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

Short title: Meta-analysis using the median and the mode.

Fernando Pires Hartwig[1,2]*, George Davey Smith[2,3], Amand Floriaan Schmidt[4,5], Jonathan AC Sterne[2,3], Julian PT Higgins[2,3], Jack Bowden[2,6]

[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.

[6]University of Exeter College of Medicine and Health, Exeter, United Kingdom

*Corresponding author. Postgraduate Program in Epidemiology, Federal University of Pelotas, Pelotas (Brazil) 96020-220. Phone: 55 53 981126807. E-mail: fernandophartwig@gmail.com.

Abstract

Meta-analyses based on systematic literature reviews are commonly used to obtain a quantitative summary of the available evidence on a given topic. However, the reliability of any meta-analysis is constrained by that of its constituent studies. One major limitation is the possibility of small study effects, when estimates from smaller and larger studies differ systematically. Small study effects may result from reporting biases (i.e., publication bias), from inadequacies of the included studies that are related to study size, or from reasons unrelated to bias. We propose two estimators based on the median and mode to increase the reliability of findings in a meta-analysis by mitigating the influence of small study effects. By re-examining data from published meta-analyses and by conducting a simulation study, we show that these estimators offer robustness to a range of plausible bias mechanisms, without making explicit modelling assumptions. They are also robust to outlying studies without explicitly removing such studies from the analysis. When meta-analyses are suspected to be at risk of bias because of small study effects, we recommend reporting the mean, median and modal pooled estimates.

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

1. Introduction

Meta-analysis is used to obtain a quantitative summary of the evidence from multiple studies on a given topic, and is often undertaken as part of a systematic review.[1,2] In its archetypal form, meta-analysis provides an overall effect estimate for a well-defined intervention that has been assessed across several independent clinical trials, although it can also be applied to other study designs. Meta-analyses also provide an opportunity to explore between-study heterogeneity, which might highlight possible explanations for variation in treatment effects.[1,2]

Systematic differences between effect estimates from different studies may also indicate the presence of bias, which we wish to understand if possible and, ultimately, seek to remove from the analysis. Such differences may be due to flaws or limitations in the design, conduct or analysis of the included studies: for example, the failure of some randomized trials to conceal the allocation sequence from those recruiting participants, or the use of inappropriate imputation methods for missing endpoint data. The seriousness of these types of limitations may be associated with the size of the study, leading to a type of heterogeneity in which estimates from larger and smaller studies differ systematically.

A further threat to the validity of meta-analyses is publication bias[2], when the probability that results are reported and published is related to their direction or magnitude,[3] so that published results are a biased sample of all results generated. This bias is less likely to affect larger than smaller studies, due to a combination of pressure to publish by external funders or collaborators, the greater inherent publishing appeal of larger studies, and because an increased sample size raises the likelihood of achieving conventional statistical significance when the true treatment effect is non-zero.[2]

It is not necessarily the case that, in the presence of systematic differences associated with study size, smaller studies are less reliable than larger ones: Systematic differences between large and small studies may be to reasons other than bias: for example if the intervention was implemented more effectively in the smaller studies.[4] Therefore, the phenomenon where the reported treatment effect is associated with study size in a meta-analysis encompasses many different mechanisms, and is referred to with the umbrella-term “small study effects”.[5] It is difficult to identify whether an association between study size and reported treatment effect is due to true

heterogeneity, biases in the results of individual studies, selective reporting (or publication), or a combination of these.[2,6]

Many methods to detect and 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 was proposed more than 30 years ago.[7] Difficulties in visual interpretation of funnel plots motivated the development of tests for funnel plot asymmetry[4,8] and approaches that "correct" for asymmetry, such as regression and trim-and-fill estimators.[9-11] However, these approaches make either implicit or explicit assumptions about the selection process, so that their performance suffers when the true bias mechanism differs from that assumed.

Here, we propose two simple estimators that are robust to small study effects, whilst making no assumptions about their precise nature. They were originally proposed for causal inference in summary data Mendelian randomization[12,13]. From a statistical perspective, this technique has strong parallels with meta-analysis[14,15].

2. Meta-analysis datasets

Before presenting the estimators, we describe four meta-analysis datasets that will be used throughout the paper to explain the proposed estimators and illustrate their application. In addition to funnel plots (Figure 1), we characterise these datasets using the following statistics: i) Asymmetry, which we defined as the Egger test's coefficient (γ) – i.e., slope in inverse variance weighted linear regression of effect estimates on standard errors.[4] P-values were calculated using t-test with $K-2$ degrees of freedom, where K is the number of studies; and ii) Between-study inconsistency, defined as the conventional I^2 statistic. Importantly, the I^2 statistic does not quantify variation in the true effect sizes across studies, but rather statistical inconsistencies in the results of the studies.[16] For example, for a given data-generating mechanism producing a given amount of variation in the true effect sizes (i.e., heterogeneity), increasing the size of the studies will generally increase I^2 (because the study-specific confidence intervals will get narrower, thus increasing the statistical power to detect inconsistencies).

- Catheter dataset (Figure 1A): this meta-analysis, originally conducted by Veenstra et al.[17] 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 (γ=3.05 [P-value=0.007]) (which translates into substantial asymmetry on the funnel plot), and high between-study inconsistency (I^2=60%).

- Aspirin dataset (Figure 1B): this meta-analysis, originally conducted by Edwards et al.,[18] evaluated 63 trials investigating the effect of a single dose of oral aspirin on pain relief (50% reduction in pain). Asymmetry was also strong in magnitude (γ=2.11 [P-value=5.2×10^{-9}]), but there was low between-study inconsistency (I^2=10%).

