BMC Medical Research Methodology

RESEARCH

Open Access

Alternative tests and measures for between-study inconsistency in meta-analysis

Zhiyuan Yu[1†], Mengli Xiao[2†], Xing Xing[3] and Lifeng Lin[4*]

Abstract

Meta-analysis is a widely used method for synthesizing results from multiple studies across diverse fields. A central challenge in meta-analysis is assessing between-study inconsistency, which can arise from differences in study populations, methodological heterogeneity, or the presence of outliers. Conventional tools such as the Q and I^2 statistics could be limited in power, especially when the number of studies is small or when the between-study distribution deviates from normality. To address these limitations, we propose a family of alternative Q-like statistics and a hybrid test that adaptively combines their strengths. We also introduce new measures to quantify inconsistency based on these statistics. Simulation studies demonstrate that the hybrid test performs robustly across a wide range of inconsistency patterns, including heavy-tailed, skewed, and contaminated distributions. We further illustrate the practical utility of our methods using three real-world meta-analyses. These approaches offer more flexible and powerful tools for detecting and quantifying inconsistency in meta-analytic practice.

Keywords Heterogeneity, Hybrid test, Inconsistency, Meta-analysis, Resampling, Statistical power

Background

Meta-analysis is widely used to combine results from multiple studies on the same research topic in a wide range of fields [1, 2]. A critical step for validating a meta-analysis is to assess the differences between studies. Because the multiple studies recruited participants with potentially different characteristics and were conducted by different research teams with different methods, their results are usually expected to have a certain extent of heterogeneity [3]. If the studies are considered mostly homogeneous, such that they share a common true effect size, a common-effect model is employed for the meta-analysis. On the other hand, if the studies are heterogeneous, such that their underlying true effect sizes differ, a random-effects model is typically used to account for the heterogeneity [4, 5]. As such, the assessment of the differences between studies offers valuable information for model selection and appraising the interpretability of meta-analysis conclusions.

The Q test is traditionally used to measure the between-study heterogeneity in a meta-analysis. Under the null hypothesis of homogeneity, the Q statistic follows a chi-squared distribution. This statistic takes the sum of squared standardized deviates from individual studies. Conceptually, it shares a similar idea with the classical least-squares regression that minimizes the sum of squared errors. The structure of the sum of squares may be suitable for the conventional assumption made by the random-effects meta-analysis models; that is, the

†Zhiyuan Yu and Mengli Xiao contributed equally.

*Correspondence:
Lifeng Lin
lifenglin@arizona.edu
[1]Citibank, Tampa, FL, USA
[2]Department of Biostatistics and Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO, USA
[3]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
[4]Department of Epidemiology and Biostatistics, University of Arizona, 1295 N. Martin Ave., Tucson, AZ 85724, USA

underlying true effect sizes of all individual studies follow a normal distribution with a certain heterogeneity variance. The theoretical properties of the Q statistic have been mostly investigated under such between-study normality [6–8]. Nevertheless, this is not always the case of discrepancies between studies' results, and the normality assumption could be questionable in many applications [9–12]. For example, a meta-analysis may include a few outlying studies whose results take extreme values compared with the majority of studies [13]. It may also inappropriately pool multiple subgroups in the same analysis; the subgroups could be distinguished based on certain study-level summaries of population characteristics, such as average ages [14]. The studies in different subgroups could have dramatically different effect sizes, while those in the same group may share more homogeneous effect sizes. In such cases, the underlying effect sizes from the whole studies may have a distribution with multiple modes. Moreover, publication bias or small-study effects can distort the observed distribution of studies. For example, if studies reporting effects in one direction (say, positive effects of an intervention) are more likely to be published, the between-study distribution may become skewed, often with a longer tail on the side favored by the bias. It is well documented in the meta-analysis literature that publication bias can interact with heterogeneity assessments, potentially exaggerating between-study variability or even masking true inconsistency [15]. Indeed, ignoring the coexistence of publication bias and heterogeneity may lead to misleading conclusions about both [16, 17]. In summary, the conventional Q test may not be suitable for the foregoing non-normal cases, as it may suffer from low statistical power, especially when the number of studies in a meta-analysis is relatively small [18, 19].

Considering the diversity of factors that may cause the discrepancies between studies on the same research topic, this article refers to such discrepancies as inconsistency in general instead of heterogeneity in particular. The term "inconsistency" was also used by Higgins et al. [20] when they introduced the famous I^2 statistic to the medical community for quantifying the discrepancies between studies. Moreover, this terminology is adopted in the GRADE (Grading of Recommendations Assessment, Development and Evaluation) framework for assessing the certainty of evidence [21]. In the literature of meta-analysis methodology, on the other hand, the term "heterogeneity" seems to be more widely used for the same purpose. It is frequently paired with the random-effects model, which assumes between-study normality; that is, the discrepancies permeate the entire meta-analysis, rather than being limited to certain subgroups or a few outlying studies. This article uses the term "inconsistency" to inclusively cover all types of discrepancies, including subgroup effects and potential outliers. It is important to note that this usage differs from evidence inconsistency in network meta-analysis, which refers to disagreement between direct and indirect comparisons of multiple interventions [22, 23].

Beyond the traditional Q test, several likelihood-based approaches, including the score test, likelihood-ratio test, and Wald test, have been proposed for assessing between-study heterogeneity [24, 25]. Similar to the Q test, these methods are generally derived under the assumption of normally distributed random effects. However, they may suffer from low power or lack robustness when the underlying random-effects distribution is skewed, heavy-tailed, or contaminated. In addition, some of these likelihood-based methods are tailored to specific effect measures, such as odds ratios [26], which can limit their general applicability across different types of meta-analyses.

Motivated by these limitations, this article presents alternative test statistics to examine the between-study inconsistency. These statistics are designed based on the sum of absolute values of standardized deviates with different mathematical powers (e.g., square, cubic, maximum, and so on). They attempt to serve as suitable candidates under various scenarios of between-study distributions. For example, consider a meta-analysis with an extremely outlying study; except for the outliers, all remaining studies are mostly homogeneous. If the conventional Q statistic is used to test for the inconsistency, the sum of squares of standardized deviates would include too much noise, given that most studies are actually homogeneous. Indeed, in such a case, the maximum of the standardized deviates would efficiently capture the inconsistency with the minimal contamination by noises from homogeneous studies, as the outlying study is expected to create the largest deviate.

By their designs, the various alternative tests for inconsistency have different statistical power under different settings that cause discrepancies between individual studies' results, so there is no universally best test. In practice, it is infeasible to justify an optimal test that fits a specific meta-analysis dataset. As such, using the idea of adaptive testing [27, 28], we derive a hybrid test based on the various tests for inconsistency. The hybrid test statistic takes the minimum P-values from various tests for inconsistency so that it can achieve relatively high power across a wide range of settings. To properly control the type I error rate of the hybrid test, we propose a parametric resampling procedure to derive the null distribution and thus calculate the empirical P-value of the hybrid test.

This article is organized as follows. We start with presenting the setup of the inconsistency problem in a meta-analysis, reviewing the existing popular Q test for

inconsistency, proposing the alternative test statistics and hybrid test, and providing the algorithm of the resampling method for deriving the tests' *P*-values. Then, we present simulation studies to compare the statistical power of the various tests and use three case studies to illustrate the real-world performance of the tests. Finally, we conclude this article with discussions about the proposed methods' limitations and potential future directions.

Methods

The conventional method for testing for inconsistency

Suppose that a meta-analysis includes a total of k independent studies. Let y_i be the observed effect size in study i (i = 1, 2, ..., k) and s_i be its standard error. Meta-analysis models typically assume that each y_i is approximately normally distributed with mean μ_i and variance s_i^2. Also, the observed standard errors s_i's are conventionally treated as fixed variables, as if there is no error in their estimation. These assumptions are generally valid if the sample sizes in the studies are sufficiently large (e.g., due to the central limit theorem and law of large numbers), but extra cautions are needed for small sample sizes [10, 29].

