Comparison of four heterogeneity measures for meta-analysis

Lifeng Lin PhD, Assistant Professor

Department of Statistics, Florida State University, Tallahassee, Florida

Correspondence
Lifeng Lin, PhD, Department of Statistics, Florida State University, Tallahassee, FL 32306.
Email: linl@stat.fsu.edu

Abstract

Rationale, aims, and objectives: Heterogeneity is a critical issue in meta-analysis, because it implies the appropriateness of combining the collected studies and impacts the reliability of the synthesized results. The Q test is a traditional method to assess heterogeneity; however, because it does not have an intuitive interpretation for clinicians and often has low statistical power, many meta-analysts alter to use some measures, such as the $I^2$ statistic, to quantify the extent of heterogeneity. This article aims at providing a summary of available tools to assess heterogeneity and comparing their performance.

Methods: We reviewed four heterogeneity measures ($I^2$, $R_I$, $R_M$, and $R_B$) and illustrated how they could be treated as test statistics like the Q statistic. These measures were compared with respect to statistical power based on simulations driven by three real-data examples. The pairwise agreement among the four measures was also evaluated using Cohen’s $\kappa$ coefficient.

Results: Generally, $R_I$ was slightly more powerful than the Q test, while its type I error rate might be slightly inflated. The power of $I^2$ was fairly close to that of Q. The $R_M$ and $R_B$ statistics might have low powers in some cases. Because the differences between the powers of $I^2$, $R_I$, and $Q$ were often tiny, meta-analysts might not expect $I^2$ and $R_I$ to yield significant heterogeneity if the Q test failed to do so. In addition, $I^2$ and $R_I$ had fairly good agreement based on the simulated meta-analyses, but all other pairs of heterogeneity measures generally had poor agreement.

Conclusion: The $I^2$ and $R_I$ statistics are recommended for measuring heterogeneity. Meta-analysts should use the heterogeneity measures as descriptive statistics which have intuitive interpretations from the clinical perspective, instead of determining the significance of heterogeneity simply based on their magnitudes.

KEYWORDS
heterogeneity, $I^2$ statistic, meta-analysis, statistical power

1 | INTRODUCTION

Meta-analysis has been a popular statistical method to synthesize evidence from different studies in clinical research. Assessing heterogeneity between the collected studies in a meta-analysis plays a critical role in examining whether the studies may be properly combined and the synthesized results are reliable. Of note, the specific meaning of heterogeneity may vary across different disciplines. For example, it could be the differences between contexts of materials or between strata in a geographical space. In this article, heterogeneity in meta-analyses of clinical studies refers to the variation in the reported treatment effect estimates across the collected studies. A traditional statistical method to detect heterogeneity in meta-analyses is the Q test, but it is generally recognized to have low statistical power in many cases. Also, it...
only produces \( P \) values that indicate a binary decision of either presence or absence of heterogeneity, and it may not be attractive for evidence users who are more interested in the magnitude of heterogeneity. Because of these, more meta-analyses in recent years have emphasized on quantifying heterogeneity using certain measures.\(^{18}\)

This article considers a class of heterogeneity measures, including the well-known \( I^2 \) statistic\(^{18}\) and the \( \hat{R}_i \), \( \hat{R}_M \), and \( \hat{R}_b \) statistics,\(^{19-22}\) that has an attractive interpretation as “the proportion of total variation caused by heterogeneity rather than within-study sampling error.” All four measures range between 0% and 100%. Conceptually, these measures describe the proportion \( \frac{\hat{\tau}^2}{\hat{\tau}^2 + \sigma^2} \), where \( \hat{\tau}^2 \) is the between-study variance caused by heterogeneity and \( \sigma^2 \) represents a summary of within-study variances. The following section introduces details about the heterogeneity measures, and Table 1 summarizes their definitions. Among them, the \( I^2 \) statistic has been used most frequently in meta-analyses. The Cochrane Handbook for Systematic Reviews of Interventions provides a rule of thumb for interpreting the \( I^2 \) statistic; that is, \( I^2 \leq 40\% \) may indicate unimportant heterogeneity, \( 30\% \leq I^2 \leq 60\% \) may represent moderate heterogeneity, \( 50\% \leq I^2 \leq 90\% \) may represent substantial heterogeneity, and \( 75\% \leq I^2 \leq 100\% \) implies that heterogeneity may be considerable.\(^{14}\) These ranges overlap because they are rough, and the true heterogeneity should be assessed not only from the statistical perspective but also from the clinical perspective.\(^{23}\) Nevertheless, in many applications, the values of 25\%, 50\%, and 75\% have been often used as the cut-off points of \( I^2 \) to differentiate the extents of heterogeneity for convenience.\(^{24}\)