- Sodium dataset (Figure 1C): this meta-analysis was originally conducted by Leyvraz et al.,[19] and assessed the effect of sodium intake on blood pressure in children and adolescents. We focused on the meta-analysis of 13 experimental studies (three of which were not randomized trials) of systolic blood pressure. Asymmetry was strong in magnitude (γ=2.60), but there was no strong statistical evidence against the null hypothesis of no asymmetry (P-value=0.679). Moreover, there was high between-study inconsistency (I^2=99%). As shown in section 4, both γ and I^2 are substantially attenuated upon removal of two studies classified as influential by Leyvraz et al.

- Streptokinase dataset (Figure 1D): this meta-analysis, originally conducted by Yusuf et al.[20] and updated by Egger et al.[4], includes 21 trials evaluating the effect of streptokinase therapy on mortality risk. These data presented moderate inconsistency (I^2=34%), but very little evidence of asymmetry (γ=-0.06, P-value=0.868). Given that in this dataset there is no strong indication of small study effects, these data were used as a positive control, where all estimators are expected to give similar answers.

3. Methods

We now give a non-technical explanation of our proposed estimators to motivate their utility. We then provide a more technical summary of our approach, by first describing the assumed data generating mechanism and the proposed estimation procedures.

3.1. Non-technical intuition

The standard way to combine studies in a meta-analysis is via a weighted mean of study-specific results, where the weight given to each study estimate is the inverse of its variance (thus reflecting its precision). Under the assumption that all included

studies provide valid estimates of the same underlying treatment effect, this “fixed effect” approach provides the summary estimate that is the most efficient – i.e., most precise and therefore with the highest power to detect a non-zero treatment effect.

Skewness affects the utility of the mean as a measure of central tendency. For example, the distribution of income is typically positively skewed due to the presence of a few individuals who are much wealthier than most of the population. In such cases, statistics such as the median or the mode are often used instead of the mean as central tendency measures to quantify “typical” income, although for some applications the mean will still be the statistic of interest.

Skewness in individual participant datasets is analogous to funnel plot asymmetry in meta-analyses. Examples of funnel plot asymmetry are shown in Figure 1 (panels A and B). In the aspirin dataset (Figure 1B), smaller studies have generally larger point estimates. Given that the mean is more sensitive to asymmetry than the median and the mode (and the same for their inverse variance weighted versions, described in section 3.3), estimates obtained using the latter two measures would be closer to the bulk of evidence in the meta-analysis. In cases where the strong asymmetry makes it implausible to discard the possibility of bias, estimators that yield combined estimates closer to the bulk of weights are likely to be more reliable.

A second situation where the mean may not be a useful central tendency statistic is when there are outliers. Using again the example of income, the mean income of a population will be largely influenced by the extreme wealth of a tiny proportion of individuals and will not reflect the typical income of the majority. Again, the median or the mode provide a central tendency statistic that is closer to most data points than the mean. The presence of a few outliers in a large population may not be problematic in typical studies using individual participant data, because their influence is diluted. However, a meta-analysis often contains a small number of data points (study results), increasing the relative influence of outliers on the combined estimate. For example, in the sodium dataset (Figure 1C), Leyvraz et al.[19], using a statistical criterion, classified two studies as outliers. In the Results section, we show that these two studies have a substantial influence in the results by pushing the weighted mean, but neither the median nor the mode, away from the bulk of the funnel plot.

We now provide a more formal justification for using the median or mode in the meta-analysis context to achieve robustness to small study effects and outlying studies, focusing on small study effects. We return to the topic of outlying studies when analysing real meta-analysis datasets.

3.2. Data generating mechanism

We first define a summary data generating mechanism with K studies indexed by j ($j = 1,2,\dots,K$) in a form that allows us to incorporate different types of small study effects (Box 1). We assume each study reports an estimated mean difference between groups (e.g., an experimental intervention and a standard intervention) denoted by $\hat{\beta}_j$, where:

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

Here:

- β is the average effect of the experimental compared with standard intervention 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_j and upper limit u_j;
- The parameters b_j, l_j, and u_j are all allowed to depend on the study size, n_j.

Standard meta-analysis models correspond to $l = -\infty$ and $u = \infty$, in which case ε_j denotes random error due to sampling variation. A conventional fixed-effects model would correspond to $b_j = 0$ for all studies, and a random-effects model to $b_j \sim N(0,\tau^2)$. This conventional random-effects distribution allows for between-study differences due to biases or due to other sources of heterogeneity; often it is not possible to distinguish one from the other.

Throughout this paper we assume a fixed treatment effect (as assumed by our proposed estimators, which do not explicitly model between-study heterogeneity), so that non-zero values of b_j occur only due to bias and not to other sources of heterogeneity. Small study effects are present if the biases b_j are correlated with study

sizes n_j. We recognize that not all systematic differences between small and large studies are due to differential bias. Small study effects also arise if b_j represents (non-bias-related) treatment effect heterogeneity that happens to be correlated with n_j. Our methods are not intended to address such situations. We discuss the practical application of the proposed estimators in the presence of heterogeneity in section 5.

Small study effects may also arise due to selective reporting and publication, which can be induced in our model by allowing the truncation limits for ε_j (i.e., l_j and u_j) to be correlated with n_j. In sections to follow, we will use b_j and the truncation limits for ε_j to induce different types of small study effects in the data, as described in Box 1. A general expression for the expected value of study j's effect estimate $\hat{\beta}_j$, based on n_j participants is:

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

3.3. Robust central tendency statistics in meta-analysis

We now introduce three estimators for β: the standard weighted mean plus two novel estimators, and discuss their ability to return consistent estimates under the assumed data generating mechanism. For the purposes of clarity only, we will assume throughout the remainder of section 3 that b_j is the sole source of bias in equation (1) – i.e., that $E[\varepsilon_j|n_j] = 0$.