This article focuses on testing for the potential inconsistency between the k studies, so the distribution of the study-specific underlying true effect sizes μ_i plays a critical role. The null hypothesis is that all studies are homogeneous, sharing a common effect size μ; that is, $\mu_1 = \mu_2 = \cdots = \mu_k = \mu$, and their distribution is a mass at μ. In such cases, the common-effect model is used. If the homogeneity does not appear to hold, meta-analysts conventionally use the random-effects model to account for the between-study inconsistency. This model typically assumes that μ_i's are random effects, following a normal distribution $N\left(\mu, \tau^2\right)$, where μ represents the overall mean effect size in the random-effects framework and τ^2 is the between-study variance [4]. Although alternative distributions are possible for the random effects [30, 31], the between-study normality assumption has dominated the current literature of meta-analyses owing to its simplicity.

The Q test is the standard approach to examining inconsistency. Its statistic is defined as

$$Q = \sum\nolimits_{i=1}^{k} \left(y_i - \widehat{\mu}_{\mathrm{CE}}\right)^2 / s_i^2,$$

Where the common-effect estimate is

$$\widehat{\mu}_{\mathrm{CE}} = \frac{\sum_{i=1}^{k} y_i / s_i^2}{\sum_{i=1}^{k} 1 / s_i^2}. \tag{1}$$

The Q statistic follows a χ^2_{k-1} distribution under the null hypothesis of homogeneity.

Alternative test statistics

It is straightforward to rewrite the conventional Q statistic as $Q = \sum_{i=1}^{k} d_i^2$, where $d_i = \left(y_i - \widehat{\mu}_{\mathrm{CE}}\right) / s_i$ may be interpreted as study-specific standardized deviates. As illustrated in the introduction section, this sum-of-squares structure is arguably suitable for the between-study normality assumption, but it may not work well for other between-study distributions. Consider the case that the meta-analysis contains an outlying study (denoting its index by o) and all remaining studies are homogeneous. We would expect d_o to take an extremely large value in absolute magnitude, while the other studies' standardized deviates d_j ($j \neq o$) are small. The between-study inconsistency can be mostly captured by d_o, and the d_j's ($j \neq o$) may add nuisance information to the sum of squares in Q. Therefore, it is sensible to test for the inconsistency based solely on d_o, i.e., the maximum value among the absolute values of all standardized deviates $|d_i|$'s.

In addition, we may consider taking the sum of the $|d_i|$'s with different mathematical powers γ to reflect different weights contributed by individual studies. For example, when γ = 1, the sum becomes $\sum_{i=1}^{k} |d_i|$, which could reduce the impact of potential outliers and make the assessment of inconsistency more robust [9]. When γ increases, the contributions of larger deviates to the sum become larger, while those of smaller deviates become smaller. If γ approaches infinity, then only the largest deviate would dominate the sum, so the sum effectively plays a similar role to the maximum of all deviates. Different values of γ could capture different patterns of the between-study distributions.

Formally, we propose the following alternative statistic for an integer value of the mathematical power γ:

$$Q_\gamma = \sum\nolimits_{i=1}^{k} \left(\left|y_i - \widehat{\mu}_{\mathrm{CE}}\right| / s_i\right)^\gamma \tag{2}$$

If γ = 2, the Q_γ becomes the conventional Q statistic. For other values of γ, the Q_γ statistic could capture different patterns of between-study inconsistency and thus be more powerful than Q. Also, because the largest deviate in absolute magnitude would dominate the sum as γ approaches infinity, we define the statistic for $\gamma = \infty$ as

$$Q_\infty = \max_{i=1,\ldots,k} \left|y_i - \widehat{\mu}_{\mathrm{CE}}\right| / s_i \tag{3}$$

This article considers the values of 1, 2, ..., 8, and ∞ for γ. Based on our empirical experiments, these values are

sufficient to capture various patterns of between-study inconsistency. We denote the *P*-value of the Q_γ statistic by P_γ.

In practice, however, it is infeasible to justify the optimal Q_γ for a specific meta-analysis because identifying the between-study distribution is challenging. Borrowing the idea of adaptive testing [27, 28, 32], we propose a hybrid test for the between-study inconsistency. Specifically, we first consider a set of candidate tests for inconsistency, say Q_γ with $\gamma \in \Gamma = \{1, 2, \ldots, 8, \infty\}$. Then, the hybrid test statistic is defined as the minimum *P*-value of all candidate tests; that is,

$$Q_{\text{hyb}} = \min_{\gamma\in\Gamma} P_\gamma \tag{4}$$

As the minimum among the *P*-values of a pool of tests, the Q_{hyb} is no longer a valid *P*-value because it cannot control the type I error rate. Indeed, we treat Q_{hyb} as a test statistic rather than a *P*-value. The following subsection proposes a resampling method to derive the null distribution of the hybrid test statistic and thus calculate its *P*-value P_{hyb}. Because it is also difficult to derive the theoretical null distribution of Q_{hyb} (except for γ = 2, which leads to the conventional Q test statistic), the resampling method is used to derive the *P*-values of Q_γ's as well.

Calculation of the *P*-values of alternative tests

The resampling method for calculating the *P*-values of the proposed tests is as follows. First, under the null hypothesis of homogeneity, the common effect size $\widehat{\mu}_{\text{CE}}$ is estimated as in Eq. (1), the test statistic Q_γ for each $\gamma \in \Gamma$ is obtained using Eqs. (2) or (3), and the hybrid test statistic Q_{hyb} is obtained using Eq. (4). Second, we generate resampled replicates of the meta-analysis for B times (say, B = 1,000). Ideally, B should be as large as computational resources allow to minimize Monte Carlo error. For each resampling iteration b (b = 1, 2, ..., B), we draw study-specific standard errors $s_i^{(b)}$ with replacement from the original standard errors s_i (i = 1, 2, ..., k), and the effect size estimates are obtained as $y_i^{(b)} \sim N\left(\widehat{\mu}_{\text{CE}}, \left(s_i^{(b)}\right)^2\right)$. For the b th resampled meta-analysis, we calculate $Q_\gamma^{(b)}$ for each $\gamma \in \Gamma$ using $y_i^{(b)}$ and $s_i^{(b)}$. These resampled test statistics form null distributions for each Q_γ; thus, the *P*-value of Q_γ can be calculated as

$$P_\gamma = \frac{\sum_{b=1}^{B} I\left(Q_\gamma^{(b)} \geq Q_\gamma\right) + 1}{B+1} \tag{5}$$

where $I(\cdot)$ is the indicator function. A constant 1 is added to both the denominator and the numerator to avoid the *P*-value being calculated as 0.

Of note, the *P*-values of Q_γ based on Eq. (5) are obtained for the original meta-analyses. To obtain the *P*-value of the hybrid test, we need to calculate the hybrid test statistic for each resampled meta-analysis, which depends on the *P*-values of Q_γ for the resampled meta-analysis. For the b th resampled meta-analysis with the test statistic $Q_\gamma^{(b)}$, the statistics $Q_\gamma^{(c)}$ in other resampled meta-analyses (c = 1, 2, ..., k but $c \neq b$) can serve as an empirical distribution for $Q_\gamma^{(b)}$. Thus, the *P*-value of $Q_\gamma^{(b)}$ can be calculated as

$$P_\gamma^{(b)} = \frac{\sum_{c\neq b} I\left(Q_\gamma^{(c)} \geq Q_\gamma^{(b)}\right) + 1}{B}$$

With these *P*-values, for the b th resampled meta-analysis, its hybrid test statistic is

$$Q_{\text{hyb}}^{(b)} = \min_{\gamma\in\Gamma} P_\gamma^{(b)}$$

Finally, based on the hybrid test statistics of the resampled meta-analyses under the null hypothesis, the *P*-value of Q_{hyb} is

$$P_{\text{hyb}} = \frac{\sum_{b=1}^{B} I\left(Q_{\text{hyb}}^{(b)} \leq Q_{\text{hyb}}\right) + 1}{B+1}$$

Alternative measures for quantifying inconsistency

In addition to testing for inconsistency, it is also of great interest to quantify it [33]. Like the conventional Q statistic, the magnitudes of the proposed Q_γ statistics depend on the number of studies k, so Q_γ cannot be directly used as measures of between-study inconsistency in different meta-analyses. Motivated by the popular I^2 statistic [20, 34], we extend the Q_γ statistics to derive alternative inconsistency measures.

