PLOS ONE

RESEARCH ARTICLE

Bias caused by sampling error in meta-analysis with small sample sizes

Lifeng Lin*

Department of Statistics, Florida State University, Tallahassee, United States of America

* linl@stat.fsu.edu

OPEN ACCESS

Citation: Lin L (2018) Bias caused by sampling error in meta-analysis with small sample sizes. PLoS ONE 13(9): e0204056. https://doi.org/10.1371/journal.pone.0204056

Editor: Zhongxue Chen, Indiana University Bloomington, UNITED STATES

Received: May 14, 2018

Accepted: August 31, 2018

Published: September 13, 2018

Data Availability Statement: The authors confirm that all data underlying the findings are fully available without restriction. All relevant data are within the Supporting Information files. The Supplementary Information contains the R code for the simulation studies.

Funding: This work was supported in part by the Agency for Healthcare Research and Quality (grant number R03 HS024743, https://www.ahrq.gov/). No additional funding was acquired for this work."

Competing interests: The authors have declared that no competing interests exist.

Abstract

Background

Meta-analyses frequently include studies with small sample sizes. Researchers usually fail to account for sampling error in the reported within-study variances; they model the observed study-specific effect sizes with the within-study variances and treat these sample variances as if they were the true variances. However, this sampling error may be influential when sample sizes are small. This article illustrates that the sampling error may lead to substantial bias in meta-analysis results.

Methods

We conducted extensive simulation studies to assess the bias caused by sampling error. Meta-analyses with continuous and binary outcomes were simulated with various ranges of sample size and extents of heterogeneity. We evaluated the bias and the confidence interval coverage for five commonly-used effect sizes (i.e., the mean difference, standardized mean difference, odds ratio, risk ratio, and risk difference).

Results

Sampling error did not cause noticeable bias when the effect size was the mean difference, but the standardized mean difference, odds ratio, risk ratio, and risk difference suffered from this bias to different extents. The bias in the estimated overall odds ratio and risk ratio was noticeable even when each individual study had more than 50 samples under some settings. Also, Hedges' *g*, which is a bias-corrected estimate of the standardized mean difference within studies, might lead to larger bias than Cohen's *d* in meta-analysis results.

Conclusions

Cautions are needed to perform meta-analyses with small sample sizes. The reported within-study variances may not be simply treated as the true variances, and their sampling error should be fully considered in such meta-analyses.

Introduction

Systematic reviews and meta-analyses have become important tools to synthesize results from various studies in a wide range of areas, especially in clinical and epidemiological research [1–3]. Sampling error is a critical issue in meta-analyses. On the one hand, it impacts the evaluation of heterogeneity between studies. For example, the popular heterogeneity measure I^2 statistic is supposed to quantify the proportion of variation due to heterogeneity rather than sampling error [4–6]; if sampling error increases, the I^2 tends to decrease, leading to a conclusion of more homogeneous studies. More troublesome, within-study sampling error may affect the derivation underlying I^2 to such an extent that the interpretation of I^2 is challenged [7]. On the other hand, sampling error may threaten the validity of a meta-analysis. The most popular meta-analysis method usually models the observed effect size in each study as a normally distributed random variable and treats the observed sample variance as if it was the true variance [8, 9]. It accounts for sampling error in the point estimate of the treatment effect within each study, but it ignores sampling error in the observed variance. This method is generally valid when the number of samples within each collected study is large: the large-sample statistical properties, such as the central limit theory and the delta method, guarantee that the distribution approximation performs well. However, ignoring sampling error in within-study variances has caused some misunderstandings about basic quantities in meta-analyses, especially when some studies have few samples. For example, the famous Q test for homogeneity does not exactly follow the chi-squared distribution due to such sampling error [10], and this problem may subvert I^2 [7].

One important purpose of performing meta-analyses is to increase precision as well as to reduce bias for the conclusions of systematic reviews [11]. For this reason, the PRISMA statement [12] recommends researchers to report both the risks of bias within individual studies and also between studies. The bias within individual studies often relates to the studies' quality [13]. Also, certain measures have been designed to reduce bias in study-level estimates. For example, Hedges' g is considered less biased than Cohen's d within studies when the effect size is the standardized mean difference (page 81 in Hedges and Olkin [14]). The bias between studies is usually introduced by publication bias or selective reporting [15–20]. Besides the bias in point estimates of treatment effects, sampling error also produces bias in the variance of the overall weighted mean estimate under the fixed-effect setting [21, 22]. Under the random-effects setting, the well-known DerSimonian–Laird estimator of the between-study variance may also have considerable bias, especially when sample sizes are small [23, 24]. Moreover, the between-study bias in the treatment effect estimates, such as publication bias, may implicate other parameters in a meta-analysis, including the between-study variance [25]. The bias in variance estimates can seriously impact the precision of the meta-analysis results.

This article focuses on the performance of meta-analyses with small sample sizes, where the sampling error in the observed within-study variances may not be ignored. Throughout this article, we refer to sample size as the number of participants in an individual study, instead of the number of studies in a meta-analysis. Studies with small or moderate sample sizes are fairly common in meta-analyses [26], especially when the treatments are expensive and the enrollments of participants are limited by studies' budgets. We demonstrate a type of bias in meta-analysis results that is completely due to sampling error; it has received relatively less attention in the existing literature compared with other types of bias [27–30]. Such bias is mainly caused by the association between the observed study-specific effect sizes y_i and their within-study variances s_i^2. This association may exist even in the absence of publication bias or selective reporting [31, 32]. When one uses the true variances instead of

the estimated variances, the association may still be present for certain effect sizes, e.g., the (log) odds ratio.

If each study's result is unbiased and its marginal expectation equals to some overall treatment effect θ, then a naïve argument for the unbiasedness of the overall effect estimate in a meta-analysis, $\hat{\theta} = \frac{\sum w_i y_i}{\sum w_i}$, is that $\mathrm{E}\left[\hat{\theta}\right] = \frac{\sum w_i \mathrm{E}[y_i]}{\sum w_i} = \theta$, where w_i is the weight of each study. The weight is usually the inverse of the within-study variance or the marginal variance incorporating heterogeneity between studies. However, this equation treats the weights w_i as fixed values, while in practice they are estimates subject to sampling error. The association between the observed effect sizes and their estimated within-study variances may be strong when the sample sizes are small, so the expectation of the overall estimate in the meta-analysis may not be directly derived without the information about such association, and its unbiasedness is largely unclear [33]. In addition, when the sample sizes are small, the sampling error in the observed within-study variances and the estimated between-study variance may be large, so the confidence interval (CI) of the overall estimate may be poor with coverage probability much lower than the nominal level.

In the following sections, we will review five common effect sizes for continuous and binary outcomes, explain how small sample sizes may introduce bias in meta-analyses, and evaluate such bias using extensive simulation studies.

Methods

Meta-analyses with continuous outcomes

Suppose that a meta-analysis contains N studies, and each study compares a treatment group with a control group. Denote n_{i0} and n_{i1} as the sample sizes in the control and treatment groups in study i. The continuous outcome measures of the participants in each group are assumed to follow normal distributions. The population means of the two groups in study i are μ_{i0} and μ_{i1}, and the sample means are denoted as $\bar{y}_{i0}$ and $\bar{y}_{i1}$ accordingly. The variances of the samples in the two groups are frequently assumed to be equal, denoted as σ_i^2; see, e.g., page 76 in Hedges and Olkin [14] and page 224 in Cooper et al. [34]. The σ_i^2 is estimated as the pooled sample variance $s_{i\mathrm{P}}^2 = \frac{(n_{i0}-1)S_{i0}^2+(n_{i1}-1)S_{i1}^2}{n_{i0}+n_{i1}-2}$, where s_{i0}^2 and s_{i1}^2 are the sample variances in the control and treatment groups, respectively.

If the outcome measures have a meaningful scale and all studies in the meta-analysis are reported on the same scale, the mean difference (MD) between the two groups, i.e., $\Delta_i = \mu_{i1} - \mu_{i0}$, is often used as the effect size to measure treatment effect (page 224 in Cooper et al. [34]). We can obtain an estimate of the MD from each study, denoted as $y_i = \bar{y}_{i1} - \bar{y}_{i0}$, and its estimated within-study variance is $s_i^2 = \left(\frac{1}{n_{i0}} + \frac{1}{n_{i1}}\right) s_{i\mathrm{P}}^2$. Traditional meta-analysis methods usually account for sampling error in the sample means y_i but ignore such error in the sample variances s_i^2; the within-study variances have been customarily treated as the true variances, which should be $\left(\frac{1}{n_{i0}} + \frac{1}{n_{i1}}\right)\sigma_i^2$ [10]. However, accurate estimates of variances may require very large sample sizes; the sample variances s_i^2 may be far away from their true values when sample sizes are small. In the following context, we will treat the sample variances as random variables like the sample means, instead of the true variances.