3.3.1. The weighted mean

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

$$\hat{\beta}_{FE} = \frac{\sum_{j=1}^{K} \hat{\beta}_j \sigma_j^{-2}}{\sum_{j=1}^{K} \sigma_j^{-2}} \quad (3).$$

If even a single study contributes a biased estimate to the meta-analysis (e.g., via a non-zero b_j), then the combined mean will also be biased (unless the biases in different studies happen to cancel out). That is, using the notation of formula (1):

$E[\hat{\beta}_{FE}] \neq \beta$ in general, whenever $b_j \neq 0$ for at least one 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. The exception would be in a situation where b_j is negative for some studies and positive for other studies, such that the net bias is zero – i.e., $\sum_{j=1}^{K} b_j \sigma_j^{-2} = 0$.

3.3.2. The weighted median

The weighted median[12] 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 $\hat{\beta}_1 \leq \hat{\beta}_2 \ldots \leq \hat{\beta}_K$. Let the standardized inverse-variance weight 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. Let $s_j = \sum_{g=1}^{j} w_g$ denote the sum of standardized 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 weighted empirical distribution, so it will be equal to $\hat{\beta}_j$if $s_j = 0.5$. In practice, no study lies exactly at the 50th percentile, so this quantity is estimated 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, provided that both $\hat{\beta}_{j^*}$ and $\hat{\beta}_{j\dagger}$ are consistent for β, the $\hat{\beta}_{WM}$ is consistent. This implies that, as the number of studies grows indefinitely large, the $\hat{\beta}_{WM}$ is consistent if up to (but not including) 50% of the total weight in the analysis comes from biased studies – i.e., $\left(\sum_{j=1}^{K} I(b_j > 0)\, w_j\right) < 50\%$. This means that the weighted median has a breakdown level of 50%. Of note, if b_j is negative for some

studies and positive for other studies, it is possible that both $\hat{\beta}_{j^*}$ and $\hat{\beta}_{j^\dagger}$ are consistent for β even if more than 50% of the weight comes from biased studies.

3.3.3. The mode-based estimate

The mode-based estimate[13] (MBE) exploits an assumption we refer to as the zero modal bias assumption (ZEMBA). This states 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 consistent for β, even if the majority of $\hat{\beta}_j$'s are biased.

More formally, $\hat{\beta}_{MBE}$ is consistent if $\omega_0 > \max(\omega_1, \omega_2, \ldots, \omega_v)$, where $\omega_0 = \sum_{j=1}^{K} w_j I(b_j = 0)$ denotes the sum of weights provided by studies with zero bias, and ω_1, ω_2 and ω_v are the sum of weights provided by studies that have the smallest, the second smallest and the largest identical bias terms, respectively. For example, suppose that there are 10 studies and $b_1 = b_2 < b_3 < b_4 = b_5 = b_6 = 0 < b_7 = b_8 = b_9 < b_{10}$. In this case, $\omega_0 = \sum_{j\in\{4,5,6\}} w_j$, $\omega_1 = \sum_{j\in\{1,2\}} w_j$, $\omega_2 = \sum w_3$, $\omega_3 = \sum_{j\in\{7,8,9\}} w_j$ and $\omega_4 = \sum w_{10}$.

It is possible to exploit ZEMBA in different ways. Here, as in Hartwig et al.,[13] we use 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:

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

where h is the smoothing bandwidth parameter.[21] This parameter regulates a bias-variance trade-off, with smaller values of h reducing both bias and precision. Given that the error terms in equation (1) were drawn from a standard truncated normal distribution, a normal kernel is expected to yield adequate density estimates.

Silverman's rule is commonly used with a normal kernel to calculate h. We used the modified Silverman's bandwidth selection rule proposed by Bickel et al.[22], which reduces the influence of outliers compared with the conventional Silverman's rule:

$$h = \frac{0.9\min\left(\mathrm{sd}(\hat{\beta}_j), 1.4826\mathrm{mad}(\hat{\beta}_j)\right)}{L^{\frac{1}{5}}} \quad (6),$$

where $\mathrm{sd}(\hat{\beta}_j)$ and $\mathrm{mad}(\hat{\beta}_j)$ respectively denote the standard deviation and the median absolute deviation of the median of the study-level point treatment effect estimates.

The exact breakdown level of the MBE depends on $\max(\omega_1, \omega_2, \ldots, \omega_v)$, which is unknown. If all biased studies estimate exactly the same effect parameter, then ZEMBA will only be satisfied if up to (but not including) 50% of the weight 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%.

3.3.4. Standard errors for the weighted median and the MBE

Standard errors of the weighted median and the MBE can be calculated using parametric bootstrap, which naturally incorporates any between-study heterogeneity. More specifically, suppose that R bootstrap iterations are to be performed. For each iteration $r \in \{1, \ldots, R\}$, the bootstrapped point estimate from the jth study ($\hat{\beta}_j^r$) is sampled from the normal distribution $\hat{\beta}_j^r \sim N(\hat{\beta}_j, \sigma_j^2)$. Then each estimator is applied to the current set $\{\hat{\beta}_j^r\}_{j=1}^K$ of resampled point estimates, generating $\hat{\beta}_{WM}^r$and $\hat{\beta}_{MBE}^r$. Repeating this step R times yields the sets $\{\hat{\beta}_{WM}^r\}_{r=1}^R$and $\{\hat{\beta}_{MBE}^r\}_{r=1}^R$, which are empirical sampling distributions of the weighted median and the MBE, respectively. We used a robust standard deviation estimator (the median absolute deviation from the median corrected for asymptotic normal consistency) to calculate the standard deviation from each empirical distribution, which is an estimate of the standard error. Finally, these can be used to calculate confidence intervals based on a normal approximation.

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

Figure 2 illustrates the assumptions underlying the combined mean, median and mode in a hypothetical meta-analysis of 10 studies, sorted in ascending order of their β_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:

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

- when 4 out of 10 studies are biased (Panel B), 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; and
- when more than 50% of the studies are biased (Panel D) and ZEMBA is violated, all estimators are biased.