Specifically, the I^2 statistic can be calculated as

$$I^2 = \max\left\{\frac{Q-(k-1)}{Q}, 0\right\} \times 100\% \tag{6}$$

where Q is the conventional Q statistic. It is interpreted as the percentage of the total variation in study estimates due to the between-study inconsistency rather than the sampling error. Thus, conceptually, the I^2 statistic can be considered as a form of

$$\frac{\tau^2}{\tau^2+\sigma^2} \tag{7}$$

where τ^2 represents the between-study variance and σ^2 is a summary of all sample variances from individual

studies [34]. The Cochrane Handbook gives a rough guide for interpreting the I^2 statistic as unimportant, moderate, or substantial inconsistency [2].

In the framework of this article, however, we do not pursue new measures with a similar interpretation as in Eq. (7) The marginal variance $\tau^2 + \sigma^2$ may be intuitive under the between-study normality assumption, but it may not be meaningful in general settings of inconsistency, e.g., the existence of outlying studies. Instead, we are motivated to construct new inconsistency measures by taking another look at the formula of I^2 in Eq. (6). Because $k-1$ in the numerator is the expectation of the Q statistic under the hypothesis, $Q-(k-1)$ describes the excess of the observed Q statistic compared with its null expectation, and thus I^2 can also be viewed as the percentage of excess inconsistency. This interpretation shares a similar idea with the concept of excess statistical significance used for assessing publication bias [35, 36].

As such, based on each alternative test statistic Q_γ, we propose to quantify inconsistency by

$$E_\gamma = \max\left\{\frac{Q_\gamma - E\left[Q_\gamma|H_0\right]}{Q_\gamma}, 0\right\} \times 100\%$$

where $E\left[Q_\gamma|H_0\right]$ is the expectation of Q_γ under the null hypothesis. The E_γ measure is interpreted as the percentage excess inconsistency based on the Q_γ statistic. When $\gamma = 2$, E_2 is identical to the I^2 statistic. In addition, like I^2 [34], the E_γ measure is scale-invariant because Q_γ is unit-free.

We can derive the theoretical null expectations of Q_γ for $\gamma < \infty$. Recall that $Q_\gamma = \sum_{i=1}^k (|y_i - \widehat{\mu}_{\mathrm{CE}}|/s_i)^\gamma$, where $\widehat{\mu}_{\mathrm{CE}} = \left[\sum_{i=1}^k y_i/s_i^2\right] / \left[\sum_{i=1}^k 1/s_i^2\right]$. Because y_i follows normal distribution, so $|y_i - \widehat{\mu}_{\mathrm{CE}}|/s_i$ follows a folded normal distribution with mean 0 if s_i's are treated as fixed values. By this property, we can derive the following formula of $E\left[Q_\gamma|H_0\right]$ using the γ th moment of the folded normal distribution:

$$E\left[Q_\gamma|H_0\right] = \frac{2^{\gamma/2}}{\sqrt{\pi}}\Gamma\left(\frac{\gamma+1}{2}\right)\sum_{i=1}^k \sigma_i^\gamma$$

where $\Gamma(\cdot)$ denotes the Gamma function and $\sigma_i = \sqrt{1 - 1/\left[s_i^2 \sum_{j=1}^k 1/s_j^2\right]}$. The detailed proof is given in Additional File 1. Here, we list the equations of $E\left[Q_\gamma|H_0\right]$ with γ = 1, 2, ..., and 8 (which are used in our numerical studies):

$$E\left[Q_1|H_0\right] = \sqrt{\tfrac{2}{\pi}}\sum_{i=1}^k \sigma_i;$$
$$E\left[Q_2|H_0\right] = \sum_{i=1}^k \sigma_i^2 = k-1;$$
$$E\left[Q_3|H_0\right] = 2\sqrt{\tfrac{2}{\pi}}\sum_{i=1}^k \sigma_i^3;$$
$$E\left[Q_4|H_0\right] = 3\sum_{i=1}^k \sigma_i^4;$$
$$E\left[Q_5|H_0\right] = 8\sqrt{\tfrac{2}{\pi}}\sum_{i=1}^k \sigma_i^5;$$
$$E\left[Q_6|H_0\right] = 15\sum_{i=1}^k \sigma_i^6;$$
$$E\left[Q_7|H_0\right] = 48\sqrt{\tfrac{2}{\pi}}\sum_{i=1}^k \sigma_i^7; \text{ and}$$
$$E\left[Q_8|H_0\right] = 105\sum_{i=1}^k \sigma_i^8.$$

It is infeasible to obtain the explicit equations of $E\left[Q_\infty|H_0\right]$ and $E\left[Q_{\mathrm{hyb}}|H_0\right]$. Nevertheless, the resampling method introduced in the previous subsection can also be used to calculate the null expectation based on the resampled meta-analyses; that is,

$$E\left[Q_\gamma|H_0\right] \approx \frac{1}{B}\sum_{b=1}^B Q_\gamma^{(b)}$$

This approximation can be readily used for obtaining $E\left[Q_\infty|H_0\right]$ with $\gamma = \infty$.

For the hybrid test statistic, it is not straightforward to derive $E\left[Q_{\mathrm{hyb}}|H_0\right]$ in a similar way for two reasons. First, a large value of Q_γ implies a large inconsistency between studies. On the contrary, a large value of Q_{hyb} implies a small inconsistency between studies. Second, Q_{hyb} only ranges from 0 to 1; therefore, it is difficult to compare its magnitudes across meta-analyses when Q_{hyb} becomes small, say, 10^{-3} and 10^{-4}. To solve these problems, we apply a log transformation with base 10 to Q_{hyb} and impose a negative sign; the transformed statistic is $L_{\mathrm{hyb}} = -\log_{10}\left(Q_{\mathrm{hyb}}\right)$. After this transformation, we can use the foregoing resampling method for L_{hyb} to approximate the $E\left[L_{\mathrm{hyb}}|H_0\right]$:

$$E\left[L_{\mathrm{hyb}}|H_0\right] \approx -\frac{1}{B}\sum_{b=1}^B \log_{10}\left(Q_{\mathrm{hyb}}^{(b)}\right)$$

Thus, the measure of inconsistency based on the hybrid statistic is calculated as

$$E_{\mathrm{hyb}} = \max\left\{\frac{L_{\mathrm{hyb}} - E\left[L_{\mathrm{hyb}}|H_0\right]}{L_{\mathrm{hyb}}}, 0\right\} \times 100\%$$

Similar to the interpretation of E_γ, E_{hyb} represents the percentage of excess inconsistency, but it is derived from the Q_{hyb} (or equivalently, L_{hyb}) statistic. For example, if Q_{hyb} = 0.001, corresponding to L_{hyb} = 3, and the expected value under the null hypothesis is $E\left[L_{\mathrm{hyb}}|H_0\right]$ = 1, then E_{hyb} = (3–1)/3 = 66.7%.

Simulation studies

We validated the performance of various tests for between-study inconsistency via simulation studies. We considered the conventional Q test, the proposed Q_γ

tests with γ = 1, 2, …, 8, and ∞, as well as the hybrid test. The tests' performance was assessed in terms of their type I error rates and statistical power.

To derive type I error rates, we generated meta-analyses with observed effect sizes as $y_i \sim N\left(\mu_{\mathrm{CE}}, s_i^2\right)$ under the null hypothesis of homogeneity (Case 0). Without loss of generality, we set μ_{CE} = 0. In addition, the within-study standard errors s_i were sampled from a uniform distribution, $U\left(1,2\right)$ or $U\left(2,3\right)$.