Several studies have evaluated these heterogeneity measures regarding their biases and confidence interval coverages.\(^{18,20,22,25,26}\) However, the concept of “the proportion of total variation caused by heterogeneity rather than within-study sampling error” is not uniquely defined, because different measures use different summaries of within-study variances \( \sigma^2 \).\(^{27,28}\) Each measure’s true value is usually defined by replacing the estimated between-study variance \( \hat{\tau}^2 \) with the true between-study variance \( \tau^2 \) (see Table 1). For example, the true proportion of total variation caused by heterogeneity is calculated as \( \frac{\tau^2}{\tau^2 + \sigma^2} \) when the \( \hat{R}_i \) statistic is used, while the true proportion is \( \frac{\tau^2}{\tau^2 + \frac{\sigma^2}{n}} \) for the \( \hat{R}_M \) statistic. Here, \( n \) is the number of studies in the meta-analysis, and \( \sigma^2_i \) are the within-study variances. Therefore, different measures have different true values, and thus, comparing the measures’ biases and confidence interval coverages may be unfair.

Instead, researchers may pay more attention to how precisely the measures describe heterogeneity in terms of statistical power. In fact, measuring heterogeneity is closely related to testing for it. Any heterogeneity measure should be able to serve as a test statistic, because it monotonically increases as heterogeneity increases. Its statistical power determines its precision as a measure of heterogeneity. Analogously, in linear regression, the coefficient of determination \( R^2 \) statistic is widely used to measure the proportion of variation in responses explained by linear predictors.\(^{29}\) If treated as a test statistic, \( R^2 \) corresponds to the hypothesis testing of regression slopes being zero.

This article explores the performance of the four heterogeneity measures serving as test statistics. We first review the derivations of the various measures and provide details about the calculations of their statistical powers. Then, three real-world examples are provided to illustrate the use of the measures, and we compare the statistical powers and type I error rates of the measures and the \( Q \) test. We also evaluate the agreement among the four heterogeneity measures using Cohen’s \( k \) coefficient. This article concludes with a brief discussion.

### 2 Methods

#### 2.1 Heterogeneity measures

The \( I^2 \) statistic has been the most popular tool to quantify heterogeneity among the four measures.\(^{18,24}\) Its motivation was based on a tentative but unrealistic assumption of equal within-study variances. Specifically, consider a meta-analysis containing \( n \) studies; each study

<table>
<thead>
<tr>
<th>Heterogeneity Measure</th>
<th>Expressed as a Function of ( \tau^2 )</th>
<th>Expressed as a Function of ( Q )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( I^2 )</td>
<td>( \frac{\hat{\tau}^2}{\hat{\tau}^2 + \frac{(n-1)\sum \sigma^2_i}{(\sum \sigma^2_i)^2} - \frac{\hat{\sigma}^2}{\hat{\tau}^2}} )</td>
<td>( Q = \frac{Q}{\frac{(n-1)\sum \sigma^2_i}{(\sum \sigma^2_i)^2} - \frac{\hat{\sigma}^2}{\hat{\tau}^2}} )</td>
</tr>
<tr>
<td>( \hat{R}_i )</td>
<td>( \frac{\hat{\tau}^2}{\hat{\tau}^2 + \frac{\hat{\sigma}^2}{\hat{\tau}^2}} )</td>
<td>( Q = \frac{Q}{\hat{\tau}^2 + \frac{\hat{\sigma}^2}{\hat{\tau}^2}} )</td>
</tr>
<tr>
<td>( \hat{R}_M )</td>
<td>( \frac{\hat{\tau}^2}{\hat{\tau}^2 + \frac{\hat{\sigma}^2}{\hat{\tau}^2}} )</td>
<td>( Q = \frac{Q}{\hat{\tau}^2 + \hat{\sigma}^2} )</td>
</tr>
<tr>
<td>( \hat{R}_b )</td>
<td>( \frac{1}{n} \frac{\sum \sigma^2_i}{\hat{\tau}^2 + \frac{\sigma^2}{\hat{\tau}^2}} )</td>
<td>( Q = \frac{Q}{\frac{(n-1)\sum \sigma^2_i}{(\sum \sigma^2_i)^2} - \frac{\hat{\sigma}^2}{\hat{\tau}^2}} )</td>
</tr>
</tbody>
</table>

