile of this distribution, so it will be equal to $\hat{\beta}_j$ if $s_j = 0.5$. In general, no study lies exactly at the 50th percentile, so this quantity is estimated in practice by linear interpolation between its neighbouring estimates $\hat{\beta}_{j^*}$ and $\hat{\beta}_{j^\dagger}$, which correspond to the effect estimates reported by the studies located immediately before and after the 50% percentile, respectively (i.e., $q_{j^*} = \max(q_1, q_2, \ldots, q_{j^*})$, $q_{j^\dagger} = \min(q_{j^\dagger}, q_{j^\dagger+1}, \ldots, q_K)$, and $q_{j^*} < 0.5 < q_{j^\dagger}$). In this case, the weighted median estimate $\hat{\beta}_{WM}$ is:

$$\hat{\beta}_{WM} = \hat{\beta}_{j^*} + \left(\hat{\beta}_{j^\dagger} - \hat{\beta}_{j^*}\right)\frac{0.5 - q_{j^*}}{q_{j^\dagger} - q_{j^*}} \qquad (4)$$

The weighted median does not require that all $\hat{\beta}_j$'s are consistent estimates for the true effect β. More specifically, $\hat{\beta}_{WM}$ is consistent if up to (but not including) 50% of the total weight in the analysis comes from biased studies – i.e., $\sum_{j=1}^{K} I(b_j > 0)\, w_j < 50\%$. This means that the weighted median has a breakdown level of 50%.

2.3.3. The mode-based estimate

The mode-based estimate[7] (MBE) works by exploiting the Zero Modal Bias Assumption (ZEMBA), which requires that the most common value of the bias parameter b_j is zero. If ZEMBA holds, the mode of all $\hat{\beta}_j$'s (hereafter referred to as $\hat{\beta}_{MBE}$) is a consistent estimate of the true effect β, even if the majority of $\hat{\beta}_j$'s are biased.

More formally, $\hat{\beta}_{MBE}$ is consistent if $w_0 > \max(w_1, w_2, \dots, w_v)$, where (in a departure from the convention introduced for the weighted median) w_0 now denotes the sum of weights provided by studies with zero bias , and w_1, w_2 and w_v are the sum of weights provided by studies that have the smallest, the second smallest and the largest identical bias terms, respectively.

It is possible to define many pooled effect estimators that exploit ZEMBA in different ways. Here, as in Hartwig et. al,[7] we used the mode of the smoothed, inverse-variance weighted empirical density function of all $\hat{\beta}_j$'s as the MBE. More specifically, $\hat{\beta}_{MBE}$ is the value of x that maximizes $f(x)$ (i.e., $f(\hat{\beta}_M) = \max\,[f(x)]$). $f(x)$ is the normal kernel density function:

$$f(x) = \frac{1}{h\sqrt{2\pi}} \sum_{j=1}^{K} w_j \exp\left[-\frac{1}{2}\left(\frac{x - \hat{\beta}_j}{h}\right)^2\right] \qquad (5),$$

where h is the smoothing bandwidth parameter.[8] This parameter regulates a bias-variance trade-off, with smaller values of h reducing both bias and precision. We used the modified Silverman's bandwidth selection rule proposed by Bickel et al.[9]

The exact breakdown level of the MBE depends on $\max(w_1, w_2, \dots, w_v)$, which is unknown. If all biased studies estimate the exact same effect parameter, then ZEMBA will only be satisfied if up to (but not including) 50% of the weights comes from biased studies. The upper limit of the breakdown level is up to (but not including) 100%, and corresponds to the situation where all invalid studies estimate different effect parameters. Therefore, the breakdown level of the MBE ranges from 50% to 100%.

The weighted median and MBE were originally proposed as robust tools for summary data Mendelian randomization,[6,7] which is analogous to a meta-analysis.

2.4. Illustrating the identifying assumptions of the mean, median and mode

Figure 1 illustrates the assumptions underlying the pooled mean, median and mode in a hypothetical meta-analysis of 10 studies, sorted in ascending order of their $\hat{\beta}_j$'s. The true effect β is zero. For simplicity, all studies have the same weight and no sources of heterogeneity other than bias are present. Chiefly:

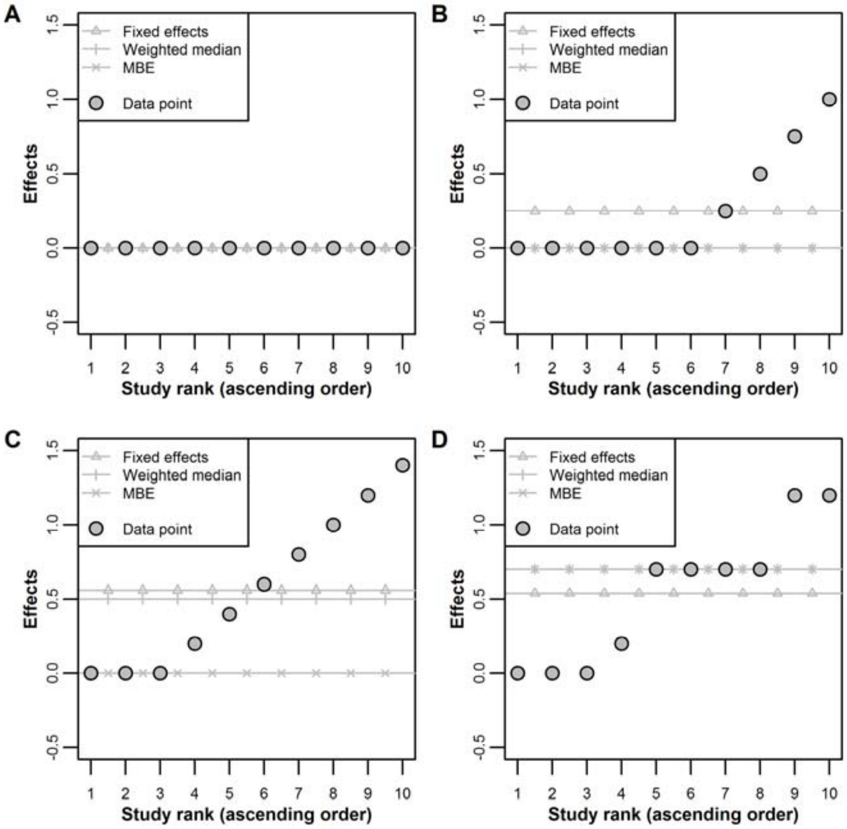

Figure 1. Illustration of the assumptions underlying the weighted median and the mode-based estimate (MBE) methods. Studies are assumed to have the same weights in the meta-analysis, and are sorted in ascending order of point estimate. The true effect is zero.
A: no heterogeneity between studies. B: 4 out of 10 studies are biased. C: 7 out of 10 studies are biased, but unbiased studies comprise the largest subgroup of studies that reported the same result. all biased studies reported different effects. D: 7 out of 10 studies are biased, and biased studies comprise the largest subgroup of studies that reported the same result.

- When all 10 studies (i.e., 100%) are unbiased (Panel A), all three methods identify the true effect (zero);

- When 4 out of 10 studies are biased (Panel B), or whenever less than 50% of studies are biased in general, the mean is biased, but the median and the mode are unbiased;
- When 7 out of 10 studies are biased (Panel C), or whenever more than 50% of studies are biased in general, and ZEMBA is satisfied, both the mean and the median are biased, but not the mode;
- When more than 50% of the studies are biased (Panel D) and ZEMBA is violated, all methods are biased.

One attractive property of the weighted median and MBE is that they are naturally robust central tendency statistics, and do not make any specific assumptions about the selection mechanism at play. Therefore, they might be robust to a range of reasonable small study effects models. However, as Figure 1 illustrates, these methods are not guaranteed to provide consistent estimates of β, failing to do so when their identifying assumptions are violated. Nevertheless, these assumptions are weaker than the assumptions required by the standard pooled mean.

2.5. Simulation study

We performed a simulation study to evaluate the performance of different meta-analysis methods in a range of small study effects models. We evaluated the following methods: fixed effects model, Egger regression, trim-and-fill, weighted median and MBE.

Summary data were generated using equation (1). We assume that each study measured a binary exposure variable $X \sim \mathrm{Bernoulli}(0.5)$ (e.g., an intervention: yes=1, no=0) and a continuous outcome variable Y with variance equal to one. Therefore, the pooled standard error of the mean difference is one for all values of j, and $\sigma_j = \sqrt{4/n_j}$, where n_j is the sample size of the jth study. We will assume studies range in size from n_1 to n_2 uniformly, so that $n_j \sim \mathrm{Uniform}(n_1, n_2)$.

2.5.1. Type (a) bias

The value of the bias term b_j was defined as the following linear function of study size: $b_j = I_{b_j}\left(0.5 - 1 \times 10^{-4}(n_j - 100)\right)$. From this model, if $n_j = 100$ (the smallest study size in our simulations), then $b_j = 0.5 I_{b_j}$. If $n_j = 5000$ (the largest study size in our simulations), then $b_j = 0.01 I_{b_j}$.

The indicator function $I_{b_j} \sim \mathrm{Bernoulli}(\delta)$, with $\delta \in [0,1]$, dictates the presence ($I_{b_j} = 1$) or absence ($I_{b_j} = 0$) of bias. Therefore, the expected number of studies suffering from bias equals δK.

2.5.2. Type (b) bias

As described above, this bias was generated through ε_j, by varying l_j according to study size. Typically, dissemination bias models assume that results that achieve conventional levels of statistical significance are more likely to be published. Therefore, in our simulations, l_j was defined to correspond the maximum one-sided P-value (null hypothesis: true mean difference ≤0) allowed for publication for a given study size (p_j). That is:

$$l_j = Q(1 - p_j) \qquad (6),$$

where $Q(p)$ is the quantile function for the Student's t distribution with $n_j - 1$ degrees of freedom. For example, if $p_j = 0.025$ and $n_j = 1000$, then $l_j = Q(1 - p_j) = Q(0.975) \approx 1.96$. This situation can be interpreted as studies with 1000 participants only being publishable if the reported one-sided P-value is ≤ 0.025.