Because the outcome measures are assumed to be normal, the sample means $\bar{y}_{i0}$ and $\bar{y}_{i1}$ are independent of the sample variances s_{i0}^2 and s_{i1}^2 (see page 218 in Casella and Berger [35]). Thus, the y_i and s_i^2 are independent in each study. Given that the observed MDs y_i are unbiased, such independence guarantees that the overall effect size estimate is unbiased in a fixed-effect meta-

analysis (which assumes that the underlying true effect sizes Δ_i in all studies equal to a common value Δ), because $\mathrm{E}\left[\frac{\sum y_i/s_i^2}{\sum 1/s_j^2}\right] = \sum \mathrm{E}[y_i]\mathrm{E}\left[\frac{1/s_i^2}{\sum 1/s_j^2}\right] = \Delta \sum \mathrm{E}\left[\frac{1/s_i^2}{\sum 1/s_j^2}\right] = \Delta$. However, in a random-effects meta-analysis, each study's weight is updated as $1/(s_i^2 + \hat{\tau}^2)$ by incorporating an estimate of the between-study variance $\hat{\tau}^2$. The between-study variance τ^2 can be estimated using many different methods [36], and its estimate depends on both y_i and s_i^2; therefore, y_i and the updated weight $1/(s_i^2 + \hat{\tau}^2)$ may be correlated to some extents. The expectation of the weighted average cannot be split in the foregoing way, so the unbiasedness of the overall MD estimate is not guaranteed in a random-effects meta-analysis.

A more commonly-used effect size for continuous outcomes is the standardized mean difference (SMD), because this unit-free measure permits different scales in the collected studies and is deemed more comparable across studies (see Normand [8] and Chapter 3 in Grissom and Kim [37]). The true SMD in study i is $\theta_i = \frac{\mu_{i1} - \mu_{i0}}{\sigma_i}$. It is usually estimated as $y_i = \frac{\bar{y}_{i1} - \bar{y}_{i0}}{s_{i\mathrm{P}}}$ by plugging in the sample means and the pooled variance, and is often referred to as Cohen's d (see page 66 in Cohen [38]). If we define a constant $q_i = n_{i0}n_{i1}/(n_{i0} + n_{i1})$, multiply Cohen's d by $\sqrt{q_i}$, and express it as $\sqrt{q_i}y_i = \frac{\sqrt{q_i}(\bar{y}_{i1} - \bar{y}_{i0})/\sigma_i}{s_{i\mathrm{P}}/\sigma_i}$, then the numerator follows a normal distribution with variance 1, and the denominator is the square root of a chi-squared random variable $(n_{i0} + n_{i1} - 2)s_{ip}^2/\sigma_i^2$ divided by its degrees of freedom $n_{i0} + n_{i1} - 2$ [35]. Also, the numerator and denominator are independent. Therefore, strictly speaking, Cohen's d (multiplied by the constant $\sqrt{q_i}$) follows a t-distribution, although it is approximated as a normal distribution in nearly all applications. If the true effect size is non-zero, the t-distribution is noncentral. The exact within-study variance of Cohen's d can be derived as a complicated form of gamma functions [39], but researchers usually use some simpler forms to approximate it. Different approximation forms for the within-study variance of Cohen's d are given in several books on meta-analyses; see, e.g., page 80 in Hedges and Olkin [14], page 226 in Cooper et al. [34], and page 290 in Egger et al. [40]. This article approximates it as $s_i^2 = \frac{1}{n_{i0}} + \frac{1}{n_{i1}} + \frac{y_i^2}{2(n_{i0}+n_{i1}-2)}$. As s_i^2 depends on y_i, they are correlated. The correlation may increase as the sample size decreases, because the coefficient of y_i^2 in the formula of s_i^2, $\frac{1}{2(n_{i0}+n_{i1}-2)}$, increases.

Furthermore, it is well-known that Cohen's d is a biased estimate of the SMD. The bias is around $\frac{3\theta_i}{4(n_{i0}+n_{i1})-9}$ (page 80 in Hedges and Olkin [14]); and it reduces toward zero as the sample sizes increase. When the sample sizes are small, a bias-corrected estimate, called Hedges' g, is usually adopted [41]. It is calculated as $y_i = \left[1 - \frac{3}{4(n_{i0}+n_{i1})-9}\right] \cdot \frac{\bar{y}_{i1} - \bar{y}_{i0}}{s_{i\mathrm{P}}}$ with an estimated variance $s_i^2 = \frac{1}{n_{i0}} + \frac{1}{n_{i1}} + \frac{y_i^2}{2(n_{i0}+n_{i1})}$ (page 86 in Hedges and Olkin [14]). Like Cohen's d, the observed data y_i and s_i^2 are also correlated when using Hedges' g as the effect size. Therefore, even if Hedges' g is (nearly) unbiased within each individual study, the overall SMD estimate in the meta-analysis may be still biased due to the correlation between y_i and s_i^2.

Meta-analyses with binary outcomes

Suppose a 2 × 2 table is available from each collected study in a meta-analysis with a binary outcome. Denote n_{i00} and n_{i01} as the numbers of participants without and with an event in the control group, respectively; n_{i10} and n_{i11} are the data cells in the treatment group. The sample sizes in the control and treatment groups are $n_{i0} = n_{i00} + n_{i01}$ and $n_{i1} = n_{i10} + n_{i11}$. Also, denote p_{i0} and p_{i1} as the population event rates in the two groups.

The odds ratio (OR) is frequently used to measure treatment effect for a binary outcome [42]; its true value in study i is $\mathrm{OR}_i = \frac{p_{i1}/(1-p_{i1})}{p_{i0}/(1-p_{i0})}$. Using the four data cells in the 2×2 table, the

OR is estimated as $\widehat{\mathrm{OR}}_i = \frac{n_{i00}n_{i11}}{n_{i01}n_{i10}}$. The ORs are usually combined on a logarithmic scale in meta-analyses, because the distribution of the estimated log OR, $y_i = \log \widehat{\mathrm{OR}}_i$ is better approximated by a normal distribution. The within-study variance of y_i is estimated as $s_i^2 = \frac{1}{n_{i00}} + \frac{1}{n_{i01}} + \frac{1}{n_{i10}} + \frac{1}{n_{i11}}$. Besides the OR, the risk ratio (RR) and risk difference (RD) are also popular effect sizes. The underlying true RR in study i is $\mathrm{RR}_i = p_{i1}/p_{i0}$, and it is also combined on the log scale in meta-analyses like the OR. The log RR is estimated as $y_i = \log\frac{n_{i11}/n_{i1}}{n_{i01}/n_{i0}}$, and its within-study variance is estimated as $s_i^2 = \frac{1}{n_{i01}} + \frac{1}{n_{i11}} - \frac{1}{n_{i0}} - \frac{1}{n_{i1}}$. Moreover, the underlying true RD in study i is $\mathrm{RD}_i = p_{i1} - p_{i0}$, estimated as $y_i = \frac{n_{i11}}{n_{i1}} - \frac{n_{i01}}{n_{i0}}$ with an estimated within-study variance $s_i^2 = \frac{n_{i00}n_{i01}}{n_{i0}^3} + \frac{n_{i10}n_{i11}}{n_{i1}^3}$. When the sample sizes are small, some data cells may be zero even if the event is not rare. If a 2 × 2 table contains zero cells, a fixed value of 0.5 is often added to each data cell to reduce bias and avoid computational error (see page 521 in the *Cochrane Handbook for Systematic Reviews of Interventions* [43] and many other papers [44–46]), although this continuity correction may not be optimal in some cases [47–50].

Like the SMD for continuous outcomes, the distributions of the sample log OR, log RR, and RD are approximated as normal distributions in conventional meta-analysis methods. Also, because both y_i and s_i^2 depend on the four cells of 2 × 2 tables for all three effect sizes, they are intrinsically correlated.

Simulation studies

We conducted simulation studies to investigate the impact of sampling error on meta-analyses with small sample sizes. The number of studies in a simulated meta-analysis was set to N = 5, 10, 20, and 50. We first generated the sample size within each study n_i from a uniform distribution $U(5, 10)$, then we gradually increased it by sampling it from $U(10, 20)$, $U(20, 30)$, $U(30, 50)$, $U(50, 100)$, $U(100, 500)$, and $U(500, 1000)$. These sample sizes n_i were generated anew for each simulated meta-analysis. The control/treatment allocation ratio was set to 1:1 in all studies, which is common in real-world applications. Specifically, $n_{i0} = \lceil \frac{n_i}{2} \rceil$ participants were assigned to the control group and $n_{i1} = n_i - \lceil \frac{n_i}{2} \rceil$ participants were assigned to the treatment group, where $\lceil x \rceil$ represents an integer that is greater than or equal to x.

When the outcome was continuous, we simulated meta-analyses based on the MD and the SMD. For the MD, each participant's outcome measure was sampled from $N(\mu_{i0}, \sigma_i^2)$ in the control group or $N(\mu_{i0} + \Delta_i, \sigma_i^2)$ in the treatment group. Without loss of generality, the baseline effect μ_{i0} of study i was generated from $N(0,1)$. The study-specific standard deviation σ_i was sampled from $U(1,5)$, and it was generated anew for each simulated meta-analysis. The mean difference Δ_i was sampled from $N(\Delta,\tau^2)$. The overall MD Δ was set to 0, 0.5, 1, 2, and 5, and the between-study standard deviation τ was set to 0, 0.5, and 1. For the SMD, each participant's outcome measure was also generated using the foregoing setting within each study, but the SMD $\theta_i = \Delta_i/\sigma_i$, not the mean difference Δ_i, was sampled from the normal distribution: $\theta_i \sim N(\theta,\tau^2)$. The overall SMD θ was set to 0, 0.2, 0.5, 0.8, and 1 to represent different magnitudes of effect size. The between-study standard deviation τ was 0, 0.2, and 0.5. Both Cohen's d and Hedges' g were used to estimate the SMD.

