

RESEARCH ARTICLE

Sensitivity analysis with iterative outlier detection for systematic reviews and meta-analyses

Zhuo Meng¹  | Jingshen Wang² | Lifeng Lin³  | Chong Wu⁴ 

¹Department of Statistics, College of Arts and Sciences, Florida State University, Tallahassee, Florida, USA

²Division of Biostatistics, School of Public Health, University of California, Berkeley, California, USA

³Department of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health, The University of Arizona, Tucson, Arizona, USA

⁴Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA

Correspondence

Lifeng Lin, Department of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health, 1200 E University Blvd, Tucson, AZ 85721, USA.
Email: lifenglin@arizona.edu

Chong Wu, Department of Biostatistics, The University of Texas MD Anderson Cancer Center, 12.3455, 7007 Bertner Avenue, Unit 1689, Houston, TX 77030, USA.
Email: cwu18@mdanderson.org

Present address

Zhuo Meng, Florida State University, 600 W College Ave, Tallahassee, Florida, 32306, USA

Funding information

US National Institutes of Health/National Institute of Mental Health, Grant/Award Number: R03 MH128727; National Institutes of Health/National Library, Grant/Award Number: R03 AG070669; US National Institutes of Health/National Institute on Aging, Grant/Award Number: R01 LM012982

Meta-analysis is a widely used tool for synthesizing results from multiple studies. The collected studies are deemed heterogeneous when they do not share a common underlying effect size; thus, the factors attributable to the heterogeneity need to be carefully considered. A critical problem in meta-analyses and systematic reviews is that outlying studies are frequently included, which can lead to invalid conclusions and affect the robustness of decision-making. Outliers may be caused by several factors such as study selection criteria, low study quality, small-study effects, and so on. Although outlier detection is well-studied in the statistical community, limited attention has been paid to meta-analysis. The conventional outlier detection method in meta-analysis is based on a leave-one-study-out procedure. However, when calculating a potentially outlying study's deviation, other outliers could substantially impact its result. This article proposes an iterative method to detect potential outliers, which reduces such an impact that could confound the detection. Furthermore, we adopt bagging to provide valid inference for sensitivity analyses of excluding outliers. Based on simulation studies, the proposed iterative method yields smaller bias and heterogeneity after performing a sensitivity analysis to remove the identified outliers. It also provides higher accuracy on outlier detection. Two case studies are used to illustrate the proposed method's real-world performance.

KEYWORDS

heterogeneity, iterative method, meta-analysis, outlier, sensitivity analysis

1 | INTRODUCTION

Meta-analysis has been an increasingly used tool for synthesizing findings from various studies on the same research topic in a systematic review.¹ Many well-designed guidelines have been developed to improve the conduct and reporting of

meta-analyses and systematic reviews and assess their quality.²⁻⁵ A fundamental step for performing a systematic review and meta-analysis is to select eligible studies based on predefined criteria. Differences in the study selection could lead to conflicting conclusions⁶⁻¹⁰; such inconsistency could be caused by the inclusion of outlying studies.¹¹ After identifying the eligible studies, it is critical to assess the heterogeneity between the studies. If the selected studies are substantially heterogeneous (eg, based on the I^2 statistic),¹² systematic reviewers need to explore the factors attributable to the heterogeneity. When some variables (eg, types of studies and summaries of population ages) may be associated with the effect results, the studies can be classified into several subgroups based on such variables.¹³

Systematic reviews frequently include studies that appear to be outlying.^{11,14-18} Such studies could seriously affect the procedures (eg, assessment of heterogeneity and publication bias) and conclusions of meta-analyses.¹⁹⁻²² They are possibly caused by reporting errors in the original studies, inappropriate study selection criteria, low study quality, small-study effects, and differences in certain characteristics (eg, study location) from other studies.²³⁻²⁶ In the presence of potential outlying studies, it is recommended to perform a sensitivity analysis of excluding these studies and validate the evidence from the meta-analysis.²⁷ In order to perform such sensitivity analyses, potential outlying studies should be detected appropriately. Outliers could be detected from clinical perspectives (eg, the possibility of extreme treatment effects) or formal statistical methods. This article focuses on the latter.

Although outlier detection has been a well-studied topic in the statistical community, relatively few efforts have been especially paid to the field of meta-analysis.²⁸⁻³⁵ Hedges and Olkin³⁶ (Chapter 12) presented several early efforts for detecting outliers in a meta-analysis, but they primarily considered the common-effect setting, where the studies are assumed to share a common true treatment effect. The leave-one-study-out procedure by Viechtbauer and Cheung³⁷ in the random-effects setting is perhaps the most widely used tool for detecting outliers in the current practice of meta-analysis. It calculates the studentized deleted residual for each individual study; a large studentized deleted residual indicates that the corresponding study may be outlying. In the standardization process for a specific study, the overall mean effect and between-study variance are estimated using all remaining studies. If this study is truly outlying, it would seriously affect the estimation and bias the studentized deleted residual.

This article introduces an iterative method of detecting outlying studies for a sensitivity analysis in meta-analysis. The iterative method refines the leave-one-study-out procedure by Viechtbauer and Cheung.³⁷ In the leave-one-study-out procedure, only a single study is removed to reduce its potential impact on the parameter estimation for calculating its studentized deleted residual. When multiple outlying studies are present in a meta-analysis, other outlying studies could still substantially affect the studentized deleted residuals. The idea of the proposed iterative method is similar to the step-wise variable selection for regression analysis. It is also used in other statistical applications, such as detecting pleiotropy in Mendelian randomization analysis with GWAS summary data.³⁸ Furthermore, the leave-one-study-out procedure ignores the impact of outlier detection (ie, model selection) when conducting inference. Model selection can yield noticeable changes in estimates in terms of discontinuities at the boundaries between model regimes. Hence, ignoring model selection usually, but not always, results in a smaller variation, which may lead to inflated Type I error rates and conservative coverage probabilities. To ensure valid inference, we adopt bagging with the nonparametric delta method,³⁹ which improves the performance for both the leave-one-study-out procedure and iterative method by reducing the variability in outlier detection when taking account of model selection.

This article is organized as follows. First, we introduce the conventional leave-one-study-out procedure and propose a new iterative method for detecting outlying studies. Second, we introduce the simulation designs and describe the meta-analyses for case studies. Third, we present the results from the simulation studies and the case studies to compare the different methods. This article concludes with a brief discussion.

2 | METHODS

2.1 | Meta-analysis models

Suppose a meta-analysis collects n independent studies. The effect size estimate of study i , denoted by y_i , is the outcome measure reported in each study, such as the mean difference, (log) odds ratio (OR), (log) relative risk (RR), etc. For each study, we assume the effect size is drawn from a sampling distribution such that

$$y_i \sim N(\mu_i, s_i^2),$$

where μ_i and s_i denote the true effect size and SE of study i , respectively. When these collected studies are homogeneous ($\mu_i = \mu$ for all collected studies), the overall effect size μ is estimated by the common-effect model as

$$\hat{\mu}_{CE} = \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i},$$

where $w_i = 1/s_i^2$. The variance of $\hat{\mu}_{CE}$ can be estimated by

$$\text{Var}(\hat{\mu}_{CE}) = \frac{1}{\sum_{i=1}^n w_i}. \quad (1)$$

The Q statistic, $Q = \sum_{i=1}^n w_i (y_i - \hat{\mu}_{CE})^2$, is commonly used for testing homogeneity. It follows χ_{n-1}^2 under the null hypothesis that the collected studies are homogeneous.

When the collected studies are considered heterogeneous, the corresponding random-effects model assumes $\mu_i \sim N(\mu, \tau^2)$, where τ^2 denotes the between-study variance and μ represents the overall mean effect size.⁴⁰ The restricted maximum likelihood (REML) method is recommended for obtaining the between-study variance estimate $\hat{\tau}^2$.⁴¹⁻⁴³ The overall mean effect size μ in the random-effects model can be estimated by

$$\hat{\mu}_{RE} = \frac{\sum_{i=1}^n w_i^* y_i}{\sum_{i=1}^n w_i^*},$$

where the weights $w_i^* = 1/(\hat{\tau}^2 + s_i^2)$. The variance of $\hat{\mu}_{RE}$ can be estimated by

$$\text{Var}(\hat{\mu}_{RE}) = \frac{1}{\sum_{i=1}^n w_i^*}. \quad (2)$$

Note that the random-effects model is reduced to the common-effect model when $\tau^2 = 0$.