We generated dissemination bias by defining p_j as various functions of n_j, as described below.

i) p_j as a continuous function of n_j (up to N).

In this model, $p_j = \min\left(f\left(\frac{n_j}{N}\right), 1\right)$, where $f\left(\frac{n_j}{N}\right)$ is some non-decreasing function of $\frac{n_j}{N}$ and N is some upper threshold study size threshold at which $p_j = 1$ (and therefore $l_j = -\infty$) for all $n_j \geq N$. We compared three distinct functions:

- Identity: $f\left(\frac{n_j}{N}\right) = \frac{n_j}{N}$;
- Square root: $f\left(\frac{n_j}{N}\right) = \sqrt{\frac{n_j}{N}}$; and
- Quadratic: $f\left(\frac{n_j}{N}\right) = \left(\frac{n_j}{N}\right)^2$.

ii) p_j as a step function of n_j.

In this model, p_j is defined as a piecewise function of n_j classifying studies into small, medium or large. That is:

$$p_j = \begin{cases} p_{small} & \text{if } n_j \leq N_{small} \\ p_{medium} & \text{if } N_{small} < n_j < N_{large} \\ p_{large} & \text{if } N_{large} \leq n_j \end{cases} \qquad (7),$$

where $1 \leq N_{small} \leq N_{large}$ and $0 \leq p_{small} \leq p_{medium} \leq p_{large} \leq 1$. Therefore, studies classified in the same group have the same P-value requirements for publication, and the relationship between p_j and n_j follows a step function.

The relationship between p_j and n_j in each one of these four models is illustrated in Supplementary Figure 1.

2.5.3. Simulation scenarios

We evaluated the meta-analysis methods described above in the simulation scenarios described below. In all cases, K was set to 5, 10, 30 or 50.

- Scenario 1: the true causal effect was zero (i.e., $\beta = 0$), as in scenarios 2-6, and no small study effects. Therefore, equation (1) simplifies to $\hat{\beta}_j = \sigma_j \varepsilon_j$, where $\varepsilon_j \sim N(0,1,-\infty,\infty)$. Study size varied as follows: i) $n_1 = 100$, $n_2 = 1000$; and ii) $n_1 = 1000$, $n_2 = 5000$.

- Scenario 2: type (a) bias only, yielding the data generating model $\hat{\beta}_j = b_j + \sigma_j \varepsilon_j$, for $\varepsilon_j \sim N(0,1,-\infty,\infty)$. Study sizes ranged between $n_1 = 100$ and $n_2 = 5000$ (these values were also used in scenarios 3-6), and the proportion of biased studies δ was varied between 0 and 1 in steps of 0.1.

- Scenarios 3-5: Type (b) bias only, yielding the data generating model $\hat{\beta}_j = \sigma_j \varepsilon_j$, $\varepsilon_j | n_j \sim N(0,1,l_j,\infty)$, and $l_j = Q\left(1 - \min\left(f\left(\frac{n_j}{N}\right),1\right)\right)$ We assumed a linear (scenario 3), square root (scenario 4) or quadratic (scenario 5) relationship between p_j and n_j. N was set to 1 (i.e., no small study effects), 1500, 3000, 4500 and 6000.

- Scenario 6: type (b) bias only, assuming a step-function relationship between p_j and n_j. This used the same model as in Scenarios 3-5 except with p_j defined following equation (7) and where $p_{small} = 0.025$, $p_{medium} = 0.15$ and $p_{large} = 1$ (the latter implying that large studies have no P-value requirements for publication) were kept constant. Cut-offs to classify studies into small, medium or large varied as follows: i) $N_{small} = N_{large} = 1$ (i.e., no small study effects); ii) $N_{small} = 500$, and $N_{large} = 1000$; iii) $N_{small} = 1000$, and $N_{large} = 2000$; iii) $N_{small} = 2000$, and $N_{large} = 4000$.

- Scenario 7: Identical to scenario 1, except with $\beta = 0.02$.

The functional relationship between the bias (i.e., $E[\hat{\beta}_j|n_j] - \beta = b_j + \sigma_j E[\varepsilon_j|n_j]$) and n_j, and between standard error (i.e., $\sigma_j\sqrt{\mathrm{Var}[\varepsilon_j|n_j]}$) and n_j, for each one of the above scenarios is illustrated in Figure 2.