Note. \( n \), the number of studies in a meta-analysis; \( \sigma^2_i \), the observed effect size and its sample variance within study \( i \); \( \hat{\tau}^2 = \frac{\sum (y_i - \bar{y})^2}{\sum 1/s^2_i} \), the method-of-moments estimate of the between-study variance \( \tau^2 \); \( Q = \sum \sigma^2_i (y_i - \bar{y})^2 \), where \( \bar{y} = \frac{\sum y_i}{\sum 1/s^2_i} \). If \( Q < n-1 \), then \( Q \) is truncated as \( n-1 \) in the expressions.
reports a point estimate of treatment effect $y_i$ with within-study variation $s_i^2$. The $I^2$ statistic was originated from the traditional Q statistic, which is defined as $Q = \sum s_i^2(y_i - \bar{y})^2$, where $\bar{y} = \frac{\sum y_i / s_i^2}{\sum 1 / s_i^2}$ is the pooled common-effect estimate of the overall treatment effect $\mu$. The expectation of the Q statistic is as follows:

$$E(Q) = r^2 \left( \frac{\sum s_i^{-2} - \sum s_i^{-4}}{\sum s_i^{-2}} \right) + n - 1,$$

where $r^2$ is the true between-study variance. Here, the within-study variances $s_i^2$ are treated as known "true values" throughout this article as in most conventional meta-analysis methods, although in practice they are sample variances and are actually random variables. If the within-study variances are assumed to be equal, ie, $s_i^2 = \sigma^2$, then the expectation of Q becomes

$$E(Q) = (n - 1) \left( \frac{r^2}{\sigma^2} + 1 \right).$$

Using the method of moments, ie, equating the observed Q with its expectation, the proportion of total variation caused by heterogeneity rather than within-study sampling error can be evaluated using the $I^2$ statistic:

$$I^2 = \frac{r^2}{\sigma^2 + r^2} = \frac{Q - (n - 1)}{Q}, \quad (1)$$

which is a function of Q. Because conceptually the proportion should be positive, Q is often truncated as $n - 1$ if $Q < n - 1$, so that $I^2$ is between 0% and 100%. Such a truncation will be applied throughout this article.

Alternatively, the $I^2$ statistic can be also expressed as a function of the between-study variance. Consider the method-of-moments estimate of the between-study variance $\tilde{r}^2$ by DerSimonian and Laird:

$$\tilde{r}^2 = \frac{Q - (n - 1)}{\sum s_i^{-2} - \sum s_i^{-4}} \sum s_i^{-2} \quad (2)$$

Again, the truncation of Q is used to guarantee the nonnegativity of this variance estimate. Expressing the Q statistic in the form of $\tilde{r}^2$ via Equation (2) and plugging it in Equation (1), we have

$$I^2 = \frac{\tilde{r}^2}{\tilde{r}^2 + \frac{(n - 1)\sum s_i^{-4}}{(\sum s_i^{-2})^2 - \sum s_i^{-4}}} \quad (3)$$

Therefore, the $I^2$ statistic effectively treats the summary of within-study variances as follows:

$$\sigma^2 = \frac{(n - 1)\sum s_i^{-2}}{(\sum s_i^{-2})^2 - \sum s_i^{-4}}$$

As the $I^2$ statistic was originally motivated by the expectation of the Q statistic, it has a simpler form expressed in terms of Q in Equation (1) than that expressed in terms of $\tilde{r}^2$ in Equation (3). The latter form is seldom used in the literature.