To derive statistical power, we considered various settings of the alternative hypothesis of between-study inconsistency. Specifically, the study-specific effect sizes y_i were generated from $N\left(\mu_i, s_i^2\right)$, where μ_i was the true effect size of study i and followed various distributions as follows to reflect different patterns of between-study inconsistency.

- Case 1: $N\left(0,2\right)$;
- Case 2: the mixture normal distribution, $0.85N\left(0,0.2\right)+0.15N\left(0,20\right)$, consisting of two normal distributions with the same mean 0 and different variances;
- Case 3: the gamma distribution with shape parameter 0.05 and rate parameter 0.1, $\mathrm{Gamma}\left(0.05,0.1\right)$;
- Case 4: the F distribution with degrees of freedom 3 and 8, $F_{3,8}$.
- Case 5: the mixture normal distribution, $0.70N\left(0,1\right)+0.30N\left(1.5,1\right)$, consisting of two normal distributions with different means but the same variance;

The normality assumption in Case 1 is conventional for most meta-analysis methods. The distribution in Case 2 had heavier tails than a single normal distribution, which could generate extreme values. The distributions in Cases 3 and 4 were skewed, arguably reflecting the potential influence of publication bias or small-study effects. In Case 5, the distribution conceptually represents scenarios in which a meta-analysis includes two subgroups, each centered at a different overall effect size. The distributions in Cases 3 to 5 have non-zero mean; to make fair comparisons, we centralized these distributions at 0 so that $E\left[\mu_i\right]$ = 0.

We additionally considered two cases of contamination:

- Case 6: all μ_i = − 0.2 expect that $\mu_{11}=0.2\left(k-1\right)$;
- Case 7: all μ_i = 0 but y_6 was artificially added by a discrepancy value of ϵ = 3.

The cases occur when most studies in a meta-analysis are homogeneous, but one study based on a dramatically different population is inappropriately included in the meta-analysis.

Each simulated meta-analysis consisted of k = 15 or 30 studies. We generated 1,000 replicates for each simulation setting. Because the resampling algorithm is computationally intensive, and considering that it needed to be repeated for 1,000 Monte Carlo replications in the simulations, we used B = 500 resampling iterations instead of a much larger number. The significance level for inconsistency was set to 0.10 because inconsistency tests typically have low power in many cases, as indicated by existing simulation studies [37].

Case studies

To illustrate the practical utility of the proposed methods, we applied the Q_γ tests and the hybrid tests to three real-world meta-analyses. For each meta-analysis, we calculated the *P*-values of the various tests and the corresponding inconsistency measures. The resampling method was used for calculating the *P*-values, and the number of resampling iterations was 10,000. Of note, we used a much larger B in these case studies than in the simulations because only three meta-analyses were analyzed, making the larger number of resampling iterations computationally feasible and allowing for smaller Monte Carlo resampling error.

Results

Results of simulation studies

Tables 1, 2, 3, and 4 present the simulation results under a range of data-generating settings that varied the distribution of the true effects μ_i, the within-study standard errors $s_i \sim U\left(1,2\right)$ or $s_i \sim U\left(2,3\right)$, and the number of studies k = 15 or 30. Monte Carlo standard errors were mostly around 2%, ensuring reliable comparisons of type I error rates or statistical power across settings. The settings covered symmetric, skewed, heavy-tailed, and contaminated distributions, offering a comprehensive assessment of test performance under diverse conditions.

Type I error rate (Case 0)

Under the null hypothesis (Case 0: μ_i = 0), all tests controlled the type I error rates well at the nominal 10% level. For example, in Table 1 with $s_i \sim U\left(1,2\right)$ and k = 15, the type I error rates ranged from 8.6% to 9.4%; in Table 2 with $s_i \sim U\left(2,3\right)$ and k = 15, the rates had similar ranges (e.g., 8.5% for Q_1, 9.4% for Q_∞). With k = 30, the type I error rates remained well-controlled (Tables 3 and 4).

Normal distribution (Case 1)

Under the conventional normality assumption for between-study inconsistency, with $\mu_i \sim N\left(0,2\right)$, the statistical power of all tests increased substantially with the number of studies. For example, in the setting $s_i \sim U\left(1,2\right)$ (Tables 1 and 3), the hybrid test Q_{hyb} achieved a power of 69.0% when k = 15, which

Table 1 Type I error rates (Case 0) and statistical power (Cases 1–7), expressed as percentages, for various tests under the setting with within-study standard errors $s_i \sim U(1, 2)$ and the number of studies $k = 15$. Monte Carlo standard errors are shown in parentheses. The significance level was set at 10%

Setting	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_∞	Q_{hyb}
Case 0	8.6 (0.9)	8.8 (0.9)	9.1 (0.9)	8.9 (0.9)	9.0 (0.9)	9.2 (0.9)	9.3 (0.9)	9.4 (0.9)	9.3 (0.9)	9.1 (0.9)
Case 1	68.2 (1.5)	71.3 (1.4)	69.9 (1.5)	67.0 (1.5)	65.3 (1.5)	64.0 (1.5)	62.9 (1.6)	62.0 (1.6)	56.2 (1.6)	69.0 (1.5)
Case 2	58.7 (1.6)	65.4 (1.5)	66.8 (1.5)	66.6 (1.5)	66.2 (1.5)	65.9 (1.5)	65.5 (1.5)	65.6 (1.5)	65.2 (1.5)	66.4 (1.5)
Case 3	42.2 (1.6)	47.3 (1.6)	49.1 (1.6)	48.6 (1.6)	48.6 (1.6)	48.8 (1.6)	48.5 (1.6)	48.7 (1.6)	48.7 (1.6)	47.9 (1.6)
Case 4	55.5 (1.6)	61.0 (1.6)	60.1 (1.6)	58.7 (1.6)	58.0 (1.6)	57.4 (1.6)	56.7 (1.6)	56.2 (1.6)	53.8 (1.6)	59.9 (1.6)
Case 5	55.9 (1.6)	60.0 (1.5)	58.2 (1.6)	56.5 (1.6)	55.2 (1.6)	53.8 (1.6)	52.4 (1.6)	51.9 (1.6)	47.1 (1.6)	56.6 (1.6)
Case 6	24.9 (1.3)	29.7 (1.4)	33.2 (1.5)	34.2 (1.5)	34.5 (1.5)	34.6 (1.5)	34.7 (1.5)	35.0 (1.5)	34.7 (1.5)	33.0 (1.5)
Case 7	23.0 (1.3)	29.2 (1.4)	32.8 (1.5)	33.1 (1.5)	33.2 (1.5)	33.3 (1.5)	32.7 (1.5)	32.3 (1.5)	32.2 (1.5)	31.4 (1.5)

Table 2 Type I error rates (Case 0) and statistical power (Cases 1–7), expressed as percentages, for various tests under the setting with within-study standard errors $s_i \sim U(2, 3)$ and the number of studies $k = 15$. Monte Carlo standard errors are shown in parentheses. The significance level was set at 10%