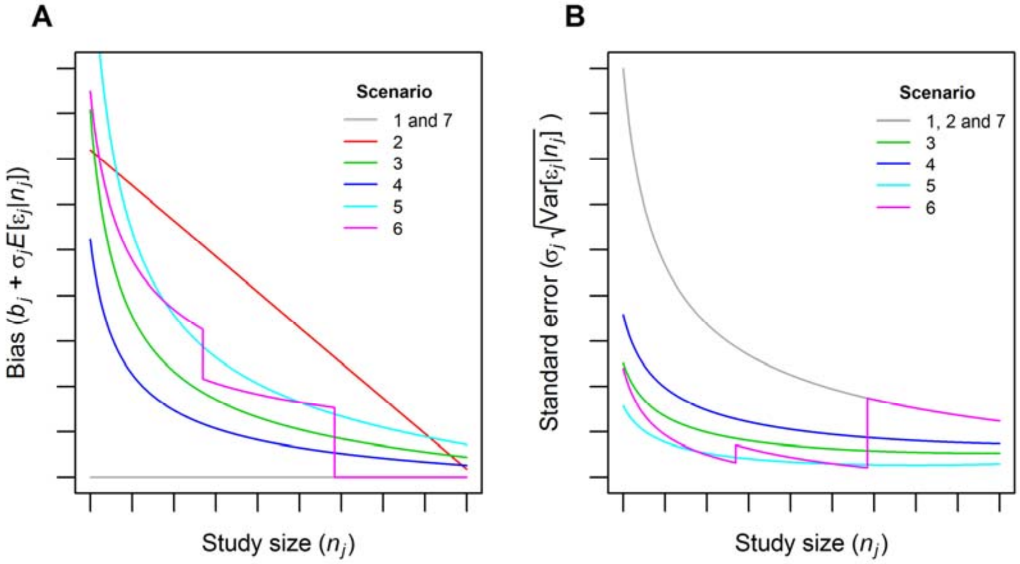

Figure 2. Illustration of the relationship between bias and n_j (panel A), and between standard error and n_j (panel B), induced by different models of small study effects.

2.5.4. Statistical analysis

In the simulation analysis, mean pooled effect estimates, standard errors, coverage and power of 95% confidence intervals were computed for the weighted mean, weighted median, MBE, Egger regression and trim-and-fill methods across 10,000 simulated datasets. Standard errors of the weighted median and the MBE are calculated using parametric bootstrap, which naturally incorporates any between-study heterogeneity into their confidence intervals.

All analyses were performed using R (www.r-project.org).

2.6. Applied examples

We further evaluated our proposed methods and illustrate their application by re-analysing three meta-analysis datasets:

- Catheter dataset: This meta-analysis, originally conducted by Veenstra et al.[10] evaluated 11 trials comparing chlorhexidine-silver sulfadiazine-impregnated vs. non-impregnated catheters with regards to risk of catheter-related bloodstream infection. These data presented a large correlation between effect estimates and their precision (r=0.76 [P-value=0.007]) (which translates into substantial asymmetry on the funnel plot), and high between-study heterogeneity (I^2=60%).

- Aspirin dataset: this meta-analysis, originally conducted by Edwards et al.,[11] evaluated 63 trials investigating the effect of a single dose of oral aspirin on pain relief (50% reduction in pain). r was also strong in magnitude (r=-0.70 [P-value=1.6×10^{-6}]), but there was low between-study heterogeneity (I^2=10%).

- Streptokinase dataset: meta-analysis originally conducted by Yusuf et al.[12] and updated by Egger et al.[3] of 21 trials evaluating the effect of streptokinase therapy on mortality risk. These data presented moderate heterogeneity (I^2=34%), but very little evidence of asymmetry (r=0.08, P-value=0.743). These data were used as a positive control, where all methods are expected to give similar answers.

3. Results

3.1. Simulation study

Simulation scenario 1 indicated that the confidence intervals of the fixed effects, weighted median and MBE are valid in the sense that they all achieve at least 95% coverage under the null (i.e., $\beta = 0$) and in absence of small study effects, although only the fixed effects method had exact 95% coverage (Supplementary Table 1). Egger regression presented under-coverage when the number of studies was small, but this attenuated as the number of studies increased. Conversely, the trim-and-fill method presented under-coverage that increased with number of studies, indicating that its confidence intervals are invalid (at least in our implementation of the method). The fixed effects method presented the smallest standard errors, followed by the trim-and-fill, which was slightly more precise than the weighted median. The MBE was less precise than the latter, but substantially more precise than Egger regression.

Supplementary Table 2 shows that scenario 2 lead to high values of I^2 and r. Under this small study effects model, the weighted median was less biased than the fixed effects model, and the MBE was the least biased among all methods (Figure 3). Those differences became more apparent as the number of studies increased. The trim-and-fill was more biased than the standard fixed effects model, and Egger regression substantially overcorrected.

Scenario 3 lead to high asymmetry, but did not substantially inflate I^2 (Supplementary Table 3), and the bias in the pooled estimates was much smaller compared to scenario 2. Again, Egger regression substantially overcorrected for small study effects, and the weighted median and MBE were less biased than the fixed effects model (Figure 4). However, the performance of the trim-and-fill relative to the weighted median and the MBE was substantially different than in scenario 2: here, it the number of studies is low ($K = 5$), the trim-and-fill performed similarly to the weighted median, but was more biased than the MBE; for $K = 5$, it outperformed the weighted median and performed similarly to the MBE; for larger values of K, the trim-and-fill was generally less biased than the other methods, unless all studies were affected by small study effects (in this case, $N = 6000$). However, as the number of studies increased, the trim-and-fill overcorrected for small study effects when $N = 1500$. In general, the differences between the weighted median, the MBE and trim-and-fill were much less marked in this scenario than in scenario 2; indeed, in scenario 3, the coverage of the weighted median and the trim-and-fill was similar for all values of K.