Unlike the $I^2$ statistic, the measures $\hat{R}_l$ and $\hat{R}_m$ were directly originated from certain summaries of within-study variances $\sigma^2$, instead of the Q statistic. The $\hat{R}_l$ statistic uses the harmonic mean of within-study variances to estimate $\sigma^2$, while the $\hat{R}_m$ statistic uses the arithmetic mean. Using the method of moments, ie, equating the observed statistic:

$$\hat{R}_l = \frac{I^2}{I^2 + \sigma^2} \quad \text{and} \quad \hat{R}_m = \frac{I^2}{I^2 + \frac{1}{n} \sum s_i^2}$$

Using Equation (2) to express $\tilde{r}^2$ in terms of Q, the two measures can be also written as functions of Q, as shown in Table 1, while such forms are not as straightforward as those in terms of $\tilde{r}^2$.

The $\hat{R}_l$ statistic was recently proposed by Crippa et al. It uses the random-effects meta-analysis result to directly impute the total variation $r^2 + \sigma^2$. Specifically, the $\hat{R}_l$ statistic is originally defined in terms of the estimated between-study variance:

$$\hat{R}_l = \frac{\tilde{r}^2}{\tilde{r}^2 + \frac{1}{n}\sum s_i^2}$$

Note that $\frac{\tilde{r}^2}{\tilde{r}^2 + \frac{1}{n}\sum s_i^2}$ represents the proportion of heterogeneity in study i; therefore, the $\hat{R}_l$ statistic can be considered as an average of study-specific proportions of heterogeneity. Again, this measure can be alternatively expressed in terms of the Q statistic via Equation (2) as in Table 1, but the alternative form is much more complicated and much less intuitive.

### 2.2 | P values of heterogeneity measures

The $P$ value of the Q test can be easily calculated because the Q statistic follows a $\chi^2$ distribution with $n - 1$ degrees of freedom under the null hypothesis of homogeneity (ie, $H_0 : r^2 = 0$). If ignoring the variation of the within-study variances $s_i^2$ and treating them as known true values. However, the four heterogeneity measures have complicated theoretical null distributions. Due to this difficulty, these measures have been rarely recognized as test statistics to produce $P$ values. To avoid the problems in deriving the measures’ exact null distributions, the parametric resampling method can be applied to calculate the $P$ values of the measures, as well as the Q test.
The steps for the parametric resampling method are as follows. First, under the null hypothesis of homogeneity, we calculate the common-effect estimate as $\hat{\mu} = \frac{\sum y_i}{\sum 1/s_i^2}$ and denote the heterogeneity measures $\Delta$ and the Q statistic in the original meta-analysis as $\Delta^{(0)}$, $\hat{\Delta}^{(0)}$, $\hat{\Delta}_M^{(0)}$, $\hat{\Delta}_R^{(0)}$, and $Q^{(0)}$, accordingly. Second, $n$ within-study variances are sampled with replacement from those ($i.e., s_i^2$) in the original meta-analysis. Third, the point estimates of the $n$ resampled studies are generated using the normal distributions with mean $\hat{\mu}$ and the sampled variances under the null hypothesis; these studies form a resampled meta-analysis. Fourth, repeat the second and third steps for $K$ (say, 1000) iterations. Finally, we calculate the four measures and Q for each resampled meta-analysis, denoted as $\Delta^{(k)}$, $\hat{\Delta}^{(k)}$, $\hat{\Delta}_M^{(k)}$, $\hat{\Delta}_R^{(k)}$, and $Q^{(k)}$ ($k = 1, 2, ... , K$). Consequently, their $P$ values are calculated as follows:

$$P_X = \sum_{k=1}^{K} \left[ 1 \left( X^{(k)} \geq X^{(0)} \right) \right] / K,$$

where $X$ represents any heterogeneity measure or the Q statistic, and $1(t)$ is the indicator function with $1(t) = 0$ if the statement $t$ is false and $1(t) = 1$ if $t$ is true.