2.2 | The leave-one-study-out procedure

One popular method to detect outliers in a meta-analysis is based on the studentized deleted residuals.³⁶ Specifically, for study i , the studentized deleted residual t_i is given by

$$t_i = \frac{y_i - \hat{\mu}_{i(-i)}}{\sqrt{\text{Var}[y_i - \hat{\mu}_{i(-i)}]}},$$

where $\hat{\mu}_{i(-i)}$ is the predicted mean effect size for study i calculated by the model that excludes study i during model fitting. Because y_i and $\hat{\mu}_{i(-i)}$ are uncorrelated, t_i can be further simplified to

$$t_i = \frac{y_i - \hat{\mu}_{i(-i)}}{\sqrt{s_i^2 + \hat{\tau}_{(-i)}^2 + \text{Var}[\hat{\mu}_{i(-i)}]}},$$

where $\hat{\tau}_{(-i)}^2$ and $\text{Var}[\hat{\mu}_{i(-i)}]$ are the estimated heterogeneity variance and variance of $\hat{\mu}_{i(-i)}$ from the model that excludes the i th study during the model fitting, respectively. When the meta-analysis considered does not incorporate study-level covariate information (which is often the case in applications), we have $\hat{\mu}_{i(-i)} = \hat{\mu}_{(-i)}$, with its variance estimated by Equations (1) and (2), respectively. Therefore, t_i becomes

$$t_i = \frac{y_i - \hat{\mu}_{(-i)}}{\sqrt{s_i^2 + \hat{\tau}_{(-i)}^2 + \text{Var}[\hat{\mu}_{(-i)}]}},$$

The studentized deleted residuals t_i follow a standard normal distribution approximately under the assumed model. One can use their magnitudes and P values to determine the outliers in a meta-analysis.

2.3 | The iterative method

The above leave-one-study-out procedure works well when only one outlier exists in a meta-analysis. However, when multiple outliers exist in a meta-analysis, $\hat{\mu}_{i(-i)}$ would deviate from the truth, and the leave-one-study-out procedure may be sub-optimal. To improve the robustness of the outlier detection, we propose an iterative approach that considers the impact of several potential outliers:

Step 1. Initialization: We obtain the initial overall mean effect estimate based on the model that includes all studies.

Step 2. Outlier test: To test whether study i is an outlier, we use a modified studentized deleted residual given by

$$\tilde{t}_i = \frac{y_i - \tilde{\mu}_{i(-i)}}{\sqrt{\text{Var}[y_i - \tilde{\mu}_{i(-i)}]}},$$

where $\tilde{\mu}_{i(-i)}$ is the predicted mean effect size for study i calculated by the model that excludes study i and the outliers detected in the previous iteration during the model fitting. Similar to the usual studentized deleted residuals, $\tilde{t}_i$ follows a standard normal distribution approximately under the assumed model, and their P values can be calculated accordingly. We apply these modified studentized deleted residuals to all n studies and treat a study as outlying if it meets a defined selection criterion, such as with a small P value or a large quantity. We adopt a selection criterion of $P < 0.05$ in the following analysis for generality. To improve the stability of the method, we only declare the study with the smallest P value as the new outlier when multiple studies yield $P < 0.05$. In application, one may consult clinicians to find the most appropriate selection criterion.

Step 3. Estimation of mean effect size: We estimate the mean effect based on the model that excludes all outlying studies detected in outlier tests (in Step 2) during the model fitting.

Step 4. Iteration: We iterate Steps 2 and 3 until there is no change in outlier detection and the mean effect estimate. Note that if a study is declared as an outlier in the previous iterations but shows a non-significant P value in the current iteration, we will bring this study back to the meta-analysis.

Generally speaking, we want to select the best possible set of outlying studies, which however is NP-hard. The proposed iterative method is one type of greedy search. According to Elenberg et al.,⁴⁴ this new method should perform within a constant factor from the best possible subset-selection solution for detecting outlying studies, partially explaining why the proposed method achieves superior performance as to be shown later.

2.4 | Bagging

To reduce variability and eliminate discontinuities in outlier detection, we further use bagging to compute the average effects and the confidence intervals (CIs).⁴⁵ Specifically, we generate bootstrap replicates

$$\mathbf{y}_j^* = (y_{j1}^*, y_{j2}^*, \dots, y_{jk}^*, \dots, y_{jn}^*),$$

for $j = 1, \dots, B$. The bootstrap samples $\mathbf{y}_j^*$ consist of n draws from $y_1, y_2, \dots, y_n$ with replacement such that for each y_{jk}^* ,

$$\Pr(y_{jk}^* = y_i) = \frac{1}{n},$$

for all i . Here, we adopt bootstrap sampling with replacement (ie, each individual study has an equal chance of being resampled) so that valid confidence can be built via the nonparametric delta method.³⁹ While other parametric bootstrap methods might be used as an alternative tool to construct a point estimate of μ , the resulting estimator may not enable valid statistical inference after outlier detection.

We then apply the proposed iterative outlier detection method and the leave-one-study-out method in each replicative bootstrap sample to remove the outliers and obtain the overall average effect estimate $\hat{\mu}_j^*$. Then, the smoothed average

effect estimate is obtained by averaging over B bootstrap replications,

$$\tilde{\mu} = \frac{1}{B} \sum_{j=1}^B \hat{\mu}_j^*.$$

We note that $\tilde{\mu}$ is a finite sample approximation of $E[\hat{\mu}_j^* | y_1, y_2, \dots, y_n]$, and it is thus correctly centered on μ as long as the majority of the bootstrap samples are correctly centered on μ .

The 95% bootstrap smoothed interval is obtained by

$$\tilde{\mu} \pm 1.96 \tilde{\text{sd}}_B,$$

where $\tilde{\text{sd}}_B$ is the nonparametric delta-method estimate of the SD for $\tilde{\mu}$.³⁹ Specifically, for the j th bootstrap replication, let

$$Y_{ji}^* = \# \{ y_{jk}^* = y_i \},$$

be the number of elements of y_j^* that equals the original effect sizes of the studies y_i . The estimate of SD for $\tilde{\mu}$ is

$$\tilde{\text{sd}}_B = \left[\sum_{i=1}^n \text{cov}_i^2 \right]^{1/2},$$

where the covariance between Y_{ji}^* and $\hat{\mu}_j^*$ can be estimated by

$$\widehat{\text{cov}}_i = \sum_{j=1}^B (Y_{ji}^* - Y_{.i}^*) (\hat{\mu}_j^* - \tilde{\mu}) / B,$$

with $Y_{.i}^* = \sum_{j=1}^B Y_{ji}^* / B$ and $\tilde{\mu} = \sum_{j=1}^B \hat{\mu}_j^* / B$.

3 | SIMULATIONS

3.1 | Simulation designs

We conducted a comprehensive simulation study to evaluate the performance of the proposed iterative outlier detection method compared with the leave-one-study-out method. For each simulated meta-analysis, the true overall effect size was $\mu = 0$. The study-specific effect sizes were generated by $y_i \sim N(\mu_i, s_i^2)$ and $\mu_i \sim N(\mu, \tau^2)$. The within-study SEs followed a uniform distribution $s_i \sim U(1, 4)$, and the between-study SD τ took values of 0, 1, and 4. The values of $\tau = 0, 1$, and 4 represented weak, moderate, and strong heterogeneity in the meta-analysis, respectively. The number of studies collected in each meta-analysis was assumed to be $n = 15$ or 30. We further assumed the number of outlier studies to be $m = 0$, $\lceil n/5 \rceil$ and $\lceil n/10 \rceil$, where $\lceil x \rceil$ denotes the smallest integer not less than x . Both the z test and t test were considered for the studentized deleted residuals; the t test statistic had $n - r - 2$ degrees of freedom, where r is the number of detected outliers. We performed 1000 simulations with 500 repetitions of bootstrap sampling. The following two scenarios were considered for generating outlier studies.