In scenario 4, small study effects resulted in a less marked asymmetry and in reduced I^2 – i.e., under-dispersion (Supplementary Table 4). In general, the results were similar to scenario 3 (as shown in Supplementary Figure 2), with two main differences. First, the weighted median presented better coverage than the trim-and-fill, unless $K = 50$ and $N = 4500$. Second, the overcorrection presented by the trim-and-fill in scenario 3 was much more apparent, especially for larger values of K. Scenario 5 was in between scenarios 2 and 3 regarding r and I^2 (Supplementary Table 5). In this scenario, the trim-and-fill was more biased than the weighted median and the MBE when the number of studies was low ($K = 5$ pr $K = 10$), and was in between them when there were more studies ($K = 30$ or $K = 50$). The difference between the weighted median and the MBE was small regardless of the number of studies (Supplementary Figure 3). In scenario 6, there was more between-study heterogeneity compared to the last scenario, but less than in scenario 2 (Supplementary Table 6). The weighted median and the MBE performed substantially better than the other methods (as shown in Supplementary Figure 4), with the MBE being close to unbiased in all cases when the number of studies was large ($K = 30$ or $K = 50$).

Supplementary Table 7 displays the performance of the methods to detect an effect in absence of small study effects (scenario 7). The fixed effects model was the most powered, followed by the trim-and-fill and the weighted median. Importantly, the trim-and-fill was slightly more precise than the weighted median, but presented substantially more power due to its under-coverage (which increased with number of studies and study size). The MBE was substantially more precise than Egger regression, but presented lower power due to under-coverage of the latter when the number of studies was low.

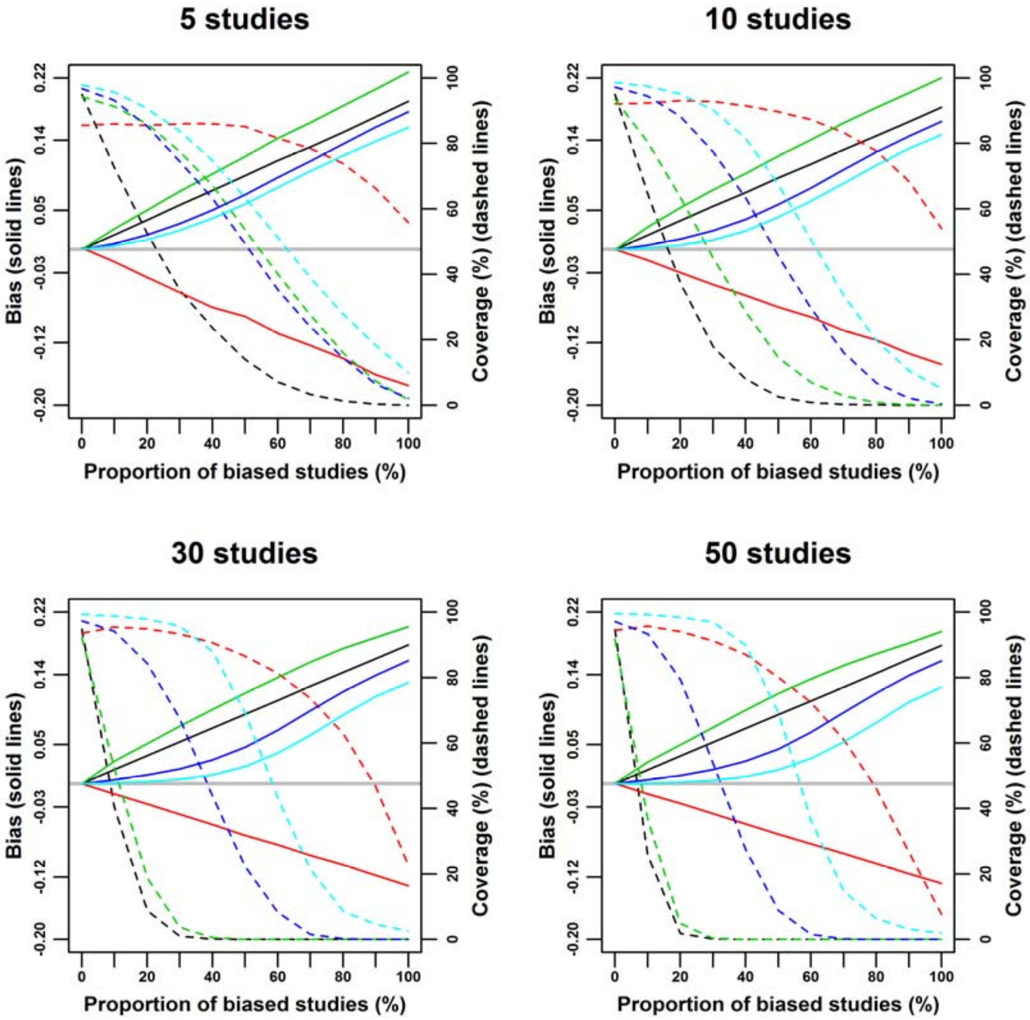

Figure 3. Bias (solid lines) and coverage (dashed lines) of the fixed effects (black), Egger regression (red), trim-and-fill (green), weighted median (dark blue) and mode-based estimate (light blue) under scenario 2: zero true effect (i.e., $\beta = 0$), small study effects through the bias term b_j, and study sizes uniformly ranging from 100 to 5000 individuals.