An attractive property of the weighted median and MBE is that they are naturally robust central tendency statistics, but do not make any specific assumptions about the selection mechanism at play. Therefore, they are robust to a range of possible causes of small-study effects. However, as Figure 2 illustrates, these estimators are not guaranteed to provide consistent estimates of β, failing to do so when their identifying assumptions are violated. Nevertheless, the assumptions they require are weaker than those of the standard weighted mean.

3.6. Regression-based extrapolation and trim-and-fill

We compared the weighted median and the MBE with two meta-analysis estimators developed to adjust for small-study bias. The first, described by Moreno et al.[11], is extrapolation to the estimated effect of intervention in a study of infinite size based on a linear regression weighted by σ_j^{-2}. This estimator assumes a linear relationship between the b_j's and σ_j's, so that $b_j = \beta_1 \sigma_j$. Plugging this expression for b_j in equation (1) to the regression-based extrapolation model yields:

$$\hat{\beta}_j = \beta_0 + \beta_1 \sigma_j + \sigma_j \varepsilon_j \qquad (6).$$

In equation (6), β_0 is the estimated effect in a study of infinite size, obtained by extrapolation based on the model assumptions. β_1 is the parameter that allows accounting for bias via non-zero b_j's, so that testing H_0: $\beta_1 = 0$ is a test for the presence of small-study effects. Indeed, this test has been shown[5] to be identical to the test of funnel plot asymmetry proposed by Egger et al.[4] For simplicity, equation (6) shows the fixed-effect regression-based extrapolation model, which can be extended into an additive or multiplicative random effects model[23]; the latter was used in the simulations

and real data examples described below. This approach is illustrated in Supplementary Figure 1 (panels A and C).

The second estimator is trim-and-fill, a non-parametric data augmentation method that estimates the number of missing studies (for example, due to publication bias) by supressing (or “trimming”) the most extreme studies from one side of the funnel plot. Then, the data is augmented so that the funnel plot is more symmetric. The augmented data is then used to calculate the combined effect[10]. In our simulations and real data examples, we used a random effects model throughout the trim-and-fill process (sometimes referred to as random-random effects trim-and-fill) and the $L0$ estimator to estimate the number of missing studies. This approach is illustrated in Supplementary Figure 1 (panels B and D).

4. Re-analysis of published meta-analyses

To illustrate the application of the proposed meta-analysis estimators and compare them with existing approaches, we re-analysed the four meta-analysis datasets described in section 2 (Table 1).

In our re-analysis of the catheter dataset (for which both r and I^2 were high), the weighted mean 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 (in magnitude) 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 weighted mean results. Regression-based extrapolation yielded a qualitatively different estimate of 1.27 (95% CI: 0.70; 2.31). This is likely an over-correction, especially given that the individual-study odds ratio estimates in the data ranged from 0.09 to 0.83 (Supplementary Figure 1B).

For the aspirin dataset (which presented low I^2 and marked asymmetry), the combined 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 weighted 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 closer to the weighted median and the MBE results than to the weighted mean. Regression-based extrapolation yielded an odds ratio of 1.03 (95% CI: 0.71; 1.48), suggesting no effect of aspirin whatsoever (and again likely over-corrected – Supplementary Figure 1D).

For the sodium dataset, removing the outlying study (as classified by Leyvraz et al.[19]) with the largest weight reduced between-study inconsistency (I^2=87%). Removing both studies eliminates between-study inconsistency (I^2=0%). Removing these studies also substantially attenuates asymmetry (γ=-0.22 and γ=0.68, respectively). This suggests that, unlike the previous examples, between-study inconsistency and asymmetry mostly stemmed from two studies (out of 13). Without removing any studies, the weighted mean, regression-based extrapolation and trim-and-fill estimators suggested an average decrease in systolic blood pressure due to sodium intake-lowering interventions of 1.48 (95% CI: 1.39; 1.57), 1.24 (95% CI: -0.72; 3.21) and 2.62 (95% CI: 0.99; 4.26) mmHg, respectively. All these results are higher than the bulk of the funnel plot (Figure 1C). Conversely, the weighted median and the MBE yielded combined estimates of 0.62 (95% CI: 0.52; 0.72) and 0.61 (95% CI: 0.51; 0.70), respectively, which is in line with the majority of the studies and located within the bulk of the plot. Indeed, these results are similar to those obtained by Leyvraz et al.[19] after they explicitly excluded these two studies from the meta-analysis.

For the streptokinase dataset (which was used as a positive control), the combined risk ratio estimate comparing treatment and control groups was 0.82 (95% CI: 0.76; 0.88) using the weighted mean. Results from the other four estimators ranged from 0.81 to 0.83. Given that the largest trial[24] corresponded to a substantial proportion of the total weight in the meta-analysis, the observed consistency between the estimators could simply be that they were all driven by this large study. However, in a sensitivity analysis where this study was removed, there was no material effect on the results. Therefore, the observed consistency between the approaches in this example was likely due to the symmetry of the data rather than to the effect of a single large trial.

The results above indicate that the weighted median and the MBE are less influenced by outlying studies compared to the weighted mean, regression-based extrapolation and trim-and-fill. This is a useful property at least for sensitivity analysis purposes, especially for meta-analyses with a small number of datapoints (and thus more sensitive to outliers). The proposed estimators appeared more robust to the presence of small study effects than the weighted mean. In a dataset with substantial asymmetry but low between-study inconsistency (the aspirin dataset), the weighted median and the MBE gave similar results to the trim-and-fill estimator. In a dataset with substantial asymmetry and between-study inconsistency (the catheter dataset), the proposed

estimators were less influenced by the left skew in the funnel plot than the trim-and-fill, which gave very similar results to the weighted mean. This suggests that presence of between-study inconsistency has a more limited effect on the robustness of the proposed estimators to small-study effects than on trim-and-fill (in the simulations, these estimators are compared in scenarios with varying degrees of between-study inconsistency). In the datasets with asymmetry, regression-based extrapolation yielded results that were likely overcorrected.