- I Assuming the outlier studies were distributed evenly on both sides of the true overall effect size. Specifically, we added/subtracted a constant $\delta = 8$ to/from the effect sizes of the m randomly selected outlier studies according to their signs, so that these m studies would have unusually large or small effect sizes compared to the rest studies. The size of δ is determined based roughly on the magnitudes of the within- and between-study variances, such that δ is about two times (the upper bound of) SEs of the effect sizes.

II Assuming the outlying studies were distributed on one side only, which is more likely to be the case in the presence of publication bias, a phenomenon that studies with plausible and significant results are more likely to be published.^{46,47} Specifically, we randomly selected m studies and added $\delta = 8$ to their absolute effect sizes, resulting in m outlying studies with large effect sizes.

We would evaluate these two methods by looking into a few statistics and diagnostic results. The details are presented in the Simulation results. We would also compare the two methods to the original meta-analyses and the meta-analyses with the true outlying studies excluded.

3.2 | Simulation results

We report statistics related to the overall effect size estimates to evaluate the outliers' impact and diagnostic statistics for binary outcomes to assess the accuracy of outlier detection. To save space, we focus on the setting with outliers on one side only (scenario II) with the number of outliers $m = n/5$ or no outlier ($m = 0$) using the z test. The complete results for other scenarios and settings are similar and are given in Tables S1–S20 in the Supporting Information. The Monte Carlo SEs of the primary results were reported in Tables S21–S24 in the Supporting Information.

We first considered the case that there were actual outlying studies. Table 1 presents results for $n = 15$ and 30, including bias, mean squared errors (MSEs), type I error rates, and coverage probabilities of 95% CIs of the overall effect size estimate $\hat{\mu}$ before and after removing the outliers by different methods. The results of performing bagging for the leave-one-study-out method and the iterative outlier detection method are presented as well. We also report the type I error rates ($\tau = 0$) and powers ($\tau > 0$) of the Q statistic for testing homogeneity and the mean of I^2 .

When the studies were homogeneous ($\tau = 0$), the biases in the overall effect size estimate $\hat{\mu}$ were negligible with small MSEs when the true outlying studies were excluded. The type I error rates of testing $\mu = 0$ were well-controlled, and the coverage probabilities were close to the nominal level (95%). The type I error rates of the Q statistic were also controlled, and the I^2 was less than 10%. These may reveal that the biases and the heterogeneity in the original meta-analyses were induced by the presence of outlying studies, given when all studies were included, the biases in $\hat{\mu}$ were significant with large MSEs, the type I error rates of both $\mu = 0$ and the Q statistic were inflated, and the I^2 values were about 70%. By conducting the sensitivity analyses of removing detected outlying studies using both outlier detection methods, the iterative outlier detection method yielded smaller biases with smaller MSEs than the leave-one-study-out method. However, the type I error rates of the overall effect size were inflated for both methods due to the randomness of outlier detection, and the coverage probabilities were also relatively low. By applying bagging, the type I error rates were well-controlled, and the coverage probabilities were increased for the iterative outlier detection method under both sample sizes and for the leave-one-study-out method when the sample size was small. The percentages of bias reduced from the original meta-analyses and the coverage probabilities after bagging were plotted in Figure 1. Besides, the iterative outlier detection method generated much smaller type I error rates for the Q statistic and produced I^2 closer to 0.

For heterogeneous studies ($\tau > 0$), the biases of the original meta-analyses increased due to the presence of outlying studies and the larger between-study variation, leading to inflated type I error rates and low coverage probabilities of the overall effect size. In the sensitivity analyses of excluding detected outlying studies, the iterative outlier detection method reduced significantly more bias than the leave-one-study-out method (see Figure 1). When the heterogeneity was moderate ($\tau = 1$), both methods controlled the type I error rates of $\hat{\mu}$ well with bagging, further highlighting the importance of considering outlier detection (model selection) when conducting inference. The Q statistic of the leave-one-study-out method was more powerful than that of the iterative outlier detection method; this was expected because the leave-one-study-out method was less effective in outlier detection and sacrificed the type I error of the Q statistic in the presence of outliers. Furthermore, the leave-one-study-out method produced much higher I^2 values than those of removing the true outliers, indicating a potential insufficiency of this detection process. On the other hand, the iterative outlier detection method shrank I^2 to around 10%, suggesting the iterative framework may be more powerful in detecting outliers. However, when heterogeneity was strong ($\tau = 4$), the type I error rates of $\hat{\mu}$ were inflated even when the true outliers were removed. Still, the iterative outlier detection method with bagging controlled the type I error rates better than that of the leave-one-study-out method. The I^2 values of the iterative outlier detection method remained substantial but were similar to that with true outliers removed, while those of the leave-one-study-out method were not.

Table 2 presents diagnostic statistics, including sensitivity, specificity, and several other diagnostic statistics that evaluate the accuracy of the outlier diagnosis. Across all settings, the iterative outlier detection method identified more outliers

TABLE 1 Results of simulation study under scenario II when the number of outliers is $m = \lceil n/5 \rceil$ using the z test.

	$n = 15$						$n = 30$					
	Bias	MSE	TIE	CP	Qp	I^2 (%)	Bias	MSE	TIE	CP	Qp	I^2 (%)
$\tau = 0$												
All studies	1.569	2.964	0.258	0.742	0.925	66.7	1.595	2.735	0.636	0.364	0.997	69.8
No outliers	0.015	0.392	0.037	0.963	0.053	9.7	0.009	0.189	0.050	0.950	0.046	7.3
LOSO	0.810	1.333	0.139	0.861	0.478	35.5	0.926	1.176	0.309	0.691	0.754	44.8
			(0.059)	(0.941)					(0.085)	(0.915)		
ITER	0.287	0.873	0.125	0.875	0.087	9.0	0.160	0.377	0.138	0.862	0.048	4.4
			(0.046)	(0.954)					(0.035)	(0.965)		
$\tau = 1$												
All studies	1.706	3.422	0.235	0.765	0.975	73.0	1.773	3.407	0.648	0.352	0.998	74.9
No outliers	-0.004	0.483	0.054	0.946	0.150	19.0	0.034	0.249	0.058	0.942	0.207	18.2
LOSO	0.865	1.511	0.149	0.851	0.613	44.9	1.077	1.604	0.343	0.657	0.877	55.4
			(0.052)	(0.948)					(0.105)	(0.895)		
ITER	0.253	1.035	0.156	0.844	0.128	12.7	0.223	0.572	0.181	0.819	0.094	8.4
			(0.036)	(0.964)					(0.050)	(0.950)		
$\tau = 4$												
All studies	2.226	6.392	0.217	0.783	0.999	87.7	2.259	5.794	0.510	0.490	1.000	88.6
No outliers	0.012	1.932	0.095	0.905	0.947	73.3	0.048	0.930	0.068	0.932	0.999	77.3
LOSO	1.733	5.160	0.237	0.763	0.990	81.6	1.924	4.718	0.442	0.558	1.000	84.7
			(0.116)	(0.884)					(0.232)	(0.768)		
ITER	1.159	5.207	0.340	0.660	0.678	57.2	1.255	3.959	0.441	0.559	0.797	62.9
			(0.083)	(0.917)					(0.112)	(0.888)		

Note: The statistics in the row “All studies” were obtained from meta-analyses containing all studies. No outlier represents the results after removing the true m outliers. LOSO represents the results after removing outliers identified by the leave-one-study-out method. ITER represents the results after removing outliers identified by the iterative outlier detection method. Bias shows the bias in the estimated overall effect size; the true effect size $\mu = 0$. MSE represents the mean squared errors. TIE represents the type I error rates; CP represents the coverage probabilities of 95% confidence intervals; those inside the parentheses were produced by bagging. Qp represents the type I error rates or powers of the Q statistic. I^2 represents the mean of I^2 among the simulation replications. All results were based on 1000 simulations with 500 bootstrap resampling.

than the leave-one-study-out method, and the number was closer to the actual number of outliers m . Besides, it had higher sensitivities, indicating that the iterative outlier detection method had a better ability to identify outliers correctly. The sensitivities of the iterative outlier detection method increased as the number of studies increased, while those of the leave-one-study-out method decreased. This reveals that the leave-one-study-out method was not optimal for multiple outliers; such an issue can be improved by identifying outliers iteratively. For both methods, it was easier to identify outliers correctly when the heterogeneity was weak or moderate. In exchange for higher sensitivities, the iterative outlier detection method shows slightly lower specificities than the leave-one-study-out method. However, its specificities remained reasonably high, especially when the studies were less heterogeneous. Of the statistics that measure the accuracy of outlier detection, they were all larger (ie, better) when the outliers were detected iteratively, except for the accuracy under the case that the studies were strongly heterogeneous. In summary, these results show that it was more efficient to detect outliers iteratively compared to the leave-one-study-out method.