Setting	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_∞	Q_{hyb}
Case 0	8.5 (0.9)	8.6 (0.9)	9.1 (0.9)	8.8 (0.9)	9.1 (0.9)	9.1 (0.9)	9.4 (0.9)	9.2 (0.9)	9.4 (0.9)	9.1 (0.9
Case 1	32.8 (1.5)	35.4 (1.5)	34.3 (1.5)	33.1 (1.5)	32.4 (1.5)	31.0 (1.5)	30.7 (1.5)	30.3 (1.5)	28.2 (1.4)	34.8 (1.5)
Case 2	36.2 (1.5)	43.5 (1.6)	43.8 (1.6)	43.3 (1.6)	43.5 (1.6)	43.5 (1.6)	43.4 (1.6)	43.0 (1.6)	41.4 (1.6)	43.2 (1.6)
Case 3	30.7 (1.5)	35.2 (1.5)	36.5 (1.6)	36.0 (1.5)	36.1 (1.5)	35.9 (1.5)	36.0 (1.5)	35.7 (1.5)	35.3 (1.5)	35.6 (1.5)
Case 4	30.0 (1.4)	35.9 (1.5)	35.9 (1.5)	35.8 (1.5)	34.4 (1.5)	34.1 (1.5)	33.7 (1.5)	33.7 (1.5)	31.9 (1.5)	35.9 (1.5)
Case 5	22.7 (1.3)	23.9 (1.3)	24.2 (1.4)	23.3 (1.3)	22.9 (1.3)	22.4 (1.3)	21.7 (1.3)	21.6 (1.3)	20.2 (1.3)	23.6 (1.3)
Case 6	14.4 (1.1)	17.2 (1.2)	17.8 (1.2)	17.8 (1.2)	18.1 (1.2)	17.8 (1.2)	17.4 (1.2)	17.4 (1.2)	17.2 (1.2)	16.6 (1.2)
Case 7	13.7 (1.0)	14.9 (1.1)	15.5 (1.1)	15.0 (1.1)	14.9 (1.1)	15.3 (1.1)	15.3 (1.1)	15.5 (1.1)	15.4 (1.1)	14.7 (1.1)

Table 3 Type I error rates (Case 0) and statistical power (Cases 1–7), expressed as percentages, for various tests under the setting with within-study standard errors $s_i \sim U(1, 2)$ and the number of studies $k = 30$. Monte Carlo standard errors are shown in parentheses. The significance level was set at 10%

Setting	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_∞	Q_{hyb}
Case 0	11.0 (1.0)	10.6 (1.0)	10.5 (1.0)	9.8 (0.9)	9.3 (0.9)	9.5 (0.9)	8.8 (0.9)	9.0 (0.9)	9.3 (0.9)	9.9 (0.9)
Case 1	86.8 (1.1)	89.4 (1.0)	88.4 (1.0)	86.6 (1.1)	84.3 (1.2)	81.4 (1.2)	79.8 (1.3)	77.7 (1.3)	68.8 (1.5)	87.4 (1.0)
Case 2	75.9 (1.3)	82.8 (1.2)	84.0 (1.2)	84.1 (1.2)	84.1 (1.2)	83.9 (1.2)	83.8 (1.2)	83.9 (1.2)	82.7 (1.2)	84.6 (1.2)
Case 3	57.5 (1.6)	65.6 (1.5)	67.6 (1.5)	67.9 (1.5)	67.8 (1.5)	67.8 (1.5)	67.7 (1.5)	67.7 (1.5)	68.0 (1.5)	67.2 (1.5)
Case 4	73.3 (1.4)	78.8 (1.3)	78.3 (1.3)	77.3 (1.3)	75.9 (1.3)	75.0 (1.3)	74.7 (1.3)	73.4 (1.4)	69.5 (1.5)	77.3 (1.3)
Case 5	75.5 (1.4)	79.4 (1.3)	77.8 (1.3)	73.8 (1.4)	70.5 (1.4)	67.3 (1.5)	65.4 (1.5)	63.3 (1.5)	54.6 (1.6)	75.5 (1.4)
Case 6	40.8 (1.6)	68.5 (1.5)	78.7 (1.3)	81.8 (1.2)	83.3 (1.2)	83.7 (1.2)	84.1 (1.1)	84.3 (1.1)	84.6 (1.1)	81.9 (1.2)
Case 7	17.9 (1.2)	22.9 (1.3)	26.6 (1.4)	27.5 (1.4)	29.2 (1.4)	28.8 (1.4)	28.8 (1.4)	28.7 (1.4)	28.4 (1.4)	27.0 (1.4)

Table 4 Type I error rates (Case 0) and statistical power (Cases 1–7), expressed as percentages, for various tests under the setting with within-study standard errors $s_i \sim U(2, 3)$ and the number of studies $k = 30$. Monte Carlo standard errors are shown in parentheses. The significance level was set at 10%

Setting	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_∞	Q_{hyb}
Case 0	11.1 (1.0)	10.5 (1.0)	10.3 (1.0)	9.6 (0.9)	9.0 (0.9)	9.5 (0.9)	8.9 (0.9)	9.0 (0.9)	9.1 (0.9)	9.8 (0.9)
Case 1	40.7 (1.6)	44.0 (1.6)	42.6 (1.6)	40.3 (1.6)	37.8 (1.6)	35.0 (1.5)	33.9 (1.5)	32.6 (1.5)	29.9 (1.4)	40.5 (1.6)
Case 2	44.8 (1.6)	55.6 (1.6)	57.0 (1.6)	57.5 (1.6)	56.9 (1.6)	56.2 (1.6)	55.5 (1.6)	55.2 (1.6)	52.8 (1.6)	57.0 (1.6)
Case 3	37.7 (1.6)	48.4 (1.6)	50.3 (1.6)	50.4 (1.6)	50.5 (1.6)	50.7 (1.6)	50.7 (1.6)	50.6 (1.6)	49.8 (1.6)	50.0 (1.6)
Case 4	39.5 (1.6)	44.9 (1.6)	46.1 (1.6)	45.5 (1.6)	44.9 (1.6)	43.7 (1.6)	43.1 (1.6)	42.6 (1.6)	38.6 (1.6)	44.9 (1.6)
Case 5	33.8 (1.5)	35.5 (1.5)	34.2 (1.5)	33.2 (1.5)	30.3 (1.5)	28.7 (1.4)	27.7 (1.4)	27.2 (1.4)	24.8 (1.4)	34.0 (1.5)
Case 6	22.2 (1.3)	30.4 (1.5)	36.3 (1.6)	38.7 (1.6)	39.7 (1.6)	40.0 (1.6)	39.8 (1.6)	40.2 (1.6)	39.6 (1.6)	38.1 (1.6)
Case 7	11.3 (1.0)	13.1 (1.0)	13.1 (1.0)	12.5 (1.0)	13.0 (1.0)	13.2 (1.0)	13.3 (1.0)	13.2 (1.0)	13.2 (1.0)	12.8 (1.0)

increased to 87.4% when k = 30. Notably, the Q_2 test demonstrated the highest power among all Q_γ tests in this scenario, reaching 71.3% for k = 15 and 89.4% for k = 30. It also slightly outperformed the hybrid test. When the within-study standard errors followed a wider distribution with $s_i \sim U(2,3)$ in Tables 2 and 4, the power of all tests declined due to increased within-study uncertainties. Nevertheless, Q_2 remained one of the best-performing methods, with power values of 35.4% (k = 15) and 44.0% (k = 30), consistently outperforming most other tests.

Heavy-tailed distribution (Case 2)

Across all settings, the hybrid test Q_{hyb} demonstrated strong robustness and consistently competitive power. For $s_i \sim U(1,2)$ and k = 15, the hybrid test achieved a power of 66.4%, which was slightly lower than the best-performing Q_γ tests, such as 66.8% for Q_3, 66.6% for Q_4, and 66.2% for Q_5. Nevertheless, it remained among the top-performing methods, and its power was notably higher than Q_1 (with power of 58.7%) and comparable to Q_2 (65.4%), as shown in Table 1. When within-study standard errors were larger, power declined for all methods. The hybrid test maintained solid performance with power of 43.2%, outperforming Q_1 (36.2%), and having similar power to Q_2 (43.5%) and Q_5 (43.5%), as shown in Table 2. With more studies (k = 30) in meta-analyses, the hybrid test showed clearer advantages. In Table 3 with $s_i \sim U(1,2)$, its power reached 84.6%, outperforming Q_1 (75.9%) and Q_2 (82.8%), and closely matching Q_5 (84.1%). In Table 4 with $s_i \sim U(2,3)$, it again led with 57.0% power, exceeding the power of Q_1 (44.8%) and Q_2 (55.6%).

Skewed distributions (Cases 3 and 4)

For the gamma (Case 3) and F-distributed (Case 4) inconsistency, the hybrid test continued to perform well. For instance, in Table 3's Case 4, the hybrid test's power was 77.3%, outperforming Q_1 (73.3%) and Q_8 (73.4%). In Table 4's Case 3, the hybrid test reached the power of 50.0%, on par with Q_8 (50.6%) and superior to Q_2 (48.4%). These demonstrated the hybrid test's adaptability to skewed distributions.