5. Simulation study

5.1. Brief description

We performed a simulation study to evaluate the performance of the weighted mean, regression-based extrapolation, 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 standard error of the mean difference is one for all values of j, and $\sigma_j = \sqrt{4/n_j}$. We assume that studies range in size from n_1 to n_2 uniformly, so that $n_j \sim \mathrm{Uniform}(n_1, n_2)$.

Data were generated to contain two forms of bias (see Box 1 for their general principles). Type (a) bias, is a fundamental property of each study (e.g., bias due to lack of intervention allocation concealment, or residual confounding in the case of meta-analyses of observational studies). For simulations under type (a) bias, the proportion of biased studies is dictated by the parameter $\delta \in [0,1]$. Among biased studies, b_j varies linearly with n_j.

Type (b) bias, is the result of publication bias, not a property of each study. For simulations under type (b) bias, we assumed (in common with most publication bias models) that results achieving conventional levels of statistical significance are more likely to be published. Therefore l_j was defined to correspond to the maximum one-sided P-value (null hypothesis: true mean difference ≤0) allowed for publication for a given study size (p_j), up to N. That is, for $n_j \geq N$, then there are no P-value requirements for publication – i.e., $p_j = 1$ for all values of j (because the study size is sufficient for publication regardless of its results). For $n_j < N$, larger studies are more

likely to be published than smaller studies, where p_j is a non-decreasing function of n_j. Therefore, N is the minimal study size required for the P-value to have no influence on the publication probability, which can be used to increase (if N is larger) or decrease (if N is smaller) the degree of type (b) bias. We evaluated four distinct functions: linear, square root, quadratic and step function. The relationship between p_j and n_j in each one of these four type (b) bias mechanisms is illustrated in Supplementary Figure 2.

In all simulations, K was set to 5, 10, 30 or 50. In Scenarios 1-6, $\beta = 0$. In scenario 1, there is neither type (a) nor type (b) bias. Scenario 2 evaluated type (a), but not type (b), bias. Scenarios 3-6 evaluated type (b) bias (but not type (a) bias), where p_j was a linear, square root, quadratic or step function (respectively) of n_j, up to N. Scenario 7 was identical to Scenario 1, except that $\beta = 0.02$. Table 2 describes the main characteristics and aims of each scenario.

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, in scenarios 1-7 is illustrated in Figure 3.

The data generating mechanism and simulation parameters are described in more detail in the Supplement. Mean combined effect estimates, standard errors, coverage and rejection rates of 95% confidence intervals were computed for the weighted mean, weighted median, MBE, regression-based extrapolation and trim-and-fill estimators across 10,000 simulated datasets. All analyses were performed using R.[25] We used the "metafor" package to calculate the I^2 statistic and to perform the weighted mean and trim-and-fill estimators.[26] The "truncnorm" package was used to generate random draws from the standard truncated normal distribution.[27] The "doParallel" package was used for parallel computing.[28]

5.2. Simulation study results

Simulation scenario 1 showed that confidence intervals for the weighted mean, weighted median and MBE are valid under the null in the sense that they all achieve at least 95% coverage when $\beta = 0$ and in absence of small study effects, although only the weighted mean had exact 95% coverage (Supplementary Figure 3 and Supplementary Table 1). Regression-based extrapolation showed under-coverage

when the number of studies was small, but this attenuated as the number of studies increased. Conversely, trim-and-fill showed under-coverage that increased with number of studies, indicating that its confidence intervals are invalid (at least in our implementation of the estimator). The weighted mean had smallest standard errors, followed by 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 regression-based extrapolation.

Supplementary Table 2 shows that scenario 2 leads to high values of I^2 and γ. Under this scenario the weighted median was less biased than the weighted mean, and the MBE was the least biased among all approaches (Figure 4). Those differences became more apparent as the number of studies increased. Trim-and-fill was more biased than the standard weighted mean, and regression-based extrapolation substantially overcorrected for the bias.

Scenario 3 leads to high asymmetry, but not a substantial inflation of I^2 (Supplementary Table 3), and the bias in the combined estimates was much smaller than in scenario 2. Again, regression-based extrapolation substantially overcorrected for small study effects, and the weighted median and MBE were less biased than the weighted mean (Figure 5). However, the performance of trim-and-fill relative to the weighted median and the MBE was substantially different than in scenario 2: if the number of studies is low ($K = 5$), trim-and-fill performed similarly to the weighted median, but was more biased than the MBE; for $K = 10$, it outperformed the weighted median and performed similarly to the MBE; for larger values of K, trim-and-fill was generally less biased than the other estimators, unless all studies were affected by small study effects (in this case, $N = 6000$). However, as the number of studies increased, 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 scenario 3 than in scenario 2; indeed, the coverage of the weighted median and trim-and-fill was similar for all values of K.

In scenario 4, small study effects resulted in less marked asymmetry than for scenario 3 and in reduced I^2 – i.e., under-dispersion (Supplementary Table 4). In general, results were similar to scenario 3 (see Supplementary Figure 4), with two main differences. First, the weighted median had better coverage than trim-and-fill, unless

$K = 50$ and $N = 4500$. Second, the overcorrection showed by trim-and-fill in scenario 3 was more apparent, especially for larger values of K. Scenario 5 was in between scenarios 2 and 3 regarding γ and I^2 (Supplementary Table 5). In this scenario, trim-and-fill was more biased than the weighted median and the MBE when the number of studies was low ($K = 5$ or $K = 10$), and 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 5). In scenario 6, there was more between-study inconsistency compared with the last scenario, but less than in scenario 2 (Supplementary Table 6). The weighted median and the MBE performed substantially better than the other estimators (as shown in Supplementary Figure 6), 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 and Supplementary Figure 7 display the performance of the estimators in detecting an effect in absence of small study effects (scenario 7). The weighted mean was the estimator with the highest power to detect a non-zero treatment effect, followed by trim-and-fill and the weighted median. Importantly, trim-and-fill was slightly more precise than the weighted median, but had substantially more power due to its under-coverage (which increased with number of studies and study size). The MBE was substantially more precise than regression-based extrapolation, but had lower power due to under-coverage of the latter when the number of studies was low.