Next, we considered the case when no outlying study was presented in the meta-analysis. Table 3 presents the results with the same statistics as in Table 1 for $n = 15$ and 30. When the studies were homogeneous or the heterogeneity was weak, the iterative outlier detection method showed similar performance to the leave-one-study-out procedure. When the heterogeneity was substantial, the iterative outlier detection method yielded higher type I error rates and lower coverage probabilities than that of the leave-one-study-out procedure. However, bagging substantially improves the performance,

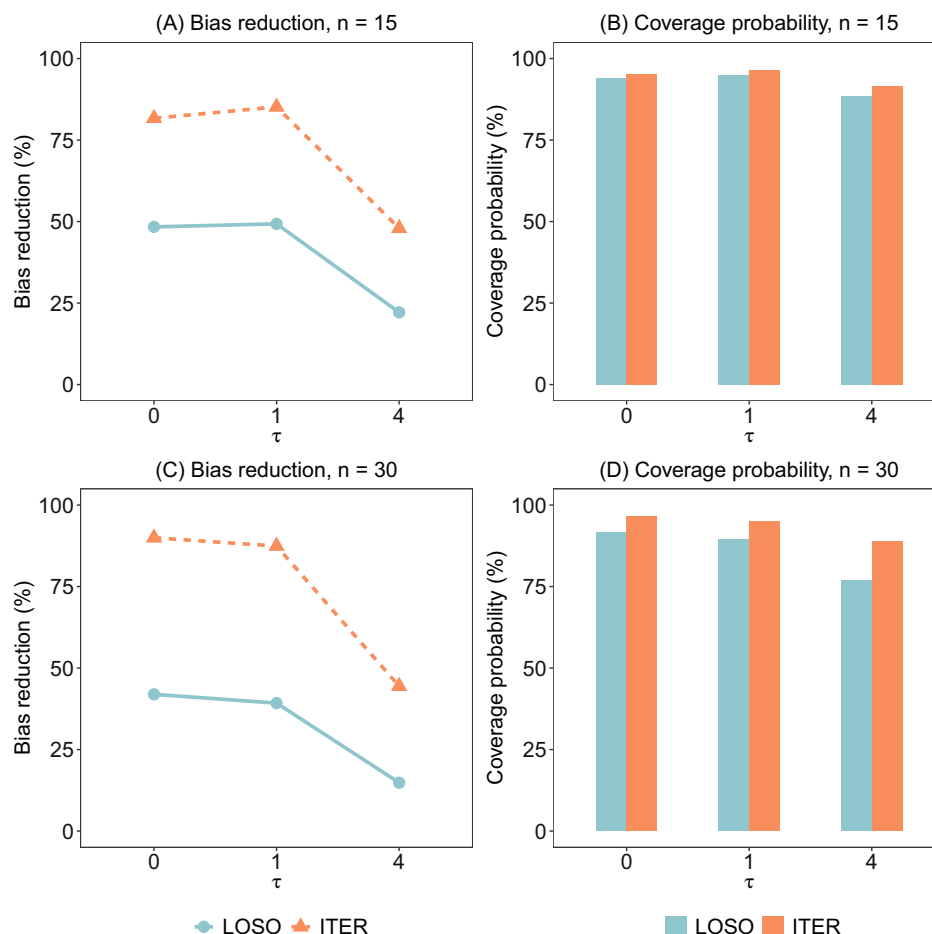


FIGURE 1 Percentage of bias reduced from the original meta-analysis and coverage probability of simulation study under scenario II when the number of outliers is $m = \lceil n/5 \rceil$ using the z test. The number of collected studies is 15 in panels (A) and (B); it is 30 in panels (C) and (D). LOSO represents the results after removing outliers identified by the leave-one-study-out method. ITER represents the results after removing outliers identified by the iterative outlier detection method. Percentages of reduction in bias from the original meta-analysis by LOSO and ITER are represented by the blue circles with solid lines and the orange triangles with dashed lines in panels (A) and (C). Coverage probability of LOSO and ITER are represented by the blue and orange bars in panels (B) and (D).

leading to valid inference for the iterative outlier detection method in most cases. Table 4 presents the number of identified outliers and their false positive rates for both methods. It appeared that the iterative outlier detection method might identify more outliers than the leave-one-study-out method when the heterogeneity was high, especially when the number of studies was large. When outliers were presented, the results were similar (Tables S13, S14, S17, and S18 in the Supporting Information). One may apply the iterative outlier detection method with caution in such a situation; a more stringent significance level may be used if substantial heterogeneity is detected.

Additional evidence further showed the bias in the overall effect size estimate $\hat{\mu}$ was small for both methods under scenario I (Tables S1–S4 in the Supporting Information). This may be because the outlying studies were evenly distributed on both sides of the true overall effect size so that the bias could be canceled out. Besides, when the number of outlying studies was small, the iterative method performed well with smaller biases, lower type I error rates, and higher coverage probabilities after bagging. The z test and t test derived similar conclusions across all settings; the z test seemed to have higher accuracy in outlier detection.

4 | CASE STUDIES

We illustrate the performance of the proposed iterative outlier detection method through two actual meta-analyses published in the *Cochrane Database of Systematic Reviews*. The first meta-analysis conducted by Fajardo-Bernal et al.⁴⁸ was aimed to evaluate the proportion of individuals tested for *Chlamydia trachomatis* (CT) and *Neisseria gonorrhoeae* (NG) in

TABLE 2 Diagnostic results of simulation study under scenario II when the number of outliers is $m = \lceil n/5 \rceil$ using the z test.

	$n = 15$						$n = 30$					
	$\tau = 0$		$\tau = 1$		$\tau = 4$		$\tau = 0$		$\tau = 1$		$\tau = 4$	
	LOSO	ITER	LOSO	ITER	LOSO	ITER	LOSO	ITER	LOSO	ITER	LOSO	ITER
Number of outliers	1.341	2.660	1.349	2.952	0.971	2.562	2.224	5.789	2.162	6.357	1.569	5.124
Sensitivity	0.405	0.734	0.415	0.775	0.240	0.502	0.339	0.790	0.333	0.822	0.181	0.499
Specificity	0.989	0.962	0.991	0.948	0.979	0.912	0.992	0.956	0.993	0.941	0.980	0.911
ACC	0.872	0.916	0.876	0.913	0.831	0.830	0.861	0.923	0.861	0.917	0.820	0.829
F ₁ score	0.535	0.741	0.548	0.751	0.334	0.461	0.481	0.792	0.476	0.786	0.272	0.470
MCC	0.536	0.715	0.551	0.721	0.319	0.413	0.498	0.757	0.496	0.748	0.271	0.420
Youden's J	0.394	0.696	0.406	0.723	0.219	0.414	0.331	0.747	0.326	0.762	0.160	0.411

Note: The accuracy (ACC) measures the proportion of correct diagnoses, including true positive and true negative, among all studies. F₁ score is the harmonic mean of the precision and recall. The Matthews correlation coefficient (MCC) measures the agreement between the observed and predicted binary classifications.

TABLE 3 Results of simulation study when there is no actual outlier ($m = 0$) using the z test.