When the outcome was binary, we first simulated meta-analyses based on the OR. The event numbers n_{i01} and n_{i11} in the control and treatment groups were sampled from Binomial (n_{i0},p_{i0}) and Binomial(n_{i1},p_{i1}), respectively. The event rate in the control group p_{i0} was sampled from $U(0.3, 0.7)$ representing a fairly common event [32], and it was generated anew for each meta-analysis. The event rate in the treatment group p_{i1} was calculated using p_{i0} and the study-specific log OR θ_i; specifically, $p_{i1} = [1 + e^{-\theta_i}\ (1 - p_{i0})/p_{i0}]^{-1}$. The study-specific log OR θ_i was

sampled from $N(\theta,\tau^2)$, where the overall log OR θ was set to 0, 0.2, 0.4, 1, and 1.5, and the between-study standard deviation τ was 0, 0.2, and 0.5. In addition to the OR, we also generated meta-analyses based on the RR and RD. The event numbers were similarly sampled from binomial distributions and the p_{i0} was from $U(0.3,0.7)$. However, for the log RR and the RD, we considered only the fixed-effect setting with all study-specific effect sizes θ_i equal to a common value θ. Specifically, if the effect size was the log RR, the event rate in the treatment group was $p_{i1} = e^{\theta}p_{i0}$, where the true log RR θ was set to 0 and 0.3 to guarantee that p_{i1} was between 0 and 1. If the effect size was the RD, $p_{i1} = p_{i0} + \theta$, where the true RD θ was set to 0 and 0.2 to guarantee that p_{i1} was between 0 and 1. The random-effects setting was not considered for the log RR and RD because it may lead to improper p_{i1}'s beyond the [0, 1] range. We could successfully generate meta-analyses by truncating such improper p_{i1}'s and constraining them to be between 0 and 1; however, this constraint would produce bias, which cannot be distinguished from the bias caused by sampling error that is of primary interest in this article.

For each simulation setting above, 10,000 meta-analyses were generated. The random-effects model was applied to each simulated meta-analysis [51], even if some meta-analyses were generated under the fixed-effect setting with $\tau = 0$. Thus, the produced CIs might be conservative. Also, the between-study variance was estimated using the popular method of moments by DerSimonian and Laird [24]. The restricted maximum likelihood method may be a better choice [8, 52, 53], but it is more computationally difficult and its solution did not converge in a noticeable number of our simulated meta-analyses. Also, there are many other alternatives for estimating the between-study variance, such as the Paule–Mandel estimator, which may be recommended in certain situations [54, 55], while they have been used less frequently compared with the DerSimonian–Laird estimator so far. Therefore, we considered only the DerSimonian–Laird estimator for the between-study variance, which was sufficient to achieve this article's purpose.

S2–S7 Files present the R code and results for the simulation studies.

Results

Fig 1–5 present the boxplots of the estimated overall effect sizes in the 10,000 simulated meta-analyses for the MD, SMD (estimated by both Cohen's *d* and Hedges' *g*), log OR, log RR, and

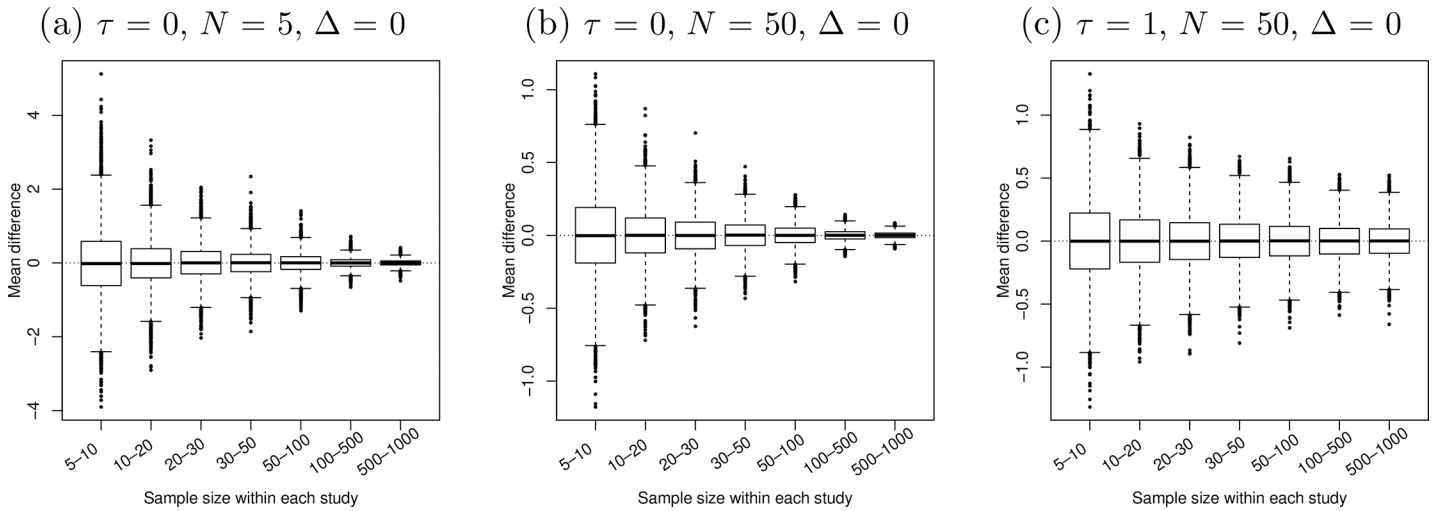

Fig 1. Boxplots of the estimated mean differences in 10,000 simulated meta-analyses. The true between-study standard deviation τ increased from 0 (panels a and b) to 1 (panel c). The number of studies in each meta-analysis N increased from 5 (panel a) to 50 (panels b and c). The true mean difference Δ (horizontal dotted line) was 0.

https://doi.org/10.1371/journal.pone.0204056.g001

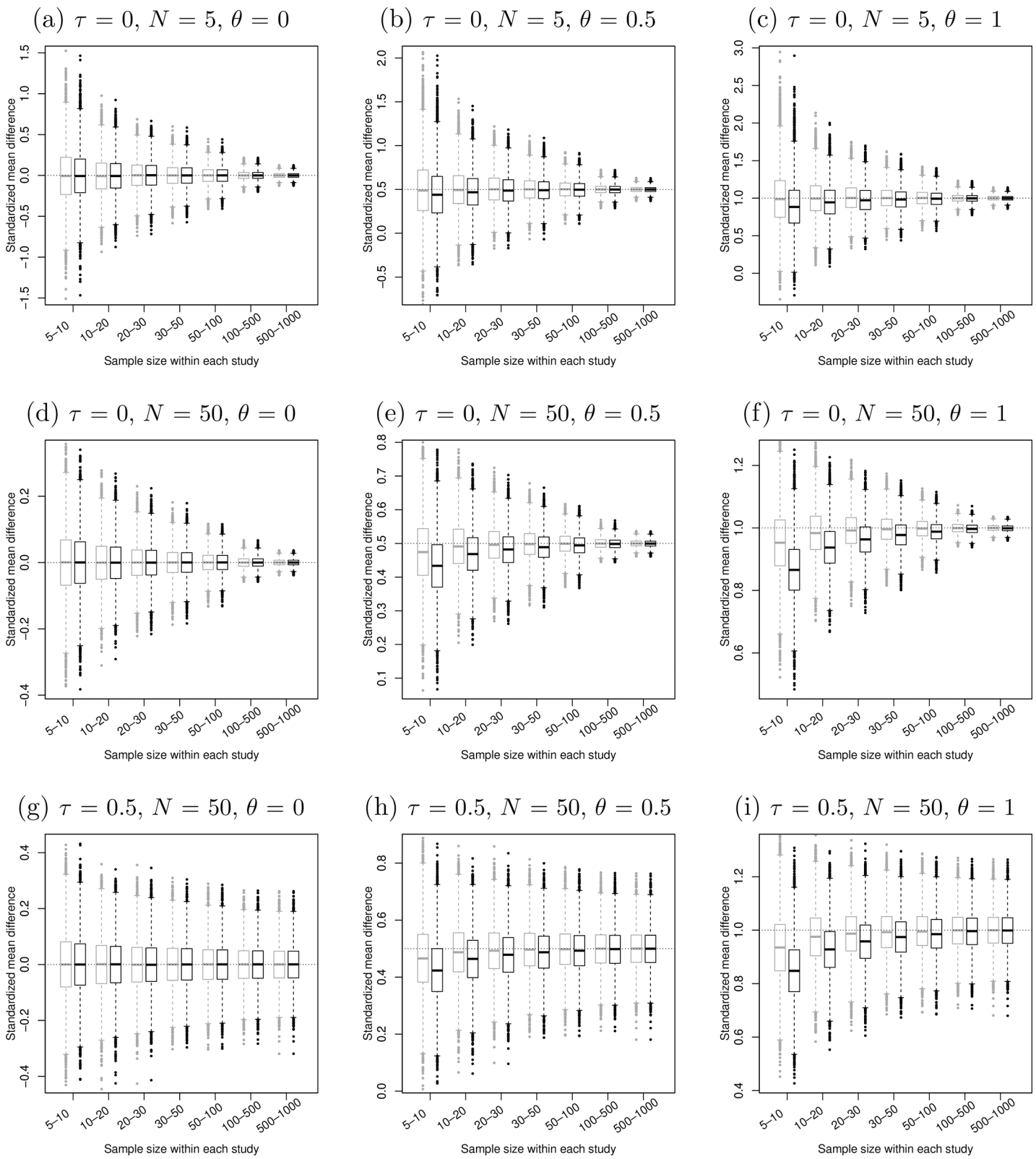

Fig 2. Boxplots of the estimated standardized mean differences in 10,000 simulated meta-analyses. For each sample size range on the horizontal axis, the left gray box was obtained using Cohen's *d*, and the right black box was obtained using Hedges' *g*. The true between-study standard deviation τ increased from 0 (upper and middle panels) to 0.5 (lower panels). The number of studies in each meta-analysis N increased from 5 (upper panels) to 50 (middle and lower panels). The true standardized mean difference θ (horizontal dotted line) increased from 0 (left panels) to 1 (right panels).