Our simulation study corroborated the well-known notion that the weighted mean is biased in the presence of small study effects (either type (a) and type (b)). In all small study effects mechanisms, regression-based extrapolation overcorrected the treatment effect (this is discussed in more detail in the Supplementary Text). Trim-and-fill was more biased in the presence of type (a) bias (which lead to substantial between-study inconsistency) than the weighted mean. Trim-and-fill was less affected by type (b) than type (a) bias, thus highlighting the dependence of this estimator to the underlying data generating mechanism. Moreover, for most variations of type (b) bias, this estimator presented more bias than the weighted median and the MBE, as well as under coverage in the absence of any small study effects. Conversely, the weighted median and the MBE had confidence intervals with coverage ≥95% in the absence of small study effects and were relatively robust to both type (a) and type (b) bias.

6. Discussion

We have proposed the weighted median and the MBE for meta-analysis as approaches that are robust to the presence of small-study effects and outliers. Application to a series of examples indicates that both approaches give sensible estimates of the intervention effect in real meta-analyses where small study effects are suspected, even when regression-based extrapolation or trim-and-fill do not. They also give similar results to the weighted mean and other meta-analysis approaches in absence of bias. Our real data examples also illustrated the robustness of the proposed estimators to outliers. Our comprehensive simulation study confirmed these findings, and showed that these estimators are less influenced by small study effects than the conventional weighted mean and previously proposed approaches to estimate intervention effects in the presence of small study effects. 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.[6,29] If, after careful examination, small study effects are suspected, we recommend that investigators apply the weighted median and the MBE in addition to standard estimators as sensitivity analyses. These estimators reduce the influence of small and/or outlying studies without excluding them formally from the meta-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 not to rely entirely on “statistical significance”, especially given that they are less precise than the weighted mean. Instead, meta-analysis authors should examine confidence intervals for the different estimators and assess their consistency with standard meta-analysis estimates. In general, the weighted median and the MBE will be robust when studies that provide consistent effect estimates receive most of the weight in the fixed-effect meta-analysis. This might occur in a meta-analysis with just one or two large studies providing consistent estimates, despite the inclusion of many other biased, smaller studies. Conversely, the weighted median and the MBE 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 2, Panel D). The Cochrane tool for assessing risk of bias in

randomized trials[30] could be used as a guide to the likely proportion of biased studies in a given meta-analysis, and to the value of applying these techniques. As such the proposed estimators is a natural extension of exploring between study heterogeneity due to perceived risk of bias.[30]

We have presented the weighted median and the MBE assuming treatment effect homogeneity. Under this assumption, any heterogeneity between effect estimates is indicative of bias. By supplementing this assumption with additional assumptions (such as the 50% rule for the weighted median or ZEMBA for the MBE) then allows consistent estimation of the treatment effect even in the presence of (some forms of) small study effects. In the absence of bias and heterogeneous treatment effects the weighted mean, weighted median and MBE estimate the inverse-variance weighted average, median and modal treatment effects (in the case of unique treatment effects for each study, the latter would simply be the most precise effect estimate). However, in this case, a more sensible approach would be to perform a random effects meta-analysis to estimate the average treatment effect. But doing so requires the assumption that there is no form of bias in any of the studies included in the meta-analysis (because all studies have non-zero weight in the meta-analysis), which may itself not be warranted at least in some applications. This illustrates the more general notion that relaxing one assumption often requires more contrived versions of one or more other assumptions. In this case, assuming absence of bias allows interpreting systematic differences between studies as being solely due to treatment effect heterogeneity (and thus a random effects weighted mean can be used to estimate the average treatment effect), while assuming treatment effect homogeneity allows interpreting such differences as being solely due to bias (and thus estimators such as the weighted median and the MBE can be used to estimate the treatment effect under some forms of bias).

Importantly, the proposed estimators here cannot be regarded as providing a general “correction” for funnel plot asymmetry or heterogeneity between studies. Heterogeneity between studies should be expected in real meta-analyses,[31] and exploring whether it is explained by measured study characteristics (e.g., via subgroup analyses and meta-regression) may yield important insights regarding treatment effect modification and/or potential sources of bias. Such insights cannot be achieved by simply applying the proposed estimators, nor any other approach that yields a single point estimate. This is especially relevant for the MBE estimator, which assumes that there is a subset of

homogeneous studies that yield consistent estimates of the treatment effect. Therefore, ideally the proposed estimators 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 (i.e., the hypothesis that the intervention has no effect whatsoever on anyone in the population). Supplementation with further assumptions would allow some learning about the average treatment effect. For example, if the true treatment effect can be assumed to be monotonic (i.e. in the same direction for all studies), then the MBE can be interpreted as a test of the direction.

As mentioned in section 3.3.3, the MBE is just one way of exploiting the ZEMBA assumption to mitigate the influence of small study effects in meta-analysis. There are many other ways of estimating the mode of continuous data, such as the half-sample mode method,[21] Grenander's estimators,[32] model-averaging[33] and explicit selection[34] methods. Even restricting to only kernel-based methods such as the MBE, there are many available choices of bandwidths and kernels. It is therefore possible that there are estimators more adequate than the MBE to exploit the ZEMBA assumption in meta-analysis, a topic that remains to be investigated. The goal of the present study was to present ZEMBA as an alternative identification assumption, and compare the performance of one estimator that relies on this assumption (the MBE) against established meta-analysis estimators.