Given the number of studies $n$, note that the $\Delta^2$ statistic depends only on the Q statistic as in Equation (1), and it is an increasing function of Q. Therefore, $\Delta^{(k)} \geq \Delta^{(0)}$ is equivalent to $Q^{(k)} \geq Q^{(0)}$, consequently, the $P$ value of $I^2$ equals to that of Q when using the resampling method, leading to the same statistical power and type I error rate. Of note, some statistical large-sample properties (eg, the delta method) may be also used to derive an approximated variance of $I^2$ as in Higgins and Thompson and thus to yield an asymptotics-based $P$ value of $I^2$. However, the large-sample approximation may be fairly poor for meta-analyses with a few (say, less than 20) studies, while such meta-analyses are common in practice; therefore, such $P$ values may not be accurate.

Because $\hat{\Delta}^{(k)}$, $\hat{\Delta}_M^{(k)}$, and $\hat{\Delta}_R^{(k)}$ must depend on the within-study variances besides Q, their mathematical forms may not clearly reveal their statistical power compared with Q. Section 2.4 explores their resampling-based $P$ values and statistical powers using real data and simulations.

2.3 Other available heterogeneity measures

Besides the aforementioned four heterogeneity measures, several other measures are available for different purposes. For example, Higgins and Thompson also proposed the $H$ and $R$ statistics, while they have been generally less popular than $I^2$. The $H$ statistic is interpreted as the ratio of the standard deviation of the summary estimate from a random-effects meta-analysis compared with that from a common-effect meta-analysis. In other words, it describes the inflation in the confidence interval of the summary estimate under a random-effects setting (heterogeneity) compared with a common-effect setting (homogeneity). The $R$ statistic is defined differently but has a similar interpretation with $H$. The $I^2$ statistic has been used more frequently than these two measures primarily because the concept of “proportion of variance explained” is widely familiar among meta-analysts.

Moreover, meta-analyses often contain some outlying studies that are inappropriately or mistakenly selected into the systematic reviews; these outliers may have substantial impact on assessing heterogeneity, and $I^2$ may overestimate heterogeneity. To reduce such impact, Lin et al proposed two alternative heterogeneity measures $I^2$ and $I^2_m$, which have the same interpretation as $I^2$ and are robust in the presence of outliers. Their robustness has been evaluated using both theoretical properties and empirical studies. In addition to these heterogeneity measures mostly used in meta-analyses of clinical studies, Wang et al introduced the $q$ statistic to assess spatial stratified heterogeneity in ecological and geographical research. Similar to $I^2$, a $q$ value of 0 indicates no spatial stratification of heterogeneity, while $q = 1$ indicates a perfect spatial stratification of heterogeneity.

These alternative measures were originally proposed for different specific purposes; $I^2$ and $I^2_m$ were specially designed for the presence of outliers, and $q$ was designed for spatial stratified heterogeneity. Their performance may be preferred in those specific cases and has been investigated in the original articles that proposed them. Consequently, in this article, the following comparisons will focus on the four heterogeneity measures reviewed in Section 2.1 that have the same interpretation for generic meta-analyses.

2.4 Data analyses

We considered three meta-analyses with different numbers of studies. O’Doherty et al presented a meta-analysis with three studies on the effect of screening on identification of women experiencing intimate partner violence in emergency clinics. Makani et al collected 16 studies to compare dual blockade of the renin-angiotensin system with monotherapy for hyperkalemia in cohorts without heart failure. One study had no events in both treatment groups, so its treatment effect was not estimable and the meta-analysis effectively contained 15 studies. Moreover, Myung et al conducted a meta-analysis with 50 studies to investigate the efficacy of vitamin and antioxidant supplements in prevention of major cardiovascular events. The effect sizes in all three meta-analyses were the risk ratios, which were analysed on a logarithmic scale. Figures 1–3 show their forest plots, which were produced using the R package "meta." We applied the four heterogeneity measures to these meta-analyses and calculated the $P$ values of these measures and the Q test based on the resampling method described in Section 2.2 with 100 000 iterations. The theoretical $\chi^2$ distribution was also used to calculate the Q test’s $P$ value. The significance level was set to 5%. The R code for the three meta-analyses is available in the Supporting Information.