The grey line indicates zero bias.

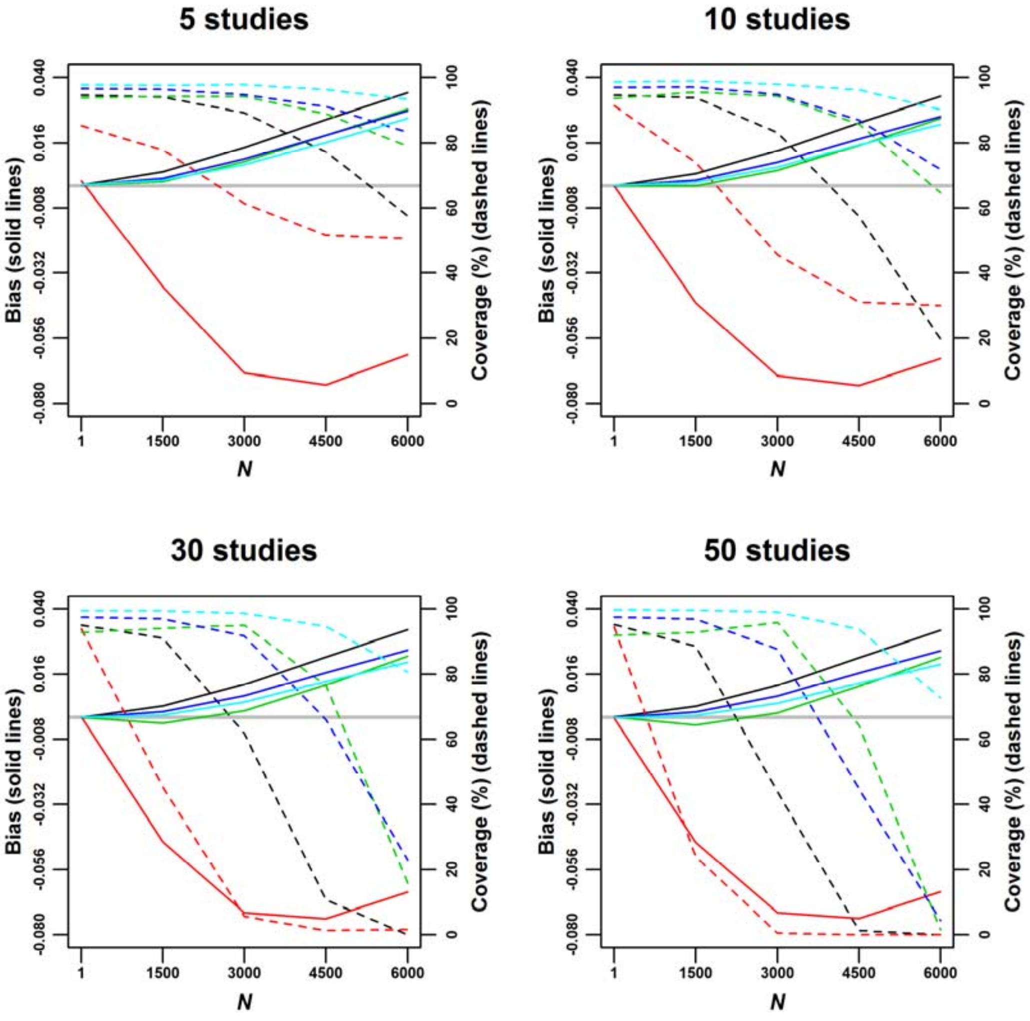

Figure 4. Bias (solid lines) and coverage (dashed lines) of the fixed effects (black), Egger regression (red), trim-and-fill (green), weighted median (dark blue) and mode-based estimate (light blue) under scenario 3: zero true effect (i.e., $\boldsymbol{\beta} = 0$), small study effects through dissemination bias (assuming a linear relationship between $\boldsymbol{p_j}$ and $\boldsymbol{n_j}$), and study sizes uniformly ranging from 100 to 5000 individuals.

p_j: maximum P-value allowed for publication for a study with n_j participants. N: study size threshold, with studies larger than or equally sized to N not being affected by small study effects.

The grey line indicates zero bias.

3.2. Real data examples

In our re-analysis of the catheter dataset (for which both r and I^2 were high), the fixed effects model yielded an odds ratio of bloodstream infection of 0.47 (95% CI: 0.38; 0.58), while the weighted median and the MBE yielded the same smaller estimate of 0.57 (95% CI: 0.44; 0.75). Trim-and-fill yielded 0.45 (95% CI: 0.31; 0.65), similar to the fixed effects results. Egger regression yielded a qualitatively different estimate of 1.27 (95% CI: 0.70; 2.31) (Figure 5, panel A). This is likely an over-correction (as in the simulation study), especially given that the individual-study odds ratio estimates in the data ranged from 0.09 to 0.83.