	$n = 15$						$n = 30$					
	Bias	MSE	TIE	CP	Qp	I^2 (%)	Bias	MSE	TIE	CP	Qp	I^2 (%)
$\tau = 0$												
All studies	0.011	0.298	0.053	0.947	0.040	9.7	−0.012	0.138	0.039	0.961	0.047	7.5
LOSO	0.005	0.342	0.064	0.936	0.008	2.5	−0.014	0.169	0.058	0.942	0.004	1.0
			(0.094)	(0.906)					(0.056)	(0.944)		
ITER	0.008	0.361	0.071	0.929	0.006	2.3	−0.014	0.191	0.082	0.918	0.004	0.7
			(0.093)	(0.917)					(0.063)	(0.937)		
$\tau = 1$												
All studies	0.007	0.416	0.089	0.911	0.139	17.6	−0.003	0.181	0.072	0.928	0.228	17.8
LOSO	0.002	0.501	0.115	0.885	0.033	5.1	0.009	0.236	0.105	0.895	0.036	5.0
			(0.096)	(0.904)					(0.064)	(0.936)		
ITER	0.005	0.585	0.144	0.856	0.024	4.0	0.010	0.306	0.147	0.853	0.027	2.9
			(0.097)	(0.903)					(0.067)	(0.933)		
$\tau = 4$												
All studies	0.025	1.554	0.080	0.920	0.966	73.9	0.041	0.733	0.061	0.939	0.999	77.1
LOSO	−0.001	1.914	0.167	0.833	0.805	60.1	0.045	0.872	0.123	0.877	0.974	67.4
			(0.066)	(0.934)					(0.053)	(0.947)		
ITER	−0.007	2.579	0.293	0.707	0.535	41.6	0.021	1.483	0.295	0.705	0.675	46.1
			(0.054)	(0.946)					(0.041)	(0.959)		

Note: The statistics in the row “All studies” were obtained from meta-analyses containing all studies. LOSO represents the results after removing outliers identified by the leave-one-study-out method. ITER represents the results after removing outliers identified by the iterative outlier detection method. Bias shows the bias in the estimated overall effect size; the true effect size $\mu = 0$. MSE represents the mean squared errors. TIE represents the type I error rates; CP represents the coverage probabilities of 95% confidence intervals; those inside the parentheses were produced by bagging. Qp represents the type I error rates or powers of the Q statistic. I^2 represents the mean of I^2 among the simulation replications. All results were based on 1000 simulations with 500 bootstrap resampling.

TABLE 4 Diagnostic results of simulation study when there is no actual outlier ($m = 0$) using the z test.

	$n = 15$						$n = 30$					
	$\tau = 0$		$\tau = 1$		$\tau = 4$		$\tau = 0$		$\tau = 1$		$\tau = 4$	
	LOSO	ITER	LOSO	ITER	LOSO	ITER	LOSO	ITER	LOSO	ITER	LOSO	ITER
Number of outliers	0.676	0.677	0.920	0.957	1.046	2.124	1.411	1.441	1.888	2.143	1.718	4.355
False positive rate	0.045	0.045	0.061	0.064	0.070	0.142	0.047	0.048	0.063	0.071	0.057	0.145

home-based specimen collection compared to clinic-based specimen collection. This meta-analysis consisted of 10 studies with RRs reported originally. The second meta-analysis proposed by Jakobsen et al.⁴⁹ was aimed to investigate the effect of direct-acting antivirals (DAAs) on the prevention of sustained virological response, which is used as a surrogate outcome of morbidity and mortality for hepatitis C virus, compared to the control/placebo group. The trials assessing the effects of a DAA over or at the median dose were collected in this meta-analysis. This meta-analysis consisted of 34 studies with RRs reported originally.

In the first meta-analysis performed by Fajardo-Bernal et al.,⁴⁸ the random-effects model was applied due to the high heterogeneity ($I^2 = 98\%$). However, study 7 was apparently outlying according to the forest plot shown in Figure 2A. If study 7 had been removed, the remaining studies would have been much more homogeneous; thus, we tried to fit both common-effect and random-effects models. Figure 2B presents the studentized deleted residuals (without iteration) using both models. Under the common-effect setting, all studies except study 5 had large studentized deleted residuals. Therefore, the leave-one-study-out procedure adopting a common-effect model may be inappropriate due to the impact of multiple and extreme outlying studies. Under the random-effects setting, the leave-one-study-out procedure identified only study 7 as a potential outlier. The iterative outlier detection method identified the same outlying studies 7, 6, 5, and 2 under both common- and random-effects settings. The following results are presented under the random-effects setting.

Table 5 presents the original meta-analysis and the meta-analyses after performing sensitivity analyses with potential outlying studies removed. The estimated overall effect size $\hat{\mu}$ for the original meta-analysis was greatly impacted by the outlying study with a large SE. The P value of the Q test was extremely small ($<.0001$); $I^2 = 98\%$ suggested substantial between-study heterogeneity. When study 7 was removed by applying the leave-one-study-out procedure, $\hat{\mu}$ became smaller with a noticeably reduced SE. Accounting for the effect of model selection, the bagging procedure provided a smaller estimated overall effect size with a wider 95% CI. The P value of the Q test remained less than .0001; the heterogeneity measure did not change much ($I^2 = 95\%$). The substantial heterogeneity in the remaining studies indicated the possibility that the leave-one-study-out procedure might not be sufficient in this meta-analysis.

We then performed the iterative outlier detection method. The outliers identified were studies 7, 6, 5, and 2. When the four studies were removed, the estimated overall effect size was less than that in the original meta-analysis but greater than that when only study 7 was removed. This was expected because these four studies were on both sides of $\hat{\mu}$. The SE was reduced due to the removal of outliers. Bagging generated a smaller $\hat{\mu}$ and a wider CI that may improve the coverage of the true overall effect size. The Q statistic turned non-significant, indicating the remaining studies were homogeneous. Also, I^2 reduced noticeably ($I^2 = 31\%$). Therefore, the heterogeneity may be caused by these potential outlying studies. However, due to the low number of studies in total, clinical justification may be needed for whether an outlying study should be removed or not. Even if only parts of the identified outliers should be removed, the order of removing certain outlying studies was clear, and the impact of outlying studies could be reduced.

The outlier detection may be able to provide some insights for clinicians. For example, study 7 was detected by both methods. This study happened to have a significantly larger sample size than the other collected studies in this meta-analysis but had relatively few events in the clinical group. The reason might be this study was conducted in an early year (in 1998), and the population was high school students. Teenagers might not want to get tested for Chlamydia trachomatis (CT) and Neisseria gonorrhoeae (NG) from their doctors or at a local clinic. Another perspective is by our proposed iterative method, studies 7, 6, and 5 were considered potential outliers. We noticed that those three studies' inclusion criteria did not require CT positive but only those who were willing to participate. Thus, the CT diagnosis status may be a factor for subgroup analysis.

In the second meta-analysis performed by Jakobsen et al.,⁴⁹ the random-effects diagnostic procedure was performed. Figure 2C shows the forest plot with the observed log RRs and their within-study 95% CIs. Study 11 seemed to be outlying; the rest of the studies tended to be heterogeneous. Figure 2D shows the study-specific studentized deleted residuals