https://doi.org/10.1371/journal.pone.0204056.g002

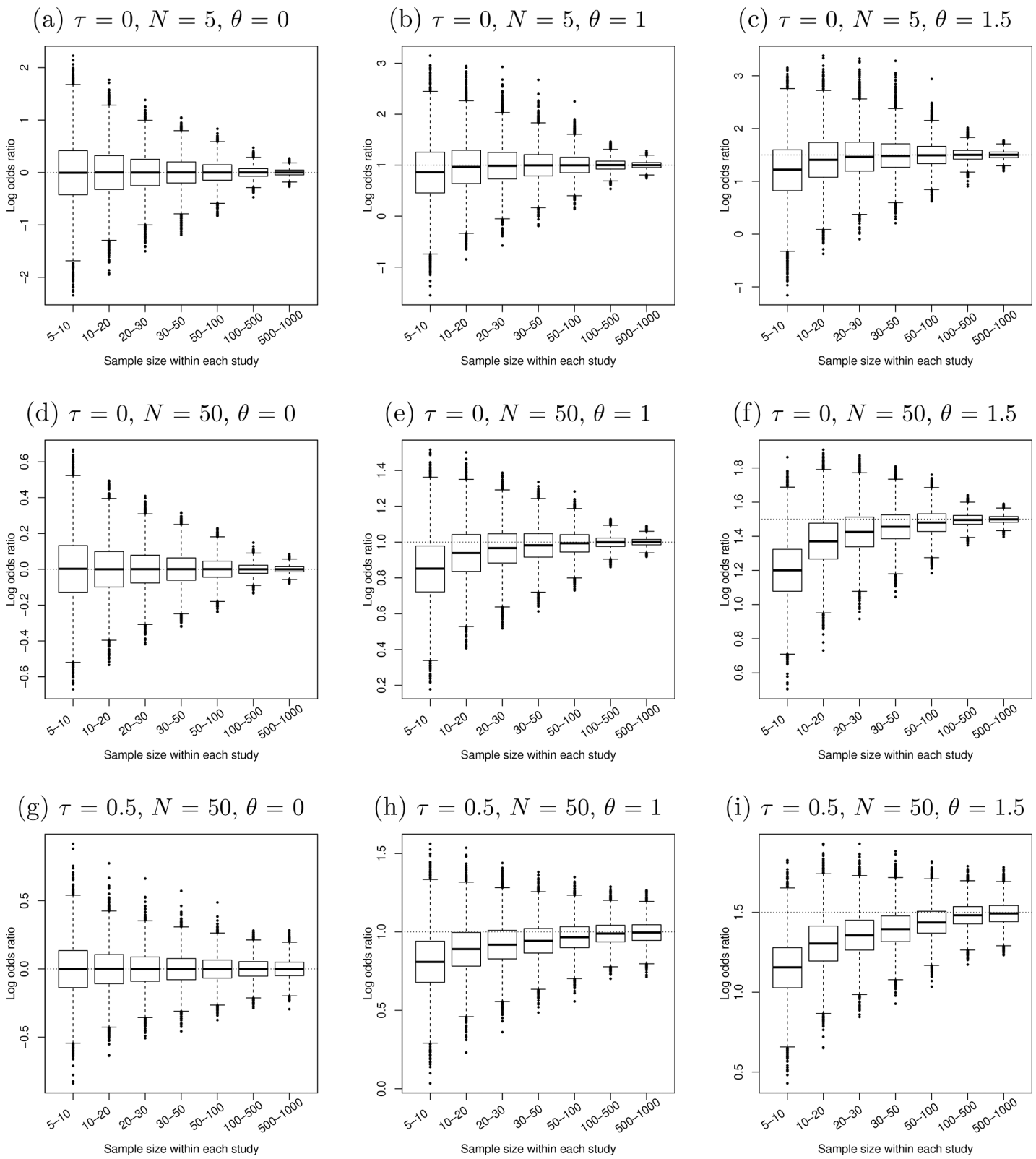

Fig 3. Boxplots of the estimated log odds ratios in 10,000 simulated meta-analyses. The true between-study standard deviation τ increased from 0 (upper and middle panels) to 0.5 (lower panels). The number of studies in each meta-analysis N increased from 5 (upper panels) to 50 (middle and lower panels). The true log odds ratio θ (horizontal dotted line) increased from 0 (left panels) to 1.5 (right panels).

https://doi.org/10.1371/journal.pone.0204056.g003

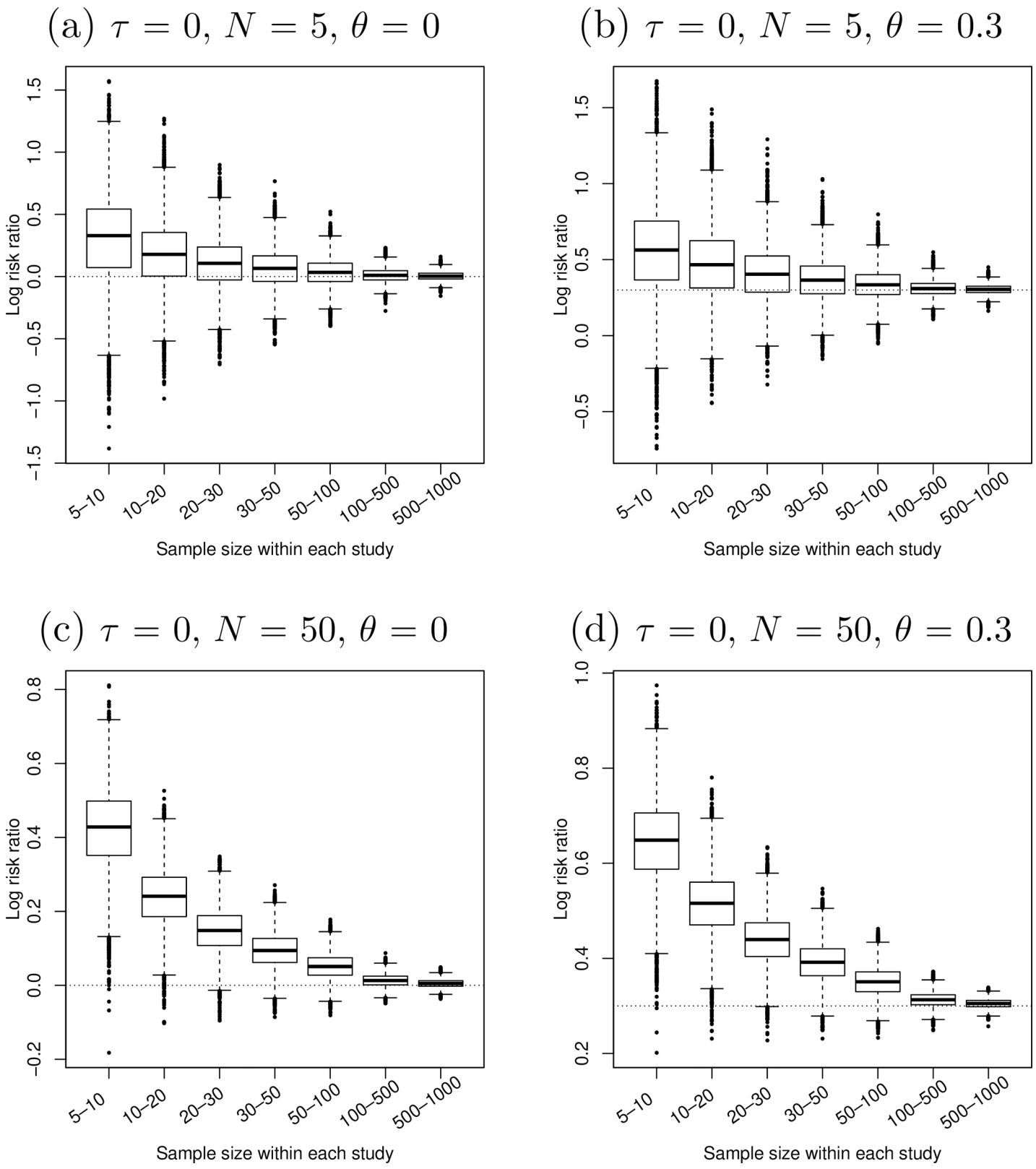

Fig 4. Boxplots of the estimated log risk ratios in 10,000 simulated meta-analyses. The true between-study standard deviation τ was 0 (i.e., the simulated studies were homogeneous). The number of studies in each meta-analysis N increased from 5 (upper panels) to 50 (lower panels). The true log risk ratio θ (horizontal dotted line) increased from 0 (left panels) to 0.3 (right panels).