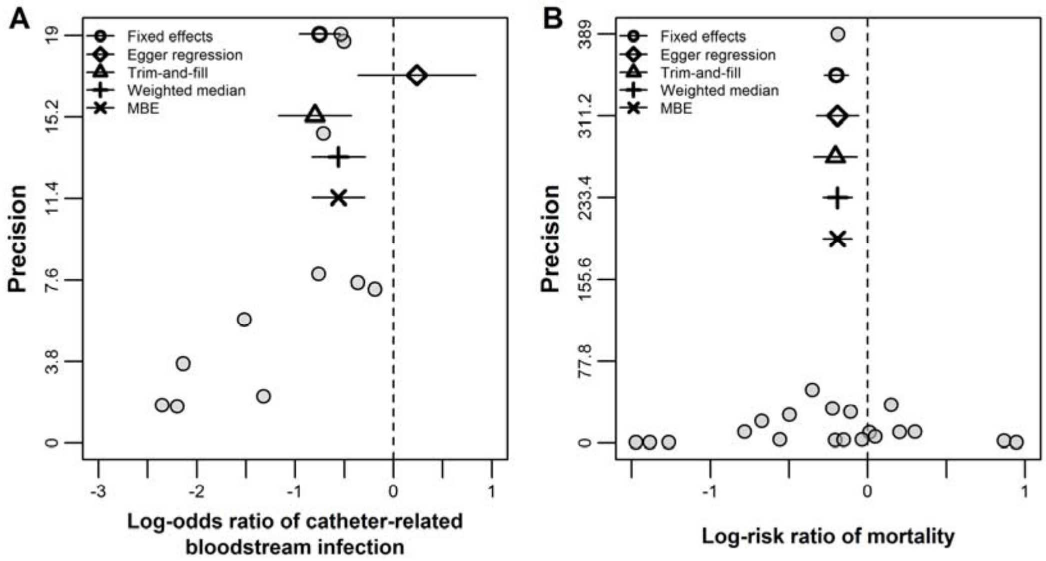

Figure 5. Funnel plots of the catheter (panel A) and streptokinase (panel B) meta-analyses, with pooled estimates and 95% confidence intervals from five meta-analysis methods.
MBE: Mode-bases estimate.

For the aspirin dataset (which presented low I^2 and marked asymmetry), the pooled odds ratio estimates of at least 50% of pain relief comparing active treatment to placebo were 3.43 (95% CI: 2.96; 3.98) for the fixed effect mean, 2.99 (95% CI: 2.41; 3.73) for the weighted median and 2.55 (95% CI: 1.78; 3.63) for the MBE. Trim-and-fill yielded an odds ratio of 2.87 (95% CI: 2.38; 3.47), which was in agreement with the weighted median and the MBE results. Egger regression yielded an odds ratio of 1.03 (95% CI: 0.71; 1.48), suggesting no effect of aspirin whatsoever (and again likely over-corrected).

For the streptokinase dataset (which was used as a positive control), the pooled risk ratio estimate comparing treatment and control groups was 0.82 (95% CI: 0.76; 0.88) for the fixed effects model. Results from the other four methods ranged from 0.81 to 0.83 (Figure 5, panel B). As a sensitivity analysis, the largest trial[13] (which corresponded to a substantial proportion of the total weight in the meta-analysis) was removed, which had no material effect on the results.

4. Discussion

The results above suggest that the weighted median and MBE give sensible answers to real meta-analyses where small study effects are suspected (even when Egger regression or trim-and-fill do not), as well as similar results to the fixed effects model and other meta-analysis methods in absence of bias. This corroborates the results from the simulation study, which indicated that these methods are less influenced by small study effects than the conventional fixed effects model and other established methods (see the Supplementary Text for a discussion on bias due to small study effects in Egger regression). Software for their implementation is provided in the Supplementary Material.

There are several strategies to investigate the presence and degree of small study effects in meta-analysis, all of which have limitations.[14,15] If, after careful examination, small study effects are suspected, we recommend that investigators apply the weighted median and the MBE in addition to standard methods as a sensitivity analyses. Our proposed approaches naturally reduce the influence of small studies without having to formally exclude them from the analysis. Exclusion often involves arbitrary study size cut-offs and artificially reduces the heterogeneity in the data.

When applying the weighted median and the MBE, it is important to not rely entirely on “statistical significance”, especially given that they will deliver estimates with less precision than the fixed effects model, but to examine their confidence intervals and assess their degree of overlap with the standard analysis. As a general rule, the weighted median and the MBE will give accurate and robust results when the majority of the weight in the analysis stems from studies that provide consistent effect estimates. This might, for example, be satisfied by just one or two large studies in a meta-analysis, despite the inclusion of many other biased studies. Conversely, they will give misleading results when the majority of the weight in the analysis stems from biased studies and, in the case of the MBE, the magnitude of the individual study biases are very similar (as illustrated in Figure 1, Panel D). The Cochrane Collaboration’s tool for assessing risk of bias[16] could be used as a guide to the likely proportion of biased studies in a given meta-analysis, and to the value of applying of these techniques. As such the proposed methodology is a natural extension of exploring between study heterogeneity due to perceived risk of bias,[16] which likely suffers from between rater subjectivity.