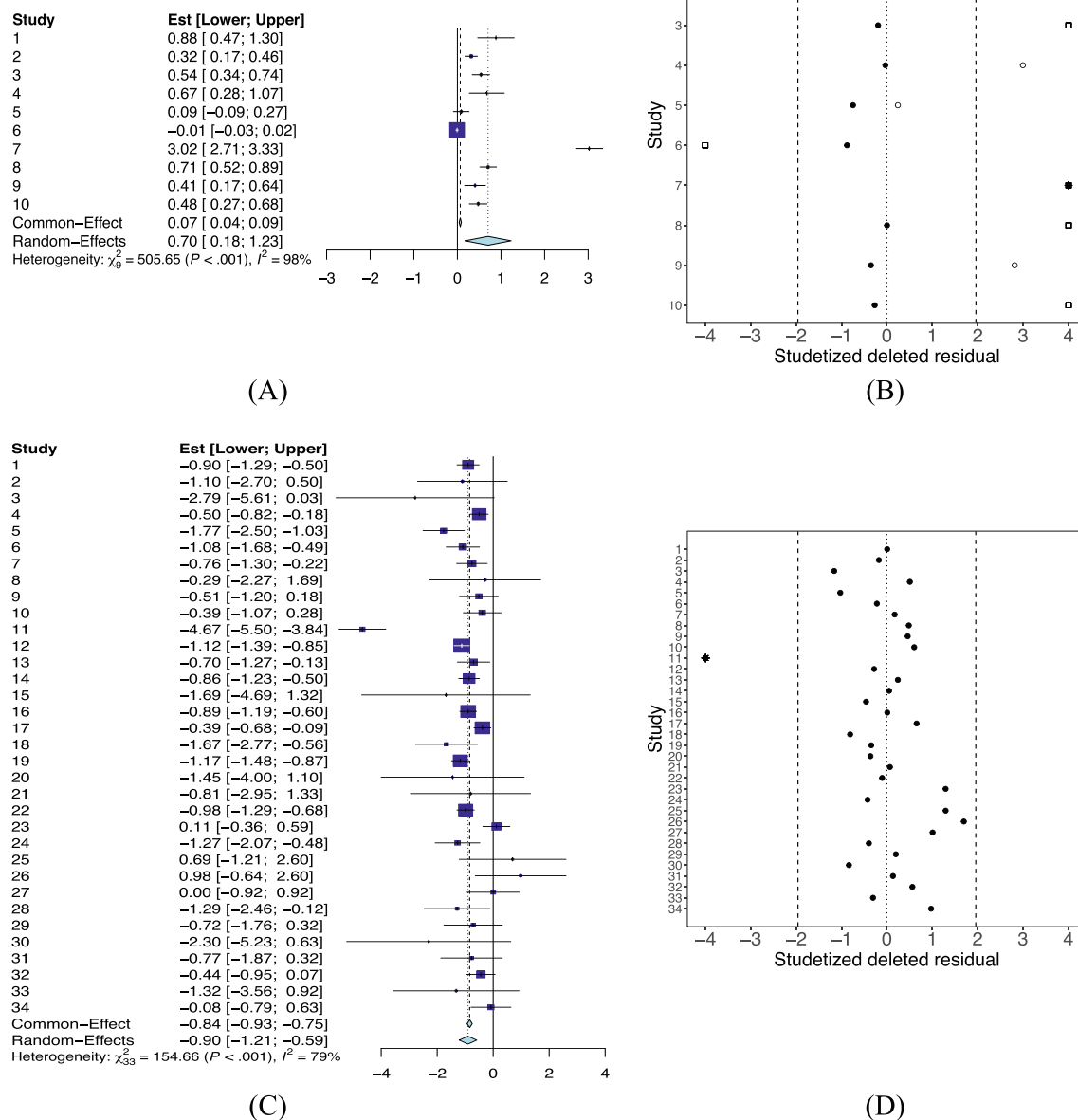


FIGURE 2 Forest plots and studentized deleted residual plots of two actual meta-analyses. The upper panels show the meta-analysis conducted by Fajardo-Bernal et al.⁴⁸; the lower panels show the meta-analysis conducted by Jakobsen et al.⁴⁹ In panels (A) and (C), the column “Est” contains log relative risks; “Lower” and “Upper” are the lower and upper bounds of 95% confidence intervals. In panels (B) and (D), the unfilled dots (squares) represent studentized deleted residuals (with absolute value greater than 4) under the common-effect setting; the filled dots (star) represent studentized deleted residuals (with absolute value greater than 4) under the random-effects setting.

achieved by the leave-one-study-out method using the random-effects model. The magnitude of the residual of study 11 was over 7, indicating that this study was an apparent outlier when the leave-one-study-out method was applied.

The results of this meta-analysis are summarized in Table 5. For the original meta-analysis, the studies were strongly heterogeneous according to the Q test and the large I^2 , with an effect size of -0.901 (95% CI: -1.213 to -0.587). We then performed the sensitivity analysis. With study 11 removed, the estimated overall effect size increased with a smaller SE, meaning that the difference between the experimental and control groups may actually be smaller. The Q statistic was still significant with $P < .0001$. The I^2 was reduced by more than 30% in absolute magnitude but remained fairly high (larger than or close to 50%), indicating substantial heterogeneity. Besides study 11, the iterative outlier detection method further identified studies 23, 5, and 26 as potential outliers. The effect sizes (-4.671 , 0.115 , -1.767 , and 0.981 accordingly)

TABLE 5 Results for the two actual meta-analyses.

		After removing outliers						
Removed studies		$\hat{\mu}$	SE	<i>P</i> value	95% CI	$\hat{\tau}^2$	<i>P</i> value of <i>Q</i> test	<i>I</i> ² (%)
Meta-analysis by Fajardo-Bernal et al. ⁴⁸								
Original	None	0.704	0.269	0.0089	0.176 to 1.232	0.709	<0.0001	98.2
LOSO	7	0.422	0.096	<0.0001	0.233 to 0.611	0.070	<0.0001	94.7
		(0.481)	(0.163)	(0.0031)	(0.160 to 0.801)			
ITER	7, 6, 5, 2	0.577	0.060	<0.0001	0.459 to 0.695	0.006	0.2006	31.3
		(0.504)	(0.177)	(0.0040)	(0.157 to 0.850)			
	7, 6	0.479	0.085	<0.0001	0.312 to 0.646	0.042	<0.0001	78.5
	7, 6, 5	0.530	0.069	<0.0001	0.395 to 0.666	0.019	0.0137	62.5
Meta-analysis by Jakobsen et al. ⁴⁹								
Original	None	−0.901	0.160	<0.0001	−1.213 to −0.587	0.600	<0.0001	78.7
LOSO	11	−0.763	0.088	<0.0001	−0.935 to −0.590	0.105	<0.0001	55.7
		(−0.772)	(0.098)	(<0.0001)	(−0.965 to −0.580)			
ITER	11, 23, 5, 26	−0.796	0.075	<0.0001	−0.942 to −0.649	0.050	0.0196	38.0
		(−0.806)	(0.126)	(<0.0001)	(−1.052 to −0.560)			

Note: The values and 95% confidence intervals inside parentheses were produced by bagging.

of the identified four studies fell outside of the 95% CI in the original meta-analysis under the random-effects setting where the heterogeneity was accounted for. With the four studies removed, the estimated overall effect size increased as well, with a reduced SE. The *P* value of the *Q* statistic changed noticeably even though it still rejected the null hypothesis of homogeneity. The *I*² reduced dramatically, indicating weak or moderate heterogeneity. Therefore, the iterative outlier detection method might perform better when the studies were heterogeneous, and the remaining studies may continue to contain statistical heterogeneity. In both the leave-one-study-out and iterative outlier detection procedures, bagging generated larger SEs with wider 95% CIs, which may provide better coverage probabilities of the truth.

We further tested our method on an additional meta-analysis conducted by Olliaro and Mussano.⁵⁰ The forest plot, studentized deleted residuals plot, results, and illustrations can be found in the Supporting Information.

5 | DISCUSSION

This article proposes an iterative procedure for detecting outlying studies. It can be broadly applied to a wide range of meta-analyses with various effect measures. Our approaches could identify outliers more accurately than the conventional leave-one-study-out procedures, as shown in the simulation studies. However, if a high heterogeneity is initially assessed, our approach may be used cautiously for the risk of false exclusion. The two case studies in the main manuscript and the additional case study in the supplementary have illustrated that different outlier detection methods could lead to fairly different conclusions. The scientific community continues to have concerns about research reproducibility and replicability. A systematic review might contain more than one study that could not be reproduced and may not replicate other studies on the same research question. The leave-one-study-out procedure could not rule out the impact of multiple outliers on the residuals used for outlier detection. The iterative outlier detection method is particularly useful in such cases.

Nevertheless, our approach has several limitations. First, the iterative outlier detection method is built on the conventional meta-analysis models, which assume within-study effect sizes follow normal distributions and treat within-study SEs as fixed values. In the random-effects setting, the between-study heterogeneity is also modeled with a normal distribution. These assumptions may be problematic, for example, in the case of small sample sizes.^{51,52} Similar ideas of iterative algorithms for outlier detection could be extended to one-stage meta-analysis methods, such as generalized linear mixed models or Bayesian hierarchical models.⁵³⁻⁵⁶ We leave these as future projects.