https://doi.org/10.1371/journal.pone.0204056.g004

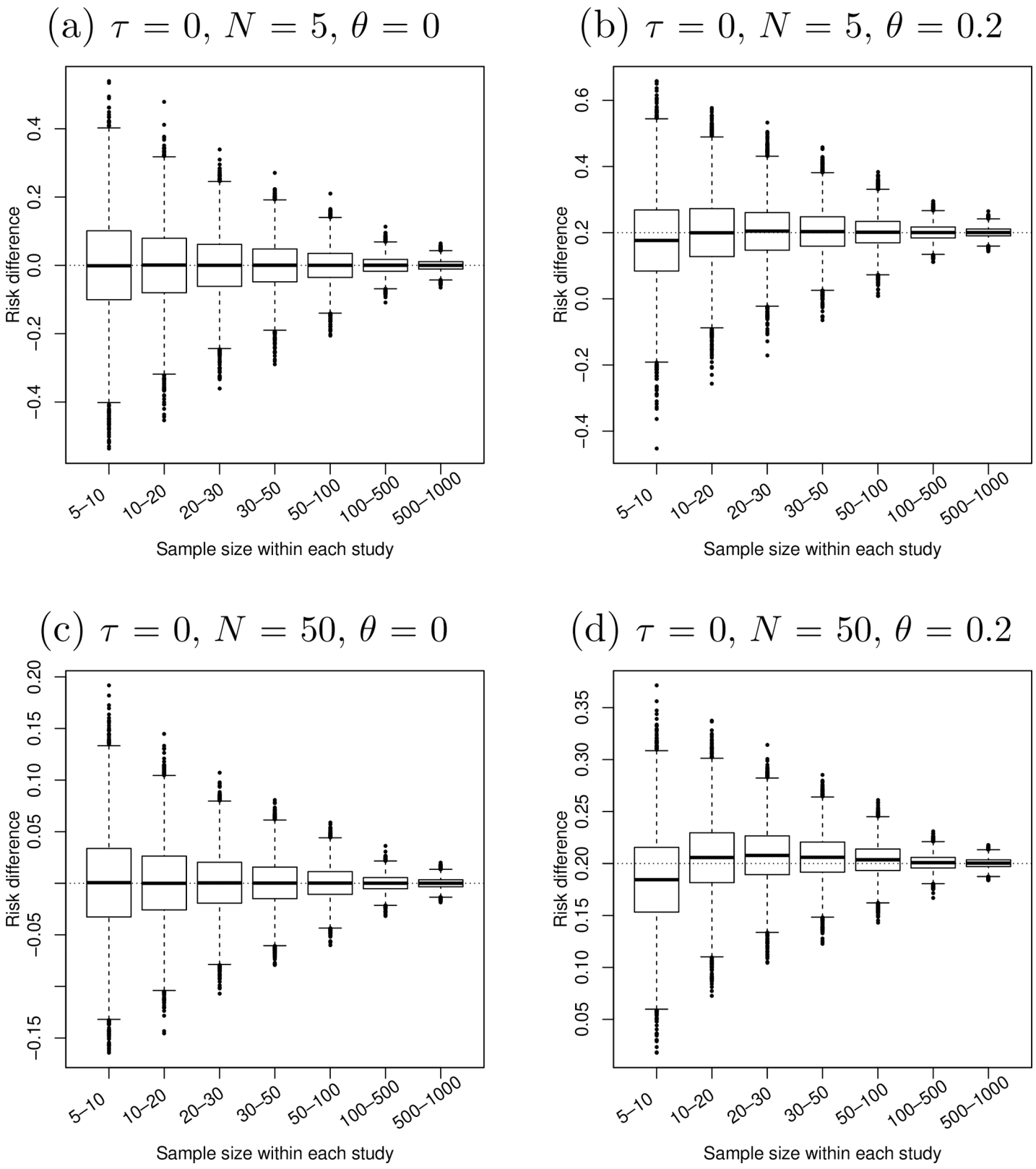

Fig 5. Boxplots of the estimated risk differences in 10,000 simulated meta-analyses. The true between-study standard deviation τ was 0 (i.e., the simulated studies were homogeneous). The number of studies in each meta-analysis N increased from 5 (upper panels) to 50 (lower panels). The true risk difference θ (horizontal dotted line) increased from 0 (left panels) to 0.2 (right panels).

https://doi.org/10.1371/journal.pone.0204056.g005

Table 1. Bias of the estimated overall effect size in the simulation studies.

Setting	Sample size						
	5–10	10–20	20–30	30–50	50–100	100–500	500–1000
Mean difference:							
$\tau = 0, N = 5, \Delta = 0$	−0.01	−0.01	0.01	0.00	0.00	0.00	0.00
$\tau = 0, N = 50, \Delta = 0$	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\tau = 1, N = 50, \Delta = 0$	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Standardized mean difference (Cohen's *d*):							
$\tau = 0, N = 5, \theta = 0$	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\tau = 0, N = 5, \theta = 0.5$	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\tau = 0, N = 5, \theta = 1$	0.01	0.00	0.01	0.00	0.00	0.00	0.00
$\tau = 0, N = 50, \theta = 0$	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\tau = 0, N = 50, \theta = 0.5$	−0.02	−0.01	0.00	0.00	0.00	0.00	0.00
$\tau = 0, N = 50, \theta = 1$	−0.05	−0.02	−0.01	0.00	0.00	0.00	0.00
$\tau = 0.5, N = 50, \theta = 0$	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\tau = 0.5, N = 50, \theta = 0.5$	−0.03	−0.01	−0.01	0.00	0.00	0.00	0.00
$\tau = 0.5, N = 50, \theta = 1$	−0.06	−0.02	−0.01	−0.01	0.00	0.00	0.00
Standardized mean difference (Hedges' *g*):							
$\tau = 0, N = 5, \theta = 0$	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\tau = 0, N = 5, \theta = 0.5$	−0.05	−0.03	−0.01	−0.01	0.00	0.00	0.00
$\tau = 0, N = 5, \theta = 1$	−0.10	−0.05	−0.02	−0.02	−0.01	0.00	0.00
$\tau = 0, N = 50, \theta = 0$	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\tau = 0, N = 50, \theta = 0.5$	−0.07	−0.03	−0.02	−0.01	−0.01	0.00	0.00
$\tau = 0, N = 50, \theta = 1$	−0.13	−0.06	−0.04	−0.02	−0.01	0.00	0.00
$\tau = 0.5, N = 50, \theta = 0$	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\tau = 0.5, N = 50, \theta = 0.5$	−0.08	−0.04	−0.02	−0.01	−0.01	0.00	0.00
$\tau = 0.5, N = 50, \theta = 1$	−0.15	−0.07	−0.04	−0.03	−0.01	0.00	0.00
Log odds ratio:							
$\tau = 0, N = 5, \theta = 0$	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\tau = 0, N = 5, \theta = 1$	−0.15	−0.03	0.00	0.00	0.00	0.00	0.00
$\tau = 0, N = 5, \theta = 1.5$	−0.30	−0.09	−0.02	−0.01	0.00	0.00	0.00
$\tau = 0, N = 50, \theta = 0$	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\tau = 0, N = 50, \theta = 1$	−0.15	−0.06	−0.04	−0.02	−0.01	0.00	0.00
$\tau = 0, N = 50, \theta = 1.5$	−0.30	−0.13	−0.07	−0.04	−0.02	0.00	0.00
$\tau = 0.5, N = 50, \theta = 0$	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\tau = 0.5, N = 50, \theta = 1$	−0.19	−0.11	−0.08	−0.06	−0.03	−0.01	0.00
$\tau = 0.5, N = 50, \theta = 1.5$	−0.35	−0.19	−0.14	−0.10	−0.06	−0.02	−0.01
Log risk ratio:							
$\tau = 0, N = 5, \theta = 0$	0.30	0.18	0.11	0.07	0.03	0.01	0.00
$\tau = 0, N = 5, \theta = 0.3$	0.26	0.17	0.11	0.07	0.04	0.01	0.00
$\tau = 0, N = 50, \theta = 0$	0.42	0.24	0.15	0.09	0.05	0.01	0.01
$\tau = 0, N = 50, \theta = 0.3$	0.35	0.22	0.14	0.09	0.05	0.01	0.01
Risk difference:							
$\tau = 0, N = 5, \theta = 0$	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\tau = 0, N = 5, \theta = 0.2$	−0.02	0.00	0.00	0.00	0.00	0.00	0.00
$\tau = 0, N = 50, \theta = 0$	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\tau = 0, N = 50, \theta = 0.2$	−0.02	0.01	0.01	0.01	0.00	0.00	0.00

https://doi.org/10.1371/journal.pone.0204056.t001

Table 2. Coverage probability (in percentage, %) of the estimated overall effect size's 95% confidence interval in the simulation studies.