Bimodal distribution (Case 5)

With k = 15, the hybrid test demonstrated stable and competitive performance. In Table 1, the hybrid test attained a power of 56.6%, closely matching Q_2 (60.0%) and outperforming both Q_5 (55.2%) and Q_∞ (47.1%). Under the scenario with larger within-study variances, the hybrid test still achieved 23.6% power, comparable to the best-performing test Q_3 (24.2%) and exceeding Q_∞ (20.2%) (Table 2). These results suggested that the hybrid test maintained reasonable sensitivity in detecting inconsistency even when the underlying effect sizes may be from different subgroups.

Contamination scenarios (Cases 6 and 7)

In scenarios involving structural or data-driven contamination, test robustness is essential. Recall that Case 6 simulated effect reversal in one study, while Case 7 introduced a single extreme value. In Case 6, with k = 15, the hybrid test achieved 33.0% power under the setting of $s_i \sim U(1,2)$ in Table 1, higher than Q_2 (29.7%) and slightly below Q_∞ (34.7%). Under the setting of $s_i \sim U(2,3)$ in Table 2, all three tests performed similarly (with power in the range of 16.6–17.2%). With k = 30, the hybrid test's advantage became more evident. Under the setting of $s_i \sim U(1,2)$, its power reached 81.9%, notably higher than Q_2 (68.5%) and close to Q_8 (84.3%) and Q_∞ (84.6%), as shown in Table 3. Under the setting of $s_i \sim U(2,3)$, the hybrid test's power achieved 38.1%, ranking between Q_2 (30.4%) and Q_∞ (39.6%), as shown in Table 4. In Case 7, under k = 15, the hybrid test had power of 31.4% under the setting of $s_i \sim U(1,2)$, outperforming Q_2 (29.2%) and slightly below Q_∞ (32.2%) (Table 1). Under the setting of $s_i \sim U(2,3)$, its power achieved 14.7%, modestly below Q_2 (17.2%) (Table 2). With k = 30, Table 3 shows the hybrid test with power at 27.0% for $s_i \sim U(1,2)$, above Q_2 (22.9%) and slightly below Q_∞ (28.4%). For $s_i \sim U(2,3)$, the hybrid test's power was 12.8%, slightly behind Q_2 (13.1%) and Q_∞ (13.2%) (Table 4).

Overall, Q_2 exhibited excellent performance under normal inconsistency, making it a strong candidate when the true effect distribution is symmetric and well-behaved. Although the hybrid test had comparable or slightly lower power in this single scenario, it provided broader robustness across other settings involving skewness, heavy tails, or contamination. These findings suggest that the hybrid test offers greater adaptability in complex inconsistency structures.

Results of case studies

Figure 1 displays the forest plots of the three real-world meta-analyses used as our case studies. Table 5 presents their results after applying the various inconsistency tests and measures; it also gives the inconsistency measures E_γ. The effect measures of all three meta-analyses were the relative risk.

Specifically, the first meta-analysis was from Hughes et al. [38]; it assessed the effectiveness of ovulation suppression agents in the treatment of endometriosis-associated subfertility in improving pregnancy outcomes, including live birth. The number of studies in this meta-analysis was 11. The *P*-value of the traditional Q_2 test was 0.221, while the *P*-value of Q_γ was lower than 0.1 for $\gamma \geq 3$. The hybrid test yielded a *P*-value of 0.065,

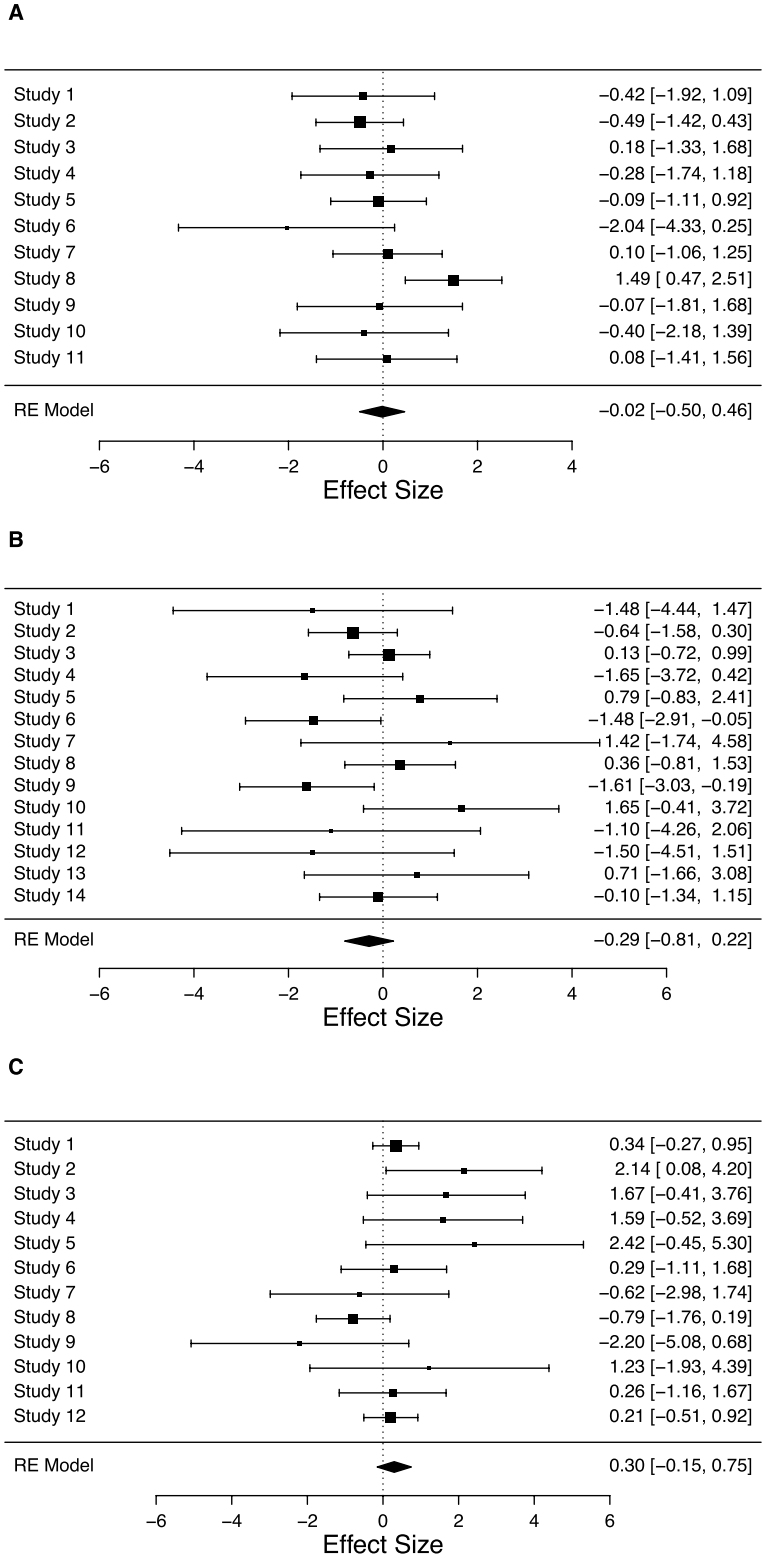

Fig. 1 The forest plots of three real-world meta-analyses

Table 5 *P*-values of the various Q_γ and hybrid tests for between-study inconsistency and their corresponding inconsistency measures E_γ for the three real-world meta-analyses

Meta-analysis		Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_∞	Q_{hyb}
Hughes et al. [38]	*P*-value	0.593	0.221	0.094	0.060	0.047	0.042	0.040	0.038	0.034	0.065
	E_γ	0	0.237	0.490	0.640	0.725	0.774	0.801	0.813	0.357	0.599
Gafter-Gvili et al. [39]	*P*-value	0.041	0.136	0.243	0.321	0.371	0.407	0.428	0.446	0.556	0.082
	E_γ	0.279	0.308	0.245	0.104	0	0	0	0	0	0.568
Saha et al. [40]	*P*-value	0.139	0.109	0.144	0.184	0.216	0.240	0.256	0.269	0.293	0.187
	E_γ	0.196	0.348	0.393	0.367	0.277	0.111	0	0	0.112	0.396

suggesting statistical significance at the 10% level (Fig. 1A and Table 5). This example highlights the hybrid test's improved sensitivity in detecting inconsistency, making it a valuable tool for enhancing the detection of inconsistency in meta-analyses.