Second, this article only considers outlying studies from a statistical perspective. While it is critical to incorporate the clinical interpretations of studies' results in the outlier assessment, our criteria for declaring an outlying study are entirely based on studentized deleted residuals and their significance. In practice, one may consider using a cutoff other than $P < .05$ for outlier detection as they see fit. Clinical experts can also provide guidance to overwrite the selection criterion. Furthermore, our method should be used in a sensitivity analysis of removing identified outliers. We may borrow opinions from clinical experts to assess whether an extremely large effect size is realistic. Besides, our method could possibly provide a perspective about whether the clinical heterogeneity is caused by a certain factor; for example, the identified outlying studies are crossover studies, while the remaining studies are randomized control trials and observational studies.

Third, our approach does not involve assessing the quality of studies. Even if some studies are detected as outliers, it does not mean they are of low quality. Clinical heterogeneity might explain the differences between the detected outliers and the remaining studies. For example, a meta-analysis might contain seven studies primarily with younger populations and three studies primarily with older populations, while effect sizes may depend greatly on population ages. The latter three studies might be detected as outliers. In such cases, systematic reviewers should perform separate subgroup analyses based on the study-level clinical characteristics.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their thought-provoking feedback, which greatly enhanced the presentation of this work. LL was supported in part by the US National Institutes of Health/National Institute of Mental Health grant R03 MH128727 and National Institutes of Health/National Library of Medicine grant R01 LM012982. CW was supported in part by the US National Institutes of Health/National Institute on Aging grant R03 AG070669. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflict of interest.

DATA AVAILABILITY STATEMENT

Data and code for the simulation study and the case studies are available in Supporting Information to reproduce the results.

ORCID

Zhuo Meng  <https://orcid.org/0009-0007-4926-2270>

Lifeng Lin  <https://orcid.org/0000-0002-3562-9816>

Chong Wu  <https://orcid.org/0000-0002-8400-1785>

REFERENCES

1. Gurevitch J, Koricheva J, Nakagawa S, Stewart G. Meta-analysis and the science of research synthesis. *Nature*. 2018;555(7695):175-182. doi:[10.1038/nature25753](https://doi.org/10.1038/nature25753)
2. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*. 2008;336(7650):924-926. doi:[10.1136/bmj.39489.470347.AD](https://doi.org/10.1136/bmj.39489.470347.AD)
3. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ*. 2009;339:b2535. doi:[10.1136/bmj.b2535](https://doi.org/10.1136/bmj.b2535)
4. Shea BJ, Reeves BC, Wells G, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ*. 2017;358:j4008. doi:[10.1136/bmj.j4008](https://doi.org/10.1136/bmj.j4008)
5. Stroup DF, Berlin JA, Morton SC, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. *Jama*. 2000;283(15):2008-2012. doi:[10.1001/jama.283.15.2008](https://doi.org/10.1001/jama.283.15.2008)
6. Chu L, Ioannidis JPA, Egilman AC, Vasiliou V, Ross JS, Wallach JD. Variation of effects in epidemiologic studies of alcohol consumption and breast cancer risk. *Int J Epidemiol*. 2020;49(2):608-618. doi:[10.1093/ije/dyz271](https://doi.org/10.1093/ije/dyz271)
7. Hacke C, Nunan D. Discrepancies in meta-analyses answering the same clinical question were hard to explain: a meta-epidemiological study. *J Clin Epidemiol*. 2020;119:47-56. doi:[10.1016/j.jclinepi.2019.11.015](https://doi.org/10.1016/j.jclinepi.2019.11.015)
8. Ioannidis JPA. The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *Milbank Q*. 2016;94(3):485-514. doi:[10.1111/1468-0009.12210](https://doi.org/10.1111/1468-0009.12210)
9. Naudet F, Schuit E, Ioannidis JPA. Overlapping network meta-analyses on the same topic: survey of published studies. *Int J Epidemiol*. 2017;46(6):1999-2008. doi:[10.1093/ije/dyx138](https://doi.org/10.1093/ije/dyx138)