Setting	Sample size						
	5–10	10–20	20–30	30–50	50–100	100–500	500–1000
Mean difference:							
$\tau = 0, N = 5, \Delta = 0$	92.0	94.7	95.4	96.0	95.9	96.1	96.6
$\tau = 0, N = 50, \Delta = 0$	92.9	94.3	95.0	95.2	95.8	96.2	96.0
$\tau = 1, N = 50, \Delta = 0$	93.8	93.6	93.8	94.5	94.2	94.1	94.2
Standardized mean difference (Cohen's *d*):							
$\tau = 0, N = 5, \theta = 0$	97.1	96.4	96.4	96.6	96.2	96.2	96.2
$\tau = 0, N = 5, \theta = 0.5$	97.2	96.6	96.3	96.4	96.3	96.2	96.3
$\tau = 0, N = 5, \theta = 1$	97.2	96.7	96.4	96.4	96.2	96.1	96.4
$\tau = 0, N = 50, \theta = 0$	97.1	96.2	96.2	95.9	96.1	95.8	95.8
$\tau = 0, N = 50, \theta = 0.5$	96.5	96.2	95.9	95.7	96.1	96.0	95.9
$\tau = 0, N = 50, \theta = 1$	94.9	95.7	95.7	95.6	96.0	95.9	96.0
$\tau = 0.5, N = 50, \theta = 0$	96.0	94.9	94.0	94.8	93.9	94.2	94.0
$\tau = 0.5, N = 50, \theta = 0.5$	95.5	94.6	93.8	94.6	94.2	94.1	94.1
$\tau = 0.5, N = 50, \theta = 1$	93.0	94.0	93.8	94.5	94.3	94.2	94.3
Standardized mean difference (Hedges' *g*):							
$\tau = 0, N = 5, \theta = 0$	98.0	97.0	96.7	96.8	96.4	96.2	96.2
$\tau = 0, N = 5, \theta = 0.5$	97.7	96.8	96.6	96.6	96.4	96.2	96.4
$\tau = 0, N = 5, \theta = 1$	96.9	96.3	96.2	96.4	96.1	96.0	96.3
$\tau = 0, N = 50, \theta = 0$	97.6	96.7	96.5	96.1	96.2	95.8	95.8
$\tau = 0, N = 50, \theta = 0.5$	94.0	94.8	95.1	95.1	95.7	95.9	96.0
$\tau = 0, N = 50, \theta = 1$	81.2	89.0	92.1	93.4	94.9	95.6	95.9
$\tau = 0.5, N = 50, \theta = 0$	96.1	94.9	94.0	94.9	94.0	94.2	94.0
$\tau = 0.5, N = 50, \theta = 0.5$	91.6	93.1	93.3	94.4	94.2	94.1	94.1
$\tau = 0.5, N = 50, \theta = 1$	76.6	88.3	91.2	93.1	93.9	94.2	94.3
Log odds ratio:							
$\tau = 0, N = 5, \theta = 0$	98.2	97.3	97.0	96.9	96.5	96.4	95.9
$\tau = 0, N = 5, \theta = 1$	98.2	97.5	96.9	96.6	96.4	96.1	96.4
$\tau = 0, N = 5, \theta = 1.5$	97.8	97.8	97.2	96.9	96.4	96.5	96.2
$\tau = 0, N = 50, \theta = 0$	98.0	97.1	96.6	96.0	95.7	95.7	95.7
$\tau = 0, N = 50, \theta = 1$	94.9	96.1	95.9	96.0	95.7	95.8	95.7
$\tau = 0, N = 50, \theta = 1.5$	82.5	91.7	93.6	94.6	95.4	96.0	95.6
$\tau = 0.5, N = 50, \theta = 0$	97.7	96.5	95.7	95.2	94.6	94.2	94.3
$\tau = 0.5, N = 50, \theta = 1$	91.7	92.7	92.5	92.9	93.5	93.9	94.3
$\tau = 0.5, N = 50, \theta = 1.5$	75.0	83.5	84.7	86.7	90.4	93.7	94.1
Log risk ratio:							
$\tau = 0, N = 5, \theta = 0$	84.0	90.2	93.1	94.4	95.3	96.2	96.0
$\tau = 0, N = 5, \theta = 0.3$	86.2	89.9	92.0	93.0	94.8	95.8	96.1
$\tau = 0, N = 50, \theta = 0$	7.4	25.6	41.6	56.2	73.0	90.3	93.5
$\tau = 0, N = 50, \theta = 0.3$	9.7	25.1	38.2	49.9	66.5	87.6	92.4
Risk difference:							
$\tau = 0, N = 5, \theta = 0$	94.0	94.2	94.9	95.6	95.9	96.3	95.8
$\tau = 0, N = 5, \theta = 0.2$	94.4	94.5	94.7	95.0	95.6	96.0	96.0
$\tau = 0, N = 50, \theta = 0$	92.7	93.5	93.9	94.5	94.9	95.7	95.6
$\tau = 0, N = 50, \theta = 0.2$	92.4	93.3	93.2	93.9	94.5	95.6	95.7

https://doi.org/10.1371/journal.pone.0204056.t002

RD, respectively. In addition, Table 1 shows the bias of the estimates and Table 2 shows their 95% CIs' coverage probabilities. When the number of studies in a meta-analysis increased from 5 to 50, the range of the estimated overall effect size shrank because their variances decreased. When the between-study heterogeneity increased in Fig 1–3, the middle and lower panels indicate that the box of the estimated overall effect sizes expanded vertically due to more heterogeneity in the meta-analyses.

Fig 1 and Table 1 indicate that the estimated MD was almost unbiased in all situations with different numbers of studies and different extents of heterogeneity, even if the studies had very small sample sizes. As the trends in the plots for $\Delta = 0.5$, 1, 2, and 5 were fairly similar to those for $\Delta = 0$, they were not displayed in Fig 1 due to space limit. Table 2 shows that the CI coverage probability of the MD was fairly close to the nominal confidence level 95% in most cases. The coverage was slightly below 95% when the number of studies was small ($N = 5$) and the sample sizes were also very small (between 5 and 10) within studies.

When the true SMD was zero in the left panels of Fig 2, both Cohen's d and Hedges' g were almost unbiased. The box of Cohen's d was slightly larger vertically than that of Hedges' g when the sample sizes within studies were small, so the point estimates of Hedges' g were more concentrated around the true SMD. The CI coverage was also close to the nominal 95% level. However, as the true SMD increased from 0 to 0.5 and to 1, both Cohen's d and Hedges'g began to have bias, and the bias increased as the sample sizes decreased within studies. Cohen's d generally produced less bias in the estimated overall SMD than Hedges' g, as shown in Table 1. The CI coverage of Cohen's d was still close to 95% when the true SMD increased, but that of Hedges' g dropped below 80% when the sample size was fairly small (between 5 and 10), the true SMD was fairly large ($\theta = 1$), and the number of studies was large ($N = 50$).

The patterns in Fig 3 of the ORs for binary outcomes were similar to those in Fig 2. The estimated overall log ORs were almost unbiased when the true log OR was zero. As the true log OR increased to 1 and to 1.5 and the sample sizes within studies decreased, the bias in the estimated overall log OR tended to be larger in the negative direction. Also, the CI coverage dropped dramatically when the number of studies and the between-study variance were large in Table 2. For example, when $\tau = 0$, $N = 5$, $\theta = 1.5$, and the sample size of each study was between 5 and 10, the bias of the estimated overall log OR was −0.30 and the CI coverage was 97.8%. The log OR underestimated the true value θ. Among the simulated meta-analyses whose CIs did not cover θ, 2.2% had CIs entirely below θ, while only one meta-analysis (0.01%) had a CI entirely above θ. As the number of studies increased to $N = 50$ and other parameters unchanged, the bias was still −0.30, but the CI coverage decreased to 82.5%. The CIs of the meta-analyses not covering θ were all below θ. Therefore, the low CI coverage was likely because the CI became shorter as the number of studies N increased while the bias remained.

Compared with the log OR, the log RR in Fig 4 was more sensitive to the sample sizes within studies. The estimated overall log RR had tiny bias and its CI coverage was close to 95% when the sample sizes within studies were large (more than 500). However, the bias was substantial and the CI coverage was fairly low even when the sample sizes were moderate (between 50 and 100). Like the situation for the log OR, the poor CI coverage for the log RR related to the bias. For example, when $\tau = 0$, $N = 5$, $\theta = 0.3$, and the sample size of each study was between 5 and 10, the bias of the estimated overall log RR was 0.26 and the CI coverage was 86.2%. The log RR overestimated the true value θ. The CIs of the simulated meta-analyses not covering θ were all above θ. When N increased to 50 and other parameters unchanged, the bias was 0.35 and the CI coverage dropped dramatically to 9.7%. The CIs of the simulated meta-analyses not covering θ were also all above θ.

Fig 5 shows that the estimated overall RD was almost unbiased when the true RD was zero and had small bias when the true RD was 0.2. The bias was relatively large when the sample sizes within studies were fairly small. The CI coverages were between 92% and 96% in all situations.

In addition, Figures A–F in S1 File present scatter plots of the sample effect sizes against their precisions (i.e., the inverse of their sample variances) in ten selected simulated meta-analyses with small sample sizes for the MD, SMD (including both Cohen's *d* and Hedges' *g*), log OR, log RR, and RD. They are plotted using the same idea of the funnel plot for assessing publication bias [56], and they roughly illustrate the association between the sample effect sizes y_i and their within-study variances s_i^2. Figure A in S1 File indicates that this association seemed tiny for the MD, which was consistent with our conclusion that the MD y_i and its variance s_i^2 are independent in theory. The other figures show different extents of association for the SMD, log OR, log RR, and RD. For example, the estimated SMDs that were closer to zero tended to have larger precisions (i.e., smaller variances) in Figures B and C in S1 File.