In addition, driven by these real-world examples, we conducted simulations to investigate the statistical powers and the type I error rates of the heterogeneity measures. Specifically, based on each example, 10 000 meta-analyses were simulated using the real dataset’s overall effect size and the number of studies $n = 3, 15, or 50$. The within-study standard errors were sampled from the uniform distribution ranging from the minimum to the maximum of the observed standard errors in each real dataset. For example, when
the simulations were based on the meta-analysis by O’Doherty et al,\textsuperscript{36} whose within-study standard errors were 1.02, 0.24, and 0.47 in its three studies, the within-study standard errors in the simulated meta-analyses were sampled from $U(0.24, 1.02)$. In addition, we considered various values for the between-study standard deviation $\tau$. Specifically, $\tau$ was from 0 to 2.5 by 0.5 for the simulations based on the meta-analysis by O’Doherty et al,\textsuperscript{36} was from 0 to 0.75 by 0.15 for those based on the meta-analysis by Makani et al,\textsuperscript{37} and was from 0 to 0.5 by 0.1 for those based on the meta-analysis by Myung et al.\textsuperscript{38} The between-study standard deviation was not further increased because the statistical power would be too high (eg, above 80\% or 90\%), and the differences between the measures would be minimal.

To calculate the $P$ values of the heterogeneity measures, 1000 resampling iterations were used for each simulated meta-analysis. The significance level remained to be 5\%.

Based on the simulated meta-analyses, we also used Cohen’s $\kappa$ coefficient to assess pairwise agreement among all four heterogeneity measures under each setting.\textsuperscript{40} The extent of heterogeneity in each simulated meta-analysis was categorized as unimportant, moderate, substantial, and considerable roughly using the cut-off points at 25\%, 50\%, and 75\% of the heterogeneity measures. The $\kappa$ coefficients less than 0.4, 0.4 to 0.75, and greater than 0.75 represented poor, fair to good, and excellent agreement, respectively; see, eg, Fleiss et al.\textsuperscript{41(p604)}

### RESULTS

#### 3.1 Real-data examples

For the meta-analysis by O’Doherty et al,\textsuperscript{36} the $Q$ test led to $P_Q = 0.070$ (based on the theoretical $\chi^2$ distribution) and 0.071 (based on the resampling method), and the heterogeneity measures were $I^2 = 62.3\%$, $R_i = 74.5\%$, $R_M = 46.3\%$, and $R_b = 58.9\%$ with $P$ values $P_{I^2} = .071$, $P_{R_i} = .057$, $P_{R_M} = .105$, and $P_{R_b} = .079$, respectively. For the meta-analysis by Makani et al,\textsuperscript{37} the $Q$ test had $P_Q = 0.054$ (based on the theoretical $\chi^2$ distribution) and 0.053 (based on the resampling method), and the heterogeneity measures were $I^2 = 40.2\%$, $R_i = 58.9\%$, $R_M = 4.4\%$, and $R_b = 19.0\%$ with $P$ values $P_{I^2} = .057$, $P_{R_i} = .058$, $P_{R_M} = .138$, and $P_{R_b} = .120$. For the meta-analysis by Myung et al.\textsuperscript{38} the $Q$ test had a $P$ value around .002 (based on both the theoretical $\chi^2$ distribution and 0.053 (based on the resampling method), and the heterogeneity measures were $I^2 = 41.3\%$, $R_i = 42.8\%$, $R_M = 2.2\%$, and $R_b = 26.1\%$ with $P$ values $P_{I^2} = .002$, $P_{R_i} = .002$, $P_{R_M} = .014$, and $P_{R_b} = .003$.

In the first two meta-analyses, all $P$ values were larger than .05, and we did not reject the null hypothesis of homogeneity, although the heterogeneity measures were mostly moderate or high, implying
nonignorable heterogeneity. In the last two meta-analyses, the \( b_{RM} \) and \( b_{Rb} \) statistics were noticeably smaller than \( I^2 \) and \( b_{RI} \). Their values (2.2% and 26.1%) seemed to seriously underestimate heterogeneity in the last meta-analysis by Myung et al, because all \( P \) values were much smaller than .05, indicating statistically significant heterogeneity.