In summary, many systematic reviews and meta-analyses contain studies that are methodologically flawed and likely biased.[35] We have proposed new weighted median and mode-based estimators that provide inferences that are robust to small study effects under a variety of reasonable simulation scenarios. Their application in real datasets supports their likely utility as a sensitivity analysis in comparison to standard mean-based meta-analytic estimates. We hope that these estimators will be used to strengthen the conclusions of systematic reviews and meta-analyses.

Data availability statement: Data sharing is not applicable to this article as no new data were created or analysed in this study.

Funding: This work was coordinated by researchers working within the Medical Research Council (MRC) Integrative Epidemiology Unit, which is funded by the MRC and the University of Bristol (grant ref: MC_UU_00011/1, MC_UU_00011/5). During this work, FPH was supported by a Brazilian National Council for Scientific and Technological Development (CNPq) postdoctoral fellowship (process number: 153134/2018). No funding body has influenced data collection, analysis, or interpretation.

Conflicts of interest: The authors declare no conflicts of interest.

7. References

1. Egger M, Smith GD, Phillips AN. Meta-analysis: principles and procedures. *BMJ*. 1997;315(7121):1533-1537.

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

3. Dickersin K, Chan S, Chalmers TC, Sacks HS, Smith H, Jr. Publication bias and clinical trials. *Control Clin Trials*. 1987;8(4):343-353.

4. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ*. 1997;315(7109):629-634.

5. Sterne JA, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *J Clin Epidemiol*. 2000;53(11):1119-1129.

6. Sterne JA, Sutton AJ, Ioannidis JP, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ*. 2011;343:d4002.

7. Light R, Pillemer D. Summing up. The science of reviewing research. Cambridge, MA, USA and London, UK Harvard University Press; 1984.

8. Harbord RM, Egger M, Sterne JA. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Stat Med*. 2006;25(20):3443-3457.

9. Copas J. What works?: selectivity models and meta analysis. *J R Stat Soc Ser A Stat Soc*. 1999;162:95-109.

10. Duval S, Tweedie R. Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*. 2000;56(2):455-463.

11. Moreno SG, Sutton AJ, Ades AE, et al. Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Med Res Methodol*. 2009;9:2.

12. Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet Epidemiol*. 2016;40(4):304-314.

13. Hartwig FP, Davey Smith G, Bowden J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *Int J Epidemiol*. 2017;46(6):1985-1998.

14. Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet Epidemiol*. 2013;37(7):658-665.

15. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol*. 2015;44(2):512-525.

16. Borenstein M, Higgins JP, Hedges LV, Rothstein HR. Basics of meta-analysis: I(2) is not an absolute measure of heterogeneity. *Res Synth Methods*. 2017;8(1):5-18.

17. Veenstra DL, Saint S, Saha S, Lumley T, Sullivan SD. Efficacy of antiseptic-impregnated central venous catheters in preventing catheter-related bloodstream infection: a meta-analysis. *JAMA*. 1999;281(3):261-267.

18. Edwards JE, Oldman A, Smith L, et al. Single dose oral aspirin for acute pain. *Cochrane Database Syst Rev*. 2000(2):CD002067.

19. Leyvraz M, Chatelan A, da Costa BR, et al. Sodium intake and blood pressure in children and adolescents: a systematic review and meta-analysis of experimental and observational studies. *Int J Epidemiol*. 2018.

20. Yusuf S, Collins R, Peto R, 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*. 1985;6(7):556-585.

21. Bickel DR, Frühwirth R. On a fast, robust estimator of the mode: Comparisons to other robust estimators with applications. *Computational Statistics & Data Analysis*. 2006;50(12):3500-3530.

22. Bickel DR. 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*. 2002;73(12):899-912.

23. Bowden J, Jackson C. Weighing Evidence "Steampunk" Style via the Meta-Analyser. *Am Stat*. 2016;70(4):385-394.

24. Gruppo Italiano per lo Studio della Streptochinasi nell'Infarto Miocardico (GISSI). Effectiveness of intravenous thrombolytic treatment in acute myocardial infarction. *Lancet*. 1986;1(8478):397-402.

25. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2018. URL: https://www.R-project.org/.

26. Viechtbauer W. Conducting meta-analyses in R with the metafor package. Journal of Statistical Software. *J Stat Soft*. 2010;36(6):1-48.

27. Mersmann O, Trautmann H, Steuer D, Bornkamp B. truncnorm: Truncated Normal Distribution. R package version 1.0-8. 2018. URL: https://CRAN.R-project.org/package=truncnorm.

28. Microsoft Corporation, Weston S. doParallel: Foreach Parallel Adaptor for the 'parallel' Package. R package version 1.0.11. 2017. URL: https://CRAN.R-project.org/package=doParallel.

29. Guyatt GH, Oxman AD, Montori V, et al. GRADE guidelines: 5. Rating the quality of evidence--publication bias. *J Clin Epidemiol*. 2011;64(12):1277-1282.

30. Sterne JAC, Savovic J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ*. 2019;366:l4898.

31. Higgins JP. Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. *Int J Epidemiol*. 2008;37(5):1158-1160.

32. Grenander U. Some direct estimates of the mode. *Ann Math Stat*. 1965;36:131-138.

33. Burgess S, Zuber V, Gkatzionis A, Foley C. Modal-based estimation via heterogeneity-penalized weighting: model averaging for consistent and efficient estimation in Mendelian randomization when a plurality of candidate instruments are valid. *Int J Epidemiol*. 2018;47(4):1242-1254.

34. Windmeijer F, Liang X, Hartwig FP, Bowden J. The Confidence Interval Method for Selecting Valid Instrumental Variables. *Discussion Paper 19/715*. 2019;Available at: https://ideas.repec.org/p/bri/uobdis/19-715.html.