The second meta-analysis was from Gafter-Gvili et al. [39]; it evaluated the effectiveness of antibiotic prophylaxis for preventing bacterial infections in afebrile neutropenic patients following chemotherapy. The number of studies was 14. The *P*-value of Q_1 test was relatively small (0.041); the *P*-value of the hybrid test was also lower than 0.1, indicating statistical significance (Fig. 1B and Table 5). The Q_2 and Q_γ with larger γ values had *P*-values greater than 0.1, failing to detect significant inconsistency.

The third meta-analysis was from Saha et al. [40], and it investigated the effects of chlorpromazine with atypical or second-generation antipsychotic drugs for the treatment of people with schizophrenia. It consisted of 12 studies. As shown in Table 5, most of the *P*-values across all tests were above the conventional significance threshold of 0.1. Specifically, the *P*-values of Q_1 and Q_2 were 0.139 and 0.109, respectively. Overall, the results indicate weak evidence of inconsistency across studies in this meta-analysis (Fig. 1C).

As a final remark, the hybrid test evaluates whether any deviation pattern among the family of candidate Q_γ statistics is convincingly present. Through resampling, it accounts for the correlations among the individual Q_γ tests and thus provides a conservative overall assessment when no evidence of inconsistency is detected, while being more sensitive when a particular γ value captures a pattern that others miss. In contrast, the individual Q_γ tests may not be interpreted collectively as formal hypothesis tests for overall inconsistency, because they do not adjust for multiple testing. A single significant Q_γ result does not necessarily indicate significant overall inconsistency; rather, these tests are intended to help explore and understand potential patterns of between-study inconsistency.

Discussion

This article introduces several alternative Q-like test statistics, denoted as Q_γ, for assessing between-study inconsistency in meta-analyses, along with a hybrid test that combines the strengths of these Q_γ tests. Our simulation studies and real-world case applications demonstrate that the hybrid test consistently achieves strong performance across a wide range of between-study inconsistency scenarios, particularly when the distribution of true effects deviates from normality.

A major strength of the proposed methods is their ability to address the challenge posed by non-normal between-study distributions. Although the assumption of normality has been a longstanding foundation in meta-analysis methodology and remains widely used in practice, its validity in any given analysis is often uncertain. Even if the assumption is approximately reasonable in many cases, it is difficult to verify whether it holds in a specific meta-analysis or whether its violation materially affects the assessment of inconsistency. This uncertainty largely stems from the fact that testing for between-study normality is inherently limited in power due to the typically small number of studies in meta-analyses.

The hybrid test is particularly well-suited to this dilemma. Without requiring knowledge of the precise form of the between-study distribution, it adaptively integrates information from a family of Q_γ statistics, each tailored to different patterns of inconsistency. As a result, it maintains relatively high power across diverse conditions, making it a robust tool for evidence synthesis.

Another strength of the hybrid test is its potential adaptability to meta-analyses affected by publication bias or small-study effects. In such cases, the between-study distribution may become skewed due to selective reporting, making the normality assumption unrealistic. Some of our simulation settings included skewed between-study distributions, which mimic scenarios involving publication bias. As discussed in the Introduction, the assessments of publication bias and between-study inconsistency are often interrelated and may influence each other. The proposed approach offers a promising tool to improve inconsistency assessment in the presence of publication bias, although further research is warranted to evaluate its performance under various

bias-generating mechanisms. In practice, researchers may first use conventional methods for detecting publication bias, such as funnel plots or Egger's regression test [41, 42]. If evidence of bias is detected, the traditional Q test may be suboptimal, and our proposed methods could serve as more robust alternatives.

Despite its strengths, the hybrid test has several limitations. First, like traditional meta-analysis models, it assumes that the within-study variances are known. This assumption may be questionable in certain contexts, such as studies with small sample sizes or rare event probabilities. More precise modeling approaches, such as generalized linear mixed models, can address this limitation [43, 44]. While our proposed test statistics and measures could be extended to these models, doing so would require replacing the sample-based within-study variances with the underlying variance parameters. Achieving this may necessitate additional methodological development, such as incorporating Bayesian hierarchical models [45], which represents a promising avenue for future research.

Second, although the methods are designed to relax the assumption of between-study normality, they still rely on the assumption of within-study normality. This can be justified by asymptotic arguments when studies have large sample sizes, but it may not hold when sample sizes are small.

Third, the choice of γ values used to construct the Q_γ tests in our study was limited to integers from 1 to 8 and infinity. In general, $\gamma = 1$ down-weights outliers and is sensitive to broad, mild dispersion; $\gamma = 2$ corresponds to the conventional Q statistic, which is most powerful under normality; larger γ values emphasize tail behavior and isolated anomalies; and $\gamma = \infty$ focuses on the maximum standardized deviate. While the set of $\gamma = 1, 2, \ldots, 8$, and ∞ performed well in our empirical examples, other γ values may be more appropriate in certain applications to better capture specific inconsistency patterns. Users are encouraged to examine the results across different Q_γ values and consider expanding the candidate set if the observed patterns suggest the need for additional sensitivity. Such adjustments may involve a trade-off between potential gains in power and computational burden, as including more γ values increases the complexity and runtime of the resampling procedure. For example, if the power of tests increases steadily from $\gamma = 1$ to 8, users may explore additional values such as 10, 15, or 20. Conversely, if the power of Q_8 and Q_∞ is substantially lower than that of tests with smaller γ values, there may be little benefit in considering larger γ values. In such cases, users may even consider excluding $\gamma = 8$ and ∞ from the candidate set, as incorporating too many γ values can introduce noise into the hybrid test and potentially reduce its overall power.

In addition, our simulation studies for validating the proposed methods included only meta-analyses with 15 or 30 studies. In practice, many meta-analyses contain fewer studies [46, 47], and the statistical power of all tests, including the proposed ones, would be expected to decrease substantially in such cases. When the number of studies is very small, relying solely on statistical testing to assess heterogeneity may not be valid or reliable because of the limited power [48]. We recommend that researchers also consider the potential sources and implications of heterogeneity from clinical perspectives, such as the presence of effect modifiers or study-level characteristics that may influence effect sizes [49, 50]. Despite our simulations being restricted to 15 and 30 studies, we believe that the results adequately demonstrate how the proposed hybrid approach integrates the strengths of various component tests and achieves competitive performance across diverse scenarios.

Conclusions

We proposed a family of alternative Q-like tests and a hybrid test to assess between-study inconsistency in meta-analyses, offering flexible and powerful tools beyond the conventional Q test. Through simulations and real-world applications, we demonstrated that the hybrid test maintains robust performance across a wide range of inconsistency scenarios, particularly when the between-study distribution deviates from normality. These methods provide practical solutions for improving the detection and quantification of inconsistency in meta-analysis, especially when traditional assumptions are questionable. Future work may extend these approaches to accommodate more complex models and further refine the choice of test statistics.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-025-02719-7.

Additional file 1: Derivation of the expectations of Q_γ statistics under the null hypothesis.

Acknowledgements

ChatGPT 4o was used solely to improve the writing of this manuscript and was not employed for any other purposes in this study.