### 3.2 Simulation studies

Table 2 presents the statistical powers and type I error rates of the heterogeneity measures and the \( Q \) test. The Monte Carlo standard error for each result was smaller than 1%. The \( I^2 \) statistic and the \( Q \) test controlled the type I error rate very well, as the rate was fairly close to 5% across different settings. The type I error rate of the \( b_{RI} \) statistic slightly inflated (by from 0.5% to 1.4%). Moreover, the type I error rates of the \( b_{RM} \) and \( b_{Rb} \) statistics did not exceed 5% under all settings, and the rates were as low as 1.9% and 3.2% under the setting based on the meta-analysis by Makani et al.

As for statistical power, the \( I^2 \) statistic produced nearly the same results with the \( Q \) test. Generally, the \( b_{RI} \) statistic was slightly more powerful than the \( I^2 \) statistic and the \( Q \) test, possibly because its type I error rate was also a bit higher. The \( b_{RM} \) statistic had the lowest power among the four heterogeneity measures, and the \( b_{Rb} \) statistic was also noticeably less powerful than the \( Q \) test and the \( I^2 \) and \( b_{RI} \) statistics.

Table 3 presents Cohen's \( \kappa \) coefficients in the simulated meta-analyses. Under the setting based on the meta-analysis by Makani et al containing only three studies, all four heterogeneity measures generally had excellent agreement because nearly all \( \kappa \) coefficients were at least 0.75. However, for larger simulated meta-analyses.
4.1 Main findings

This article has reviewed four measures that can be used to quantify heterogeneity in meta-analyses of clinical studies. The $\hat{R}_b$ statistic was found to be slightly more powerful than the Q test and the $I^2$ statistic in some cases, while its type I error rate was also slightly inflated. The Q test and the $I^2$ statistic had nearly the same power. The power of the $\hat{R}_b$ statistic was noticeably lower than that of the foregoing measures, and the $\hat{R}_M$ statistic was generally the least powerful one among the four measures. In addition, based on the simulated meta-analyses, the agreement between $I^2$ and $\hat{R}_b$ was fairly good, while all other pairs of heterogeneity measures generally had poor agreement.

4.2 Strengths and limitations

Most existing studies compared heterogeneity measures in terms of biases, mean squared errors, confidence interval coverages, etc. Although all four heterogeneity measures reviewed in this article can be interpreted as the proportion of total variation due to heterogeneity, such a proportion is not a uniquely defined concept. Different measures use different estimates to summarize the within-study variability, such a proportion is not a uniquely defined concept. Different measures use different estimates to summarize the within-study variability, and it may be unfair to compare the measures with respect to characteristics such as biases. Alternatively, this article compared the four measures with respect to their statistical powers (and also type I error rates) via simulations driven by real data, which may better reflect the performance of the measures for detecting heterogeneity. The simulation settings covered various scenarios with different numbers of studies and different extents of heterogeneity.
4.3 | Recommendations

In summary, based on their statistical powers and agreement, both the \( I^2 \) and \( \hat{R}_I \) statistics may be preferred measures of heterogeneity, while the \( \hat{R}_2 \) and \( \hat{R}_C \) statistics may not be recommended.

In the current literature, some meta-analysts depend solely on the magnitude of certain heterogeneity measures (usually \( I^2 \)) for assessing heterogeneity and choosing either the common-effect or random-effects model. For example, Myung et al.\(^{38} \) used the \( I^2 \) statistic “for the test of heterogeneity”; however, they reported only the magnitude of the \( I^2 \) statistic, without any information about the \( P \) value of \( I^2 \) or the \( Q \) test. This article has shown that the measures were not dramatically more powerful than the \( Q \) test; therefore, they may not yield statistically significant heterogeneity if \( Q \) cannot. It is untenable to assess heterogeneity based only on the measures’ magnitude. In addition, meta-analysts have been recommended to report the measures along with their confidence intervals, rather than simply using them as absolute measures of heterogeneity.\(^{16-49} \)

Despite that the heterogeneity measures may not greatly outperform the traditional \( Q \) test with respect to statistical power, the \( I^2 \) and \( \hat{R}_I \) statistics still provide valuable information about heterogeneity in a meta-analysis from the epidemiological or clinical perspective, and they can be easily understood by evidence users.


SUPPORTING INFORMATION

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