Discussion

This article has shown that the bias in the overall estimates of the SMD, log OR, log RR, and RD may be substantial in meta-analyses with small sample sizes. The estimated overall MD was almost unbiased in nearly all simulation settings, mainly because its point estimate and within-study variance were independent. However, for the other four effect sizes except the MD, the intrinsic association between their point estimates and estimated variances within studies may be strong, so the meta-analysis results were biased in many simulation settings. Therefore, when the collected studies have small sample sizes, researchers need to choose a proper effect size and perform the meta-analysis with great cautions.

Surprisingly, to estimate the overall SMD, using Cohen's *d* led to noticeably less bias than using Hedges' *g* in our simulation studies, although Hedges' *g* was designed as a bias-corrected estimate of the SMD within individual studies. For example, in one of our simulated fixed-effect meta-analyses with 50 studies and 5 to 10 samples in each study (the true SMD was 1), the average of Cohen's *d* in the 50 studies was around 1.29, while the average of Hedges' *g* 1.07 was closer to the true value 1. This was consistent with the fact that Hedges' *g* was generally less biased within individual studies. However, the meta-analytic overall Cohen's *d* was 0.98, which was much closer to 1 compared with the meta-analytic overall Hedges' *g* 0.89, because of the sampling error in these effect sizes' variances that caused the association between the effect sizes and the variances. Note that, instead of advocating that Cohen's *d* is always preferred than Hedges' *g* in meta-analyses, this article only reminds researchers that Cohen's *d* may be less biased in at least some meta-analytic results, and the argument for the use of Hedges' *g* in the presence of small sample sizes needs to be carefully examined.

In addition, there are alternative methods to estimate the within-study variance of Hedges' *g* besides the one used in our article. Specifically, our simulation studies used $s_{gi}^2 = \frac{1}{n_{i0}} + \frac{1}{n_{i1}} + \frac{g_i^2}{2(n_{i0}+n_{i1})}$, where g_i is the point estimate of Hedges' *g* in study *i*; this calculation was introduced on page 86 in Hedges and Olkin [14]. Recall that Hedges' *g* is calculated by multiplying Cohen's *d* by a bias-correction coefficient; that is, $g_i = J_i d_i$, where $J_i = 1 - \frac{3}{4(n_{i0}+n_{i1})-9}$ and d_i is the point estimate of Cohen's in study *i*. Therefore, the variance of Hedges' *g* can be alternatively estimated as $s_{gi}^2 = J_i^2 s_{di}^2$, where s_{di}^2 is the within-study variance of Cohen's *d*; see, e.g., page 226 in Cooper et al. [34]. Using this alternative calculation for the within-study variances of Hedges' *g*, the combined SMD may remain biased. For example, consider a special case that all *N* studies in a meta-analysis have the same sample size *n*, so the bias-correction coefficients in

all studies are equal: $J_i = J$. Using the fixed-effect model, the expectation of the combined Cohen's d is

$$\mu_d = E\left[\frac{\sum d_i/s_{di}^2}{\sum 1/s_{di}^2}\right],$$

and the expectation of the combined Hedges' g is

$$\mu_g = E\left[\frac{\sum g_i/s_{gi}^2}{\sum 1/s_{gi}^2}\right] = E\left[\frac{\sum (Jd_i)/(J^2 s_{di}^2)}{\sum 1/(J^2 s_{di}^2)}\right] = J\mu_d.$$

Because J is a coefficient always less than 1, we have $\mu_g < \mu_d$ if assuming μ_d is positive. If the true overall SMD θ is also positive and the combined Cohen's d underestimates it (as in our simulation studies), then $\mu_g < \mu_d < \theta$, indicating that the combined Hedges' g is more biased. However, if the combined Cohen's d overestimates the overall SMD (i.e., $\mu_d > \theta$), then the combined Hedges' g might be less biased.

This article helps explain the phenomenon of the inflated type I error rates for testing for publication bias. To detect potential publication bias in meta-analyses, it has been popular to check for the association between the study-specific effect sizes and their standard errors using the funnel plot or Egger's regression test [15]. However, it is well known that such association may be intrinsic for binary outcomes even if no publication bias appears, so Egger's test may have an inflated type I error rate [31, 32]. In addition to the intrinsic association for binary outcomes, this article indicates that such a problem also exists when using the SMD for continuous outcomes. Although the meta-analyses with false positive results do not truly have publication bias, they may still suffer from bias due to sampling error.

Moreover, our findings imply that the magnitude of sample size may not be viewed as an absolute concept in meta-analyses; we may not determine whether a sample size is small or large without taking other parameters into account. For example, using the log OR as the effect size, Fig 3(A), 3(D) and 3(G) show that a sample size of 10 to 20 may be large enough to produce desirable meta-analysis results when the true log OR is zero. However, when the heterogeneity, the number of studies, and the true log OR are large, Fig 3(I) shows that a sample size of 50 to 100 may not be adequate.

The bias of the estimated overall log RR was particularly substantial in Fig 4; this may be related to the effect of the weighting bias for binary outcomes [57]. However, unlike the purpose of Tang [57], this article focused on the bias completely due to sample error which exists for both continuous and binary outcomes.

This article performed the simulated meta-analyses using the popular inverse-of-variance method in a frequentist way. Alternatively, several exact models have been proposed for binary outcomes; they do not require the normal approximation to estimate the study-specific effect sizes and their within-study variances [58–62]. The event numbers in the compared groups can be directly modeled as binomial distributions, thus accounting for sampling error in both point estimates of effect sizes and their variances. Similar exact models are also needed for continuous outcomes to avoid treating the within-study variances as if they were the true variances; we leave them as future work.

Supporting information

S1 File. Scatter plots of the sample effect sizes against their precisions (i.e., the inverse of their sample variances) in some simulated meta-analyses with small sample sizes.
(PDF)

S2 File. R code for the simulation studies.
(ZIP)

S3 File. Simulation results for the mean difference.
(ZIP)

S4 File. Simulation results for the standardized mean difference.
(ZIP)

S5 File. Simulation results for the log odds ratio.
(ZIP)

S6 File. Simulation results for the log risk ratio.
(ZIP)

S7 File. Simulation results for the risk difference.
(ZIP)

Author Contributions

Conceptualization: Lifeng Lin.

Formal analysis: Lifeng Lin.

Funding acquisition: Lifeng Lin.

Methodology: Lifeng Lin.

Software: Lifeng Lin.

Writing – original draft: Lifeng Lin.

References

1. Sutton AJ, Higgins JPT. Recent developments in meta-analysis. Statistics in Medicine. 2008; 27 (5):625–50. https://doi.org/10.1002/sim.2934 PMID: 17590884
2. Berlin JA, Golub RM. Meta-analysis as evidence: building a better pyramid. JAMA. 2014; 312(6):603–6. https://doi.org/10.1001/jama.2014.8167 PMID: 25117128
3. Gurevitch J, Koricheva J, Nakagawa S, Stewart G. Meta-analysis and the science of research synthesis. Nature. 2018; 555:175–82. https://doi.org/10.1038/nature25753 PMID: 29517004
4. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ. 2003; 327(7414):557–60. https://doi.org/10.1136/bmj.327.7414.557 PMID: 12958120
5. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Statistics in Medicine. 2002; 21(11):1539–58. https://doi.org/10.1002/sim.1186 PMID: 12111919
6. Lin L, Chu H, Hodges JS. Alternative measures of between-study heterogeneity in meta-analysis: reducing the impact of outlying studies. Biometrics. 2017; 73(1):156–66. https://doi.org/10.1111/biom.12543 PMID: 27167143
7. Hoaglin DC. Practical challenges of I^2 as a measure of heterogeneity. Research Synthesis Methods. 2017; 8(3):254. https://doi.org/10.1002/jrsm.1251 PMID: 28631294
8. Normand S-LT. Meta-analysis: formulating, evaluating, combining, and reporting. Statistics in Medicine. 1999; 18(3):321–59. PMID: 10070677
9. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Research Synthesis Methods. 2010; 1(2):97–111. https://doi.org/10.1002/jrsm.12 PMID: 26061376
10. Hoaglin DC. Misunderstandings about Q and 'Cochran's Q test' in meta-analysis. Statistics in Medicine. 2016; 35(4):485–95. https://doi.org/10.1002/sim.6632 PMID: 26303773
11. Egger M, Davey Smith G. Meta-analysis: potentials and promise. BMJ. 1997; 315(7119):1371–4. PMID: 9432250

12. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLOS Medicine. 2009; 6(7):e1000100. https://doi.org/10.1371/journal.pmed.1000100 PMID: 19621070
13. Ioannidis JPA, Patsopoulos NA, Rothstein HR. Reasons or excuses for avoiding meta-analysis in forest plots. BMJ. 2008; 336(7658):1413–5. https://doi.org/10.1136/bmj.a117 PMID: 18566080
14. Hedges LV, Olkin I. Statistical Methods for Meta-Analysis. Orlando, FL: Academic Press; 1985.
15. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ. 1997; 315(7109):629–34. PMID: 9310563
16. Sutton AJ, Song F, Gilbody SM, Abrams KR. Modelling publication bias in meta-analysis: a review. Statistical Methods in Medical Research. 2000; 9(5):421–45. https://doi.org/10.1177/096228020000900503 PMID: 11191259
17. Kirkham JJ, Dwan KM, Altman DG, Gamble C, Dodd S, Smyth R, et al. The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. BMJ. 2010; 340:c365. https://doi.org/10.1136/bmj.c365 PMID: 20156912
18. Lin L, Chu H. Quantifying publication bias in meta-analysis. Biometrics. 2017:In press. https://doi.org/10.1111/biom.12817 PMID: 29141096
19. Lin L, Chu H, Murad MH, Hong C, Qu Z, Cole SR, et al. Empirical comparison of publication bias tests in meta-analysis. Journal of General Internal Medicine. 2018; 33(8):1260–7. https://doi.org/10.1007/s11606-018-4425-7 PMID: 29663281
20. Murad MH, Chu H, Lin L, Wang Z. The effect of publication bias magnitude and direction on the certainty in evidence. BMJ Evidence-Based Medicine. 2018; 23(3):84–6. https://doi.org/10.1136/bmjebm-2018-110891 PMID: 29650725
21. Cochran WG, Carroll SP. A sampling investigation of the efficiency of weighting inversely as the estimated variance. Biometrics. 1953; 9(4):447–59.
22. Hedges LV. An unbiased correction for sampling error in validity generalization studies. Journal of Applied Psychology. 1989; 74(3):469–77.
23. Böhning D, Malzahn U, Dietz E, Schlattmann P, Viwatwongkasem C, Biggeri A. Some general points in estimating heterogeneity variance with the DerSimonian–Laird estimator. Biostatistics. 2002; 3(4):445–57. https://doi.org/10.1093/biostatistics/3.4.445 PMID: 12933591
24. DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled Clinical Trials. 1986; 7(3):177–88. PMID: 3802833
25. Jackson D. The implications of publication bias for meta-analysis' other parameter. Statistics in Medicine. 2006; 25(17):2911–21. https://doi.org/10.1002/sim.2293 PMID: 16345059
26. Davey J, Turner RM, Clarke MJ, Higgins JPT. Characteristics of meta-analyses and their component studies in the *Cochrane Database of Systematic Reviews*: a cross-sectional, descriptive analysis. BMC Medical Research Methodology. 2011; 11:160. https://doi.org/10.1186/1471-2288-11-160 PMID: 22114982
27. Easterbrook PJ, Gopalan R, Berlin JA, Matthews DR. Publication bias in clinical research. The Lancet. 1991; 337(8746):867–72.
28. Gøtzsche PC. Reference bias in reports of drug trials. BMJ. 1987; 295(6599):654–6. PMID: 3117277
29. Gilbert JR, Williams ES, Lundberg GD. Is there gender bias in *JAMA*'s peer review process? JAMA. 1994; 272(2):139–42. PMID: 8015126
30. Egger M, Zellweger-Zähner T, Schneider M, Junker C, Lengeler C, Antes G. Language bias in randomised controlled trials published in English and German. The Lancet. 1997; 350(9074):326–9.
31. Macaskill P, Walter SD, Irwig L. A comparison of methods to detect publication bias in meta-analysis. Statistics in Medicine. 2001; 20(4):641–54. https://doi.org/10.1002/sim.698 PMID: 11223905
32. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Comparison of two methods to detect publication bias in meta-analysis. JAMA. 2006; 295(6):676–80. https://doi.org/10.1001/jama.295.6.676 PMID: 16467236
33. Shuster JJ. Empirical vs natural weighting in random effects meta-analysis. Statistics in Medicine. 2010; 29(12):1259–65. https://doi.org/10.1002/sim.3607 PMID: 19475538
34. Cooper H, Hedges LV, Valentine JC. The Handbook of Research Synthesis and Meta-Analysis. 2nd ed. New York, NY: Russell Sage Foundation; 2009.
35. Casella G, Berger RL. Statistical Inference. 2nd ed. Belmont, CA: Duxbury Press; 2001.
36. Veroniki AA, Jackson D, Viechtbauer W, Bender R, Bowden J, Knapp G, et al. Methods to estimate the between-study variance and its uncertainty in meta-analysis. Research Synthesis Methods. 2016; 7(1):55–79. https://doi.org/10.1002/jrsm.1164 PMID: 26332144

37. Grissom RJ, Kim JJ. Effect Sizes for Research: A Broad Practical Approach. Mahwah, NJ: Lawrence Erlbaum Associates; 2005.
38. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
39. Malzahn U, Böhning D, Holling H. Nonparametric estimation of heterogeneity variance for the standardised difference used in meta-analysis. Biometrika. 2000; 87(3):619–32.
40. Egger M, Davey Smith G, Altman DG. Systematic Reviews in Health Care: Meta-Analysis in Context. 2nd ed. London, UK: BMJ Publishing Group; 2001.
41. Hedges LV. Distribution theory for Glass's estimator of effect size and related estimators. Journal of Educational Statistics. 1981; 6(2):107–28.
42. Bland JM, Altman DG. The odds ratio. BMJ. 2000; 320(7247):1468. PMID: 10827061
43. Higgins JPT, Green S. Cochrane Handbook for Systematic Reviews of Interventions. Chichester, UK: John Wiley & Sons; 2008.
44. Haldane JBS. The estimation and significance of the logarithm of a ratio of frequencies. Annals of Human Genetics. 1956; 20(4):309–11. PMID: 13314400
45. Gart JJ, Pettigrew HM, Thomas DG. The effect of bias, variance estimation, skewness and kurtosis of the empirical logit on weighted least squares analyses. Biometrika. 1985; 72(1):179–90.
46. Pettigrew HM, Gart JJ, Thomas DG. The bias and higher cumulants of the logarithm of a binomial variate. Biometrika. 1986; 73(2):425–35.
47. Sweeting MJ, Sutton AJ, Paul LC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. Statistics in Medicine. 2004; 23(9):1351–75. https://doi.org/10.1002/sim.1761 PMID: 15116347
48. Bradburn MJ, Deeks JJ, Berlin JA, Localio AR. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. Statistics in Medicine. 2007; 26(1):53–77. https://doi.org/10.1002/sim.2528 PMID: 16596572
49. Cai T, Parast L, Ryan L. Meta-analysis for rare events. Statistics in Medicine. 2010; 29(20):2078–89. https://doi.org/10.1002/sim.3964 PMID: 20623822
50. Rücker G, Schwarzer G, Carpenter J, Olkin I. Why add anything to nothing? The arcsine difference as a measure of treatment effect in meta-analysis with zero cells. Statistics in Medicine. 2009; 28(5):721–38. https://doi.org/10.1002/sim.3511 PMID: 19072749
51. Viechtbauer W. Conducting meta-analyses in R with the metafor package. Journal of Statistical Software. 2010; 36:3.
52. Jackson D, Bowden J, Baker R. How does the DerSimonian and Laird procedure for random effects meta-analysis compare with its more efficient but harder to compute counterparts? Journal of Statistical Planning and Inference. 2010; 140(4):961–70.
53. Sidik K, Jonkman JN. A comparison of heterogeneity variance estimators in combining results of studies. Statistics in Medicine. 2007; 26(9):1964–81. https://doi.org/10.1002/sim.2688 PMID: 16955539
54. Paule RC, Mandel J. Consensus values and weighting factors. Journal of Research of the National Bureau of Standards. 1982; 87(5):377–85.
55. van Aert RCM, Jackson D. Multistep estimators of the between-study variance: the relationship with the Paule-Mandel estimator. Statistics in Medicine. 2018; 37(17):2616–29. https://doi.org/10.1002/sim.7665 PMID: 29700839
56. Sterne JAC, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. Journal of Clinical Epidemiology. 2001; 54(10):1046–55. PMID: 11576817
57. Tang J-L. Weighting bias in meta-analysis of binary outcomes. Journal of Clinical Epidemiology. 2000; 53(11):1130–6. PMID: 11106886
58. Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: a comparative study. Statistics in Medicine. 1995; 14(24):2685–99. PMID: 8619108
59. Warn DE, Thompson SG, Spiegelhalter DJ. Bayesian random effects meta-analysis of trials with binary outcomes: methods for the absolute risk difference and relative risk scales. Statistics in Medicine. 2002; 21(11):1601–23. https://doi.org/10.1002/sim.1189 PMID: 12111922
60. Chu H, Nie L, Chen Y, Huang Y, Sun W. Bivariate random effects models for meta-analysis of comparative studies with binary outcomes: methods for the absolute risk difference and relative risk. Statistical Methods in Medical Research. 2012; 21(6):621–33. https://doi.org/10.1177/0962280210393712 PMID: 21177306
61. Stijnen T, Hamza TH, Özdemir P. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. Statistics in Medicine. 2010; 29(29):3046–67. https://doi.org/10.1002/sim.4040 PMID: 20827667

62. Jackson D, Law M, Stijnen T, Viechtbauer W, White IR. A comparison of seven random-effects models for meta-analyses that estimate the summary odds ratio. Statistics in Medicine. 2018; 37:1059–85. https://doi.org/10.1002/sim.7588 PMID: 29315733