10. Palpacuer C, Hammas K, Duprez R, Laviolle B, Ioannidis JPA, Naudet F. Vibration of effects from diverse inclusion/exclusion criteria and analytical choices: 9216 different ways to perform an indirect comparison meta-analysis. *BMC Med*. 2019;17(1):174. doi:[10.1186/s12916-019-1409-3](https://doi.org/10.1186/s12916-019-1409-3)
11. Schöttker B, Jorde R, Peasey A, et al. Vitamin D and mortality: meta-analysis of individual participant data from a large consortium of cohort studies from Europe and the United States. *BMJ*. 2014;348:g3656. doi:[10.1136/bmj.g3656](https://doi.org/10.1136/bmj.g3656)
12. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ*. 2003;327(7414):557-560. doi:[10.1136/bmj.327.7414.557](https://doi.org/10.1136/bmj.327.7414.557)
13. Sun X, Ioannidis JPA, Agoritsas T, Alba AC, Guyatt G. How to use a subgroup analysis: users' guide to the medical literature. *Jama*. 2014;311(4):405-411. doi:[10.1001/jama.2013.285063](https://doi.org/10.1001/jama.2013.285063)
14. Aune D, Saugstad OD, Henriksen T, Tonstad S. Maternal body mass index and the risk of fetal death, stillbirth, and infant death: a systematic review and meta-analysis. *Jama*. 2014;311(15):1536-1546. doi:[10.1001/jama.2014.2269](https://doi.org/10.1001/jama.2014.2269)
15. de Souza RJ, Mente A, Maroleanu A, et al. Intake of saturated and trans unsaturated fatty acids and risk of all cause mortality, cardiovascular disease, and type 2 diabetes: systematic review and meta-analysis of observational studies. *BMJ*. 2015;351:h3978. doi:[10.1136/bmj.h3978](https://doi.org/10.1136/bmj.h3978)
16. Eyding D, Lelgemann M, Grouven U, et al. Risperidone for acute treatment of major depression: systematic review and meta-analysis of published and unpublished placebo and selective serotonin reuptake inhibitor controlled trials. *BMJ*. 2010;341:c4737. doi:[10.1136/bmj.c4737](https://doi.org/10.1136/bmj.c4737)
17. Jones L, Bellis MA, Wood S, et al. Prevalence and risk of violence against children with disabilities: a systematic review and meta-analysis of observational studies. *Lancet*. 2012;380(9845):899-907. doi:[10.1016/S0140-6736\(12\)60692-8](https://doi.org/10.1016/S0140-6736(12)60692-8)
18. Nelson JP, Kennedy PE. The use (and abuse) of meta-analysis in environmental and natural resource economics: an assessment. *Environ Resource Econ*. 2009;42(3):345-377. doi:[10.1007/s10640-008-9253-5](https://doi.org/10.1007/s10640-008-9253-5)
19. Hartwig FP, Davey Smith G, Schmidt AF, Sterne JAC, Higgins JPT, Bowden J. The median and the mode as robust meta-analysis estimators in the presence of small-study effects and outliers. *Res Synth Methods*. 2020;11(3):397-412. doi:[10.1002/jrsm.1402](https://doi.org/10.1002/jrsm.1402)
20. 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):156-166. doi:[10.1111/biom.12543](https://doi.org/10.1111/biom.12543)
21. Petitti DB. Approaches to heterogeneity in meta-analysis. *Stat Med*. 2001;20(23):3625-3633. doi:[10.1002/sim.1091](https://doi.org/10.1002/sim.1091)
22. Shi L, Lin L. The trim-and-fill method for publication bias: practical guidelines and recommendations based on a large database of meta-analyses. *Medicine (Baltimore)*. 2019;98(23):e15987. doi:[10.1097/MD.00000000000015987](https://doi.org/10.1097/MD.00000000000015987)
23. Desai K, Carroll I, Asch S, Hernandez-Boussard T, Ioannidis JPA. Extremely large outlier treatment effects may be a footprint of bias in trials from less developed countries: randomized trials of gabapentinoids. *J Clin Epidemiol*. 2019;106:80-87. doi:[10.1016/j.jclinepi.2018.10.012](https://doi.org/10.1016/j.jclinepi.2018.10.012)
24. Panagiotou OA, Contopoulos-Ioannidis DG, Ioannidis JPA. Comparative effect sizes in randomised trials from less developed and more developed countries: meta-epidemiological assessment. *BMJ*. 2013;346:f707. doi:[10.1136/bmj.f707](https://doi.org/10.1136/bmj.f707)
25. Pereira TV, Horwitz RI, Ioannidis JPA. Empirical evaluation of very large treatment effects of medical interventions. *JAMA*. 2012;308(16):1676-1684. doi:[10.1001/jama.2012.13444](https://doi.org/10.1001/jama.2012.13444)
26. Meng Z, Wu C, Lin L. The effect direction should be taken into account when assessing small-study effects. *J EvidBased Dent Pract*. 2023;23(1):101830. doi:[10.1016/j.jebdp.2022.101830](https://doi.org/10.1016/j.jebdp.2022.101830)
27. Deeks JJ, Higgins JP, Altman DG. Analysing Data and Undertaking Meta-Analyses. *Cochrane Handbook for Systematic Reviews of Interventions*. Chichester, UK: John Wiley & Sons, Ltd; 2008:243-296. doi:[10.1002/9780470712184.ch9](https://doi.org/10.1002/9780470712184.ch9)
28. Baker R, Jackson D. A new approach to outliers in meta-analysis. *Health Care Manag Sci*. 2008;11(2):121-131. doi:[10.1007/s10729-007-9041-8](https://doi.org/10.1007/s10729-007-9041-8)
29. Baker R, Jackson D. New models for describing outliers in meta-analysis. *Res Synth Methods*. 2016;7(3):314-328. doi:[10.1002/jrsm.1191](https://doi.org/10.1002/jrsm.1191)
30. Beath KJ. A finite mixture method for outlier detection and robustness in meta-analysis. *Res Synth Methods*. 2014;5(4):285-293. doi:[10.1002/jrsm.1114](https://doi.org/10.1002/jrsm.1114)
31. Gumedze FN, Jackson D. A random effects variance shift model for detecting and accommodating outliers in meta-analysis. *BMC Med Res Methodol*. 2011;11(1):19. doi:[10.1186/1471-2288-11-19](https://doi.org/10.1186/1471-2288-11-19)
32. Huffcutt AI, Arthur W. Development of a new outlier statistic for meta-analytic data. *J Appl Psychol*. 1995;80(2):327-334. doi:[10.1037/0021-9010.80.2.327](https://doi.org/10.1037/0021-9010.80.2.327)
33. Matsushima Y, Noma H, Yamada T, Furukawa TA. Influence diagnostics and outlier detection for meta-analysis of diagnostic test accuracy. *Res Synth Methods*. 2020;11(2):237-247. doi:[10.1002/jrsm.1387](https://doi.org/10.1002/jrsm.1387)
34. Noma H, Goshio M, Ishii R, Oba K, Furukawa TA. Outlier detection and influence diagnostics in network meta-analysis. *Res Synth Methods*. 2020;11(6):891-902. doi:[10.1002/jrsm.1455](https://doi.org/10.1002/jrsm.1455)
35. Zhang J, Fu H, Carlin BP. Detecting outlying trials in network meta-analysis. *Stat Med*. 2015;34(19):2695-2707. doi:[10.1002/sim.6509](https://doi.org/10.1002/sim.6509)
36. Hedges LV, Olkin I. *Statistical Methods for Meta-Analysis*. Orlando, FL: Academic Press; 2014.
37. Viechtbauer W, Cheung MWL. Outlier and influence diagnostics for meta-analysis. *Res Synth Methods*. 2010;1(2):112-125. doi:[10.1002/jrsm.11](https://doi.org/10.1002/jrsm.11)
38. Zhu X, Li X, Xu R, Wang T. An iterative approach to detect pleiotropy and perform mendelian randomization analysis using GWAS summary statistics. *Bioinformatics*. 2021;37(10):1390-1400. doi:[10.1093/bioinformatics/btaa985](https://doi.org/10.1093/bioinformatics/btaa985)
39. Efron B. Estimation and accuracy after model selection. *J Am Stat Assoc*. 2014;109(507):991-1007. doi:[10.1080/01621459.2013.823775](https://doi.org/10.1080/01621459.2013.823775)
40. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res Synth Methods*. 2010;1(2):97-111. doi:[10.1002/jrsm.12](https://doi.org/10.1002/jrsm.12)
41. 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? *J Stat Plan Inference*. 2010;140(4):961-970. doi:[10.1016/j.jspi.2009.09.017](https://doi.org/10.1016/j.jspi.2009.09.017)

42. Langan D, Higgins JPT, Jackson D, et al. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Res Synth Methods*. 2019;10(1):83-98. doi:[10.1002/jrsm.1316](https://doi.org/10.1002/jrsm.1316)
43. Veroniki AA, Jackson D, Viechtbauer W, et al. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Res Synth Methods*. 2016;7(1):55-79. doi:[10.1002/jrsm.1164](https://doi.org/10.1002/jrsm.1164)
44. Elenberg ER, Khanna R, Dimakis AG, Negahban S. Restricted strong convexity implies weak submodularity. *Ann Stat*. 2018;46(6B):3539-3568.
45. Breiman L. Bagging predictors. *Mach Learn*. 1996;24(2):123-140. doi:[10.1007/BF00058655](https://doi.org/10.1007/BF00058655)
46. Begg CB, Berlin JA. Publication bias: a problem in interpreting medical data. *J R Stat Soc A Stat Soc*. 1998;151(3):419-463.
47. Dickersin K. The existence of publication bias and risk factors for its occurrence. *Jama*. 1990;263(10):1385-1389. doi:[10.1001/jama.1990.03440100097014](https://doi.org/10.1001/jama.1990.03440100097014)
48. Fajardo-Bernal L, Aponte-Gonzalez J, Vigil P, et al. Home-based versus clinic-based specimen collection in the management of chlamydia trachomatis and Neisseria gonorrhoeae infections. *Cochrane Database Syst Rev*. 2015;2015:9. doi:[10.1002/14651858.CD011317.pub2](https://doi.org/10.1002/14651858.CD011317.pub2)
49. Jakobsen JC, Nielsen EE, Feinberg J, et al. Direct-acting antivirals for chronic hepatitis C. *Cochrane Database Syst Rev*. 2017;2017(9). doi:[10.1002/14651858.CD012143.pub3](https://doi.org/10.1002/14651858.CD012143.pub3)
50. Olliaro PL, Mussano P. Amodiaquine for treating malaria. *Cochrane Database Syst Rev*. 2003;(2). doi:[10.1002/14651858.CD000016](https://doi.org/10.1002/14651858.CD000016)
51. Jackson D, White IR. When should meta-analysis avoid making hidden normality assumptions? *Biom J*. 2018;60(6):1040-1058. doi:[10.1002/bimj.201800071](https://doi.org/10.1002/bimj.201800071)
52. Lin L. Bias caused by sampling error in meta-analysis with small sample sizes. *PloS One*. 2018;13(9):e0204056. doi:[10.1371/journal.pone.0204056](https://doi.org/10.1371/journal.pone.0204056)
53. 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. *Stat Med*. 2018;37(7):1059-1085. doi:[10.1002/sim.7588](https://doi.org/10.1002/sim.7588)
54. Lin L, Chu H. Meta-analysis of proportions using generalized linear mixed models. *Epidemiology*. 2020;31(5):713-717. doi:[10.1097/EDE.0000000000001232](https://doi.org/10.1097/EDE.0000000000001232)
55. Simmonds MC, Higgins JP. A general framework for the use of logistic regression models in meta-analysis. *Stat Methods Med Res*. 2016;25(6):2858-2877. doi:[10.1177/0962280214534409](https://doi.org/10.1177/0962280214534409)
56. Xu C, Furuya-Kanamori L, Lin L. Synthesis of evidence from zero-events studies: a comparison of one-stage framework methods. *Res Synth Methods*. 2022;13(2):176-189. doi:[10.1002/jrsm.1521](https://doi.org/10.1002/jrsm.1521)

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Meng Z, Wang J, Lin L, Wu C. Sensitivity analysis with iterative outlier detection for systematic reviews and meta-analyses. *Statistics in Medicine*. 2024;43(8):1549-1563. doi: [10.1002/sim.10008](https://doi.org/10.1002/sim.10008)