35. Ioannidis JP. The Mass Production of Redundant, Misleading, and Conflicted Systematic Reviews and Meta-analyses. *Milbank Q*. 2016;94(3):485-514.

Table

Table 1. Combined estimates with 95% confidence intervals for different meta-analysis datasets and estimators.

Estimator	**Dataset**			
	Catheter[A]	**Aspirin[A]**	**Sodium[B]**	**Streptokinase[C]**
	I^2=60%	I^2=10%	I^2=99%[D]	I^2=34%
	γ=-3.05 (P=0.007)	γ=2.11 (P=5.2×10^{-9})	γ=2.60 (P=0.679)[D]	γ=-0.06 (P=0.868)
Weighted mean	0.47 (0.38; 0.58)	3.43 (2.96; 3.98)	1.48 (1.39; 1.57)	0.82 (0.76; 0.88)
Regression-based extrapolation	1.27 (0.70; 2.31)	1.03 (0.71; 1.48)	1.24 (-0.72; 3.21)	0.83 (0.72; 0.94)
Trim-and-fill	0.45 (0.31; 0.65)	2.87 (2.38; 3.47)	2.62 (0.99; 4.26)	0.81 (0.71; 0.93)
Weighted median	0.57 (0.43; 0.75)	2.99 (2.41; 3.72)	0.62 (0.52; 0.72)	0.83 (0.75; 0.91)
Weighted mode	0.57 (0.44; 0.75)	2.55 (1.82; 3.56)	0.61 (0.51; 0.70)	0.83 (0.75; 0.90)

[A]Odds ratio. [B]Mean difference. [C]Risk ratio. [D]Removing outlying studies substantially attenuates both γ and I^2 (see section 4).

I^2: between-study inconsistency. γ: Egger test's coefficient (i.e., slope in inverse variance weighted linear regression of effect estimates on standard errors).

Table 2. Brief description of the simulation scenarios.

Scenario	β[A]	Bias type	Aim
1	0	None	Assess bias and FRR[B] under neither treatment effect nor small study effects, for various numbers of studies included in the meta-analysis.
2	0	Type (a)	Asses bias and FRR[B] in the presence of type (a) bias, for various numbers of studies included in the meta-analysis.
3	0	Type (b)	Asses bias and FRR[B] in the presence of type (b) bias, for various numbers of studies included in the meta-analysis. In scenario 3, the minimum P-value required for publication is a linear function of study size.
4	0	Type (b)	Same as scenario 3, except that the minimum P-value required for publication is a square root function of study size.
5	0	Type (b)	Same as scenario 3, except that the minimum P-value required for publication is a quadratic function of study size.
6	0	Type (b)	Same as scenario 3, except that the minimum P-value required for publication is a step function of study size.
7	0.02	None	Assess power to detect a non-zero treatment effect in the absence of small study effects, for various numbers of studies included in the meta-analysis.

[A]Treatment effect.
[B]Since $\beta = 0$ in scenarios 1-6, the FRR under small study effects is simply the overall rejection rate.
FRR: false-rejection rate.

Figure legends

Figure 1. Funnel plots of the catheter (panel A), aspirin (panel B), sodium (panel C) and streptokinase (panel D) meta-analyses.

Figure 2. Illustration of the assumptions underlying the weighted mean, weighted median and the mode-based estimate (MBE) estimators. 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. D: 7 out of 10 studies are biased, and biased studies comprise the largest subgroup of studies that reported the same result.

Figure 3. 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.

Figure 4. Bias (solid lines) and coverage (dashed lines) of the weighted mean (black), regression-based extrapolation (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. The dashed grey line indicates 95% coverage.

Figure 5. Bias (solid lines) and coverage (dashed lines) of the weighted mean (black), regression-based extrapolation (red), trim-and-fill (green), weighted median (dark blue) and mode-based estimate (light blue) under scenario 3: zero true effect (i.e., $\beta = 0$), small study effects through publication bias (assuming a

linear relationship between p_j and 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. The dashed grey line indicates 95% coverage.

Box 1. General principles of our small study effects models

We use model (1) to explore two types of small study effects: 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 differences between small and large published studies are due to fundamental properties of each study that are correlated with study size (n_j). For this, the fixed bias parameter b_j is a function of n_j such that:

$0 \leq b_j \leq b_k$ or $0 \geq b_j \geq b_k$, whenever $n_k \leq n_j$.

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:

$b_j \to 0$ as $n_j \to \infty$, 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., publication bias). We induce this through the random error component of model (1), ε_j, by defining l_j or u_j as functions of n_j such that, whenever $n_k \leq n_j$:

$l_j \leq l_k \leq 0$, and therefore $0 \leq E[\varepsilon_j|n_j] \leq E[\varepsilon_k|n_k]$; or

$u_j \geq u_k \geq 0$, and therefore $0 \geq E[\varepsilon_j|n_j] \geq E[\varepsilon_k|n_k]$.

For example, assume that type (b) bias is always positive, so that and $E[\varepsilon_j|n_j] \geq 0$. This corresponds to a situation where the selection process favours the publication of studies that reported positive effect estimates. This could be achieved defining l_j as a non-increasing function of n_j.

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

$l_j \to -\infty$ and $u_j \to \infty \Rightarrow E[\varepsilon_j|n_j] \to 0$ as $n_j \to \infty$; or

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

Box 1. General principles of our small study bias models (continued)

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, $\mathrm{Var}\left[\varepsilon_j | n_j\right]$ will generally be less than 1, and $\mathrm{Var}\left[\varepsilon_j | n_j\right] \geq \mathrm{Var}[\varepsilon_k | n_k]$ whenever $n_k \leq n_j$.This phenomenon leads to under-dispersion across the set of study-specific estimates constituting the meta-analysis.