Authors' contributions

ZY: methodology, formal analysis, investigation, writing—original draft, visualization; MX: validation, investigation, formal analysis, writing—review & editing, visualization; XX: validation, formal analysis, writing—original draft, visualization; LL: conceptualization, data curation, writing—original draft, writing—review & editing, supervision, funding acquisition.

Funding

This study was supported in part by the US National Institute of Mental Health grant R03 MH128727, the US National Institute on Aging grant R03 AG093555, the US National Library of Medicine grants R21 LM014533 and R01 LM012982, and the Arizona Biomedical Research Centre grant RFGA2023-008–11. The

content is solely the responsibility of the authors and does not necessarily represent the official views of the US National Institutes of Health and the Arizona Department of Health Services.

Data availability
The R code for the case studies is available at https://osf.io/4t9wj/. The main R function is also included in our R package "altmeta," which is available on CRAN.

Declarations

Ethics approval and consent to participate
Ethics approval and consent to participate were not required for this study because it used published data in the existing literature.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 29 July 2025 / Accepted: 27 October 2025
Published online: 20 November 2025

References

1. Gurevitch J, Koricheva J, Nakagawa S, Stewart G. Meta-analysis and the science of research synthesis. Nature. 2018;555(7695):175–82.
2. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane handbook for systematic reviews of interventions. Chichester, UK: John Wiley & Sons; 2019.
3. Higgins JPT. Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. Int J Epidemiol. 2008;37(5):1158–60.
4. 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.
5. Riley RD, Higgins JPT, Deeks JJ. Interpretation of random effects meta-analyses. BMJ. 2011;342:d549.
6. Biggerstaff BJ, Tweedie RL. Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. Stat Med. 1997;16(7):753–68.
7. Jackson D. The power of the standard test for the presence of heterogeneity in meta-analysis. Stat Med. 2006;25(15):2688–99.
8. Biggerstaff BJ, Jackson D. The exact distribution of Cochran's heterogeneity statistic in one-way random effects meta-analysis. Stat Med. 2008;27(29):6093–110.
9. 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.
10. Jackson D, White IR. When should meta-analysis avoid making hidden normality assumptions? Biom J. 2018;60(6):1040–58.
11. Wang C-C, Lee W-C. Evaluation of the normality assumption in meta-analyses. Am J Epidemiol. 2020;189(3):235–42.
12. Al Amer FM, Lin L. Prediction performance of earlier studies for later studies in Cochrane reviews. J Eval Clin Pract. 2025;31(4):e70172.
13. Viechtbauer W, Cheung MWL. Outlier and influence diagnostics for meta-analysis. Res Synth Methods. 2010;1(2):112–25.
14. Sun X, Briel M, Busse JW, You JJ, Akl EA, Mejza F, et al. The influence of study characteristics on reporting of subgroup analyses in randomised controlled trials: systematic review. BMJ. 2011;342:d1569.
15. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L, Moreno SG. Assessing publication bias in meta-analyses in the presence of between-study heterogeneity. J R Stat Soc Ser A Stat Soc. 2010;173(3):575–91.
16. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. Stat Med. 2007;26(25):4544–62.
17. Terrin N, Schmid CH, Lau J, Olkin I. Adjusting for publication bias in the presence of heterogeneity. Stat Med. 2003;22(13):2113–26.
18. Huedo-Medina TB, Sánchez-Meca J, Marín-Martínez F, Botella J. Assessing heterogeneity in meta-analysis: Q statistic or I^2 index? Psychol Methods. 2006;11(2):193–206.
19. Jia P, Lin L, Kwong JSW, Xu C. Many meta-analyses of rare events in the Cochrane database of systematic reviews were underpowered. J Clin Epidemiol. 2021;131:113–22.
20. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ. 2003;327(7414):557–60.
21. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 7. Rating the quality of evidence—inconsistency. J Clin Epidemiol. 2011;64(12):1294–302.
22. Higgins JPT, Jackson D, Barrett JK, Lu G, Ades AE, White IR. Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. Res Synth Methods. 2012;3(2):98–110.
23. White IR, Barrett JK, Jackson D, Higgins JPT. Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression. Res Synth Methods. 2012;3(2):111–25.
24. Viechtbauer W. Hypothesis tests for population heterogeneity in meta-analysis. Br J Math Stat Psychol. 2007;60(1):29–60.
25. Zhang C, Wang X, Chen M, Wang T. A comparison of hypothesis tests for homogeneity in meta-analysis with focus on rare binary events. Res Synth Methods. 2021;12(4):408–28.
26. Kulinskaya E, Dollinger MB. An accurate test for homogeneity of odds ratios based on Cochran's *Q*-statistic. BMC Med Res Methodol. 2015;15(1):49.
27. Pan W, Kim J, Zhang Y, Shen X, Wei P. A powerful and adaptive association test for rare variants. Genetics. 2014;197(4):1081–95.
28. Xu G, Lin L, Wei P, Pan W. An adaptive two-sample test for high-dimensional means. Biometrika. 2016;103(3):609–24.
29. Lin L. Bias caused by sampling error in meta-analysis with small sample sizes. PLoS ONE. 2018;13(9):e0204056.
30. Baker R, Jackson D. A new approach to outliers in meta-analysis. Health Care Manag Sci. 2008;11(2):121–31.
31. Lee KJ, Thompson SG. Flexible parametric models for random-effects distributions. Stat Med. 2008;27(3):418–34.
32. Lin L. Hybrid test for publication bias in meta-analysis. Stat Methods Med Res. 2020;29(10):2881–99.
33. Lin L. Comparison of four heterogeneity measures for meta-analysis. J Eval Clin Pract. 2020;26(1):376–84.
34. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002;21(11):1539–58.
35. Ioannidis JPA, Trikalinos TA. An exploratory test for an excess of significant findings. Clin Trials. 2007;4(3):245–53.
36. Stanley TD, Doucouliagos H, Ioannidis JPA, Carter EC. Detecting publication selection bias through excess statistical significance. Res Synth Methods. 2021;12(6):776–95.
37. Hardy RJ, Thompson SG. Detecting and describing heterogeneity in meta-analysis. Stat Med. 1998;17(8):841–56.
38. Hughes E, Brown J, Collins JJ, Farquhar C, Fedorkow DM, Vanderkerchove P. Ovulation suppression for endometriosis for women with subfertility. Cochrane Database Syst Rev. 2007;3:CD000155.
39. Gafter-Gvili A, Fraser A, Paul M, Vidal L, Lawrie TA, van de Wetering MD, Kremer LCM, Leibovici L. Antibiotic prophylaxis for bacterial infections in afebrile neutropenic patients following chemotherapy. Cochrane Database Syst Rev. 2012;1:CD004386.
40. Saha KB, Bo L, Zhao S, Xia J, Sampson S, Zaman RU. Chlorpromazine versus atypical antipsychotic drugs for schizophrenia. Cochrane Database Syst Rev. 2016;4:CD010631.
41. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. J Clin Epidemiol. 2008;61(10):991–6.
42. 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.
43. Ju K, Lin L, Chu H, Cheng L-L, Xu C. Laplace approximation, penalized quasi-likelihood, and adaptive Gauss-Hermite quadrature for generalized linear mixed models: towards meta-analysis of binary outcome with sparse data. BMC Med Res Methodol. 2020;20(1):152.
44. 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):176–89.
45. Shi L, Chu H, Lin L. A bayesian approach to assessing small-study effects in meta-analysis of a binary outcome with controlled false positive rate. Res Synth Methods. 2020;11(4):535–52.

46. 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 Med Res Methodol. 2011;11:160.
47. von Hippel PT. The heterogeneity statistic I^2 can be biased in small meta-analyses. BMC Med Res Methodol. 2015;15:35.
48. Bender R, Friede T, Koch A, Kuss O, Schlattmann P, Schwarzer G, et al. Methods for evidence synthesis in the case of very few studies. Res Synth Methods. 2018;9(3):382–92.
49. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. BMJ. 1994;309(6965):1351–5.
50. Sedgwick P. Meta-analyses: heterogeneity and subgroup analysis. BMJ. 2013;346:f4040.

Publisher's Note