Importantly, the methods proposed here do not “correct” for asymmetry or heterogeneity between studies. Indeed, heterogeneity between studies should be expected,[17] and exploring if measured study characteristics account for the latter (e.g., via subgroup analyses and meta-regression) may yield important insights regarding treatment effect modification and/or potential sources of bias. This cannot be achieved by simply applying the proposed methods, nor any other method that yields a single pooled point estimate. This is especially relevant for the MBE method, which assumes that there is a subset of homogeneous studies that yield consistent estimates of the treatment effect.

Therefore, ideally the proposed methods would be applied if plausible effect modifiers do not account for observed heterogeneity between studies, or if there is residual heterogeneity within subgroups (although in this case the number of studies per subgroup may be prohibitive for meaningful comparisons between different estimators). Otherwise, the MBE can be used as a sensitivity analysis and interpreted as a test of the sharp null hypothesis; and, if the treatment effect can be assumed to be monotonic, as a test of the direction of the treatment effect. Finally, there are other waits to exploit the ZEMBA assumption (for example, by a model averaging approach[18]), and comparing their performance in plausible meta-analysis settings remains to be done.

Given its importance in summarising and interpreting the totality of available evidence, and their low cost compared to conducting a new study involving data collection, meta-analyses are particularly appealing to both researchers and journals, and a cornerstone of evidence-based medicine. Unfortunately, many systematic reviews and meta-analysis contain studies that are methodologically flawed and likely biased.[19] We are confident that the weighted median and MBE provide inferences that are robust to small study effects under a variety of reasonable simulation models and in real datasets likely affected by this bias. We hope that these simple and intuitive methods will be used to strengthen the conclusions of meta-analyses.

5. References

1. Egger, M., Smith, G.D. & Phillips, A.N. Meta-analysis: principles and procedures. *BMJ* 315, 1533-7 (1997).

2. Egger, M., Davey Smith, G. & Altman, D.G. *Systematic Reviews in Health Care: Meta-Analysis in Context* (John Wiley & Sons, New York, USA 2008).

3. Egger, M., Davey Smith, G., Schneider, M. & Minder, C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 315, 629-34 (1997).

4. Duval, S. & Tweedie, R. Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics* 56, 455-63 (2000).

5. Copas, J. What works?: selectivity models and meta analysis. *J R Stat Soc Ser A Stat Soc* 162, 95-109 (1999).

6. Bowden, J., Davey Smith, G., Haycock, P.C. & Burgess, S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet Epidemiol* 40, 304-14 (2016).

7. Hartwig, F.P., Davey Smith, G. & Bowden, J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *Int J Epidemiol* In press. doi: 10.1093/ije/dyx102(2017).

8. Bickel, D.R. & Frühwirth, R. On a fast, robust estimator of the mode: Comparisons to other robust estimators with applications. *Computational Statistics & Data Analysis* 50, 3500-3530 (2006).

9. Bickel, D.R. Robust and efficient estimation of the mode of continuous data: the mode as a viable measure of central tendency. *Journal of Statistical Computation and Simulation* 73, 899-912 (2002).

10. Veenstra, D.L., Saint, S., Saha, S., Lumley, T. & Sullivan, S.D. Efficacy of antiseptic-impregnated central venous catheters in preventing catheter-related bloodstream infection: a meta-analysis. *JAMA* 281, 261-7 (1999).

11. Edwards, J.E. *et al.* Single dose oral aspirin for acute pain. *Cochrane Database Syst Rev*, CD002067 (2000).

12. Yusuf, S. *et al.* Intravenous and intracoronary fibrinolytic therapy in acute myocardial infarction: overview of results on mortality, reinfarction and side-effects from 33 randomized controlled trials. *Eur Heart J* 6, 556-85 (1985).

13. Gruppo Italiano per lo Studio della Streptochinasi nell'Infarto Miocardico (GISSI). Effectiveness of intravenous thrombolytic treatment in acute myocardial infarction. *Lancet* 1, 397-402 (1986).

14. Guyatt, G.H. *et al.* GRADE guidelines: 5. Rating the quality of evidence--publication bias. *J Clin Epidemiol* 64, 1277-82 (2011).

15. Sterne, J.A. *et al.* Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ* 343, d4002 (2011).

16. Higgins, J.P. *et al.* The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 343, d5928 (2011).

17. Higgins, J.P. Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. *Int J Epidemiol* 37, 1158-60 (2008).

18. Burgess, S., Zuber, V., Gkatzionis, A., Rees, J.M.B. & Foley, C. Improving on a modal-based estimation method: model averaging for consistent and efficient estimation in Mendelian randomization when a plurality of candidate instruments are valid. *bioRxiv* (2017).

19. Ioannidis, J.P. The Mass Production of Redundant, Misleading, and Conflicted Systematic Reviews and Meta-analyses. *Milbank Q* 94, 485-514 (2016).