Received: 22 January 2019 | Revised: 8 April 2019 | Accepted: 9 April 2019

DOI: 10.1111/jep.13159

ORIGINAL PAPER

WILEY Journal of Evaluation in Clinical Practice
International Journal of Public Health Policy and Health Services Research

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, $\widehat{R}_I$, $\widehat{R}_M$, and $\widehat{R}_b$) 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 κ coefficient.

Results: Generally, $\widehat{R}_I$ 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 $\widehat{R}_M$ and $\widehat{R}_b$ statistics might have low powers in some cases. Because the differences between the powers of I^2, $\widehat{R}_I$, and Q were often tiny, meta-analysts might not expect I^2 and $\widehat{R}_I$ to yield significant heterogeneity if the Q test failed to do so. In addition, I^2 and $\widehat{R}_I$ 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 $\widehat{R}_I$ 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.[1,2] 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.[3-12] Of note, the specific meaning of heterogeneity may vary across different disciplines. For example, it could be the differences between contextures of materials or between strata in a geographical space.[13] In this article, heterogeneity in meta-analyses of clinical studies refers to the variation in the reported treatment effect estimates across the collected studies.[14] A traditional statistical method to detect heterogeneity in meta-analyses is the Q test,[15,16] but it is generally recognized to have low statistical power in many cases.[17] Also, it

 © 2019 John Wiley & Sons, Ltd.

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 $\widehat{R}_I$, $\widehat{R}_M$, and $\widehat{R}_b$ 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 $\frac{\tau^2}{\tau^2+\sigma^2}$, where τ^2 is the between-study variance caused by heterogeneity and σ^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 σ^2.[27,28] Each measure's true value is usually defined by replacing the estimated between-study variance $\widehat{\tau}^2$ with the true between-study variance τ^2 (see Table 1). For example, the true proportion of total variation caused by heterogeneity is calculated as $\frac{\tau^2}{\tau^2+n/\sum s_i^{-2}}$ when the $\widehat{R}_I$ statistic is used, while the true proportion is $\frac{\tau^2}{\tau^2+\sum s_i^2/n}$ for the $\widehat{R}_M$ statistic. Here, n is the number of studies in the meta-analysis, and s_i^2 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 κ 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 1 Heterogeneity measures interpreted as the proportion of total variation in a meta-analysis caused by heterogeneity rather than sampling error

Heterogeneity Measure	Expressed as a Function of $\widehat{\tau}^2$	Expressed as a Function of Q
I^2	$\frac{\widehat{\tau}^2}{\widehat{\tau}^2+\frac{(n-1)\sum s_i^{-2}}{(\sum s_i^{-2})^2-\sum s_i^{-4}}}$	$\frac{Q-(n-1)}{Q}$
$\widehat{R}_I$	$\frac{\widehat{\tau}^2}{\widehat{\tau}^2+\frac{n}{\sum s_i^{-2}}}$	$\frac{Q-(n-1)}{Q+1-n\sum s_i^{-4}/(\sum s_i^{-2})^2}$
$\widehat{R}_M$	$\frac{\widehat{\tau}^2}{\widehat{\tau}^2+\frac{1}{n}\sum s_i^2}$	$\frac{Q-(n-1)}{Q-(n-1)+\left[(\sum s_i^{-2})^2-\sum s_i^{-4}\right]\sum s_i^2/(n\sum s_i^{-2})}$
$\widehat{R}_b$	$\frac{1}{n}\sum\frac{\widehat{\tau}^2}{\widehat{\tau}^2+s_i^2}$	$\frac{1}{n}\sum\frac{Q-(n-1)}{Q-(n-1)+s_i^2\left[\left(\sum s_j^{-2}\right)^2-\sum s_j^{-4}\right]/\sum s_j^{-2}}$

Note. n, the number of studies in a meta-analysis; y_i and s_i^2, the observed effect size and its sample variance within study i; $\widehat{\tau}^2=\frac{Q-(n-1)}{\sum s_i^{-2}-\sum s_i^{-4}/\sum s_i^{-2}}$, the method-of-moments estimate of the between-study variance τ^2; $Q=\sum s_i^{-2}(y_i-\overline{\mu})^2$, where $\overline{\mu}=\frac{\sum y_i/s_i^2}{\sum 1/s_i^2}$. 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 variance s_i^2. The I^2 statistic was originated from the traditional Q statistic, which is defined as $Q = \sum s_i^{-2}(y_i - \bar{\mu})^2$, where $\bar{\mu} = \frac{\sum y_i/s_i^2}{\sum 1/s_i^2}$ is the pooled common-effect estimate of the overall treatment effect μ. The expectation of the Q statistic is as follows[30]:

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

where τ^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.[16] If the within-study variances are assumed to be equal, ie, $s_i^2 = \sigma^2$, then the expectation of Q becomes

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

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:

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

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 τ^2 by DerSimonian and Laird[31]:

$$\hat{\tau}^2 = \frac{Q - (n-1)}{\sum s_i^{-2} - \frac{\sum s_i^{-4}}{\sum s_i^{-2}}}. \quad (2)$$

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

$$I^2 = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \frac{(n-1)\sum s_i^{-2}}{\left(\sum s_i^{-2}\right)^2 - \sum s_i^{-4}}}. \quad (3)$$

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

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

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 $\hat{\tau}^2$ in Equation (3). The latter form is seldom used in the literature.

Unlike the I^2 statistic, the measures $\hat{R}_I$ and $\hat{R}_M$ were directly originated from certain summaries of within-study variances σ^2, instead of the Q statistic. The $\hat{R}_I$ statistic uses the harmonic mean of within-study variances to estimate σ^2, while the $\hat{R}_M$ statistic uses the arithmetic mean.[19-21] The $\hat{R}_I$ statistic was proposed by Takkouche et al[19] prior to the introduction of the I^2 statistic, although it has been used less frequently than I^2 so far. Specifically, the $\hat{R}_I$ statistic treats the summary of within-study variances as $\sigma^2 = \frac{n}{\sum s_i^{-2}}$, and the $\hat{R}_M$ uses $\sigma^2 = \frac{1}{n}\sum s_i^2$. Consequently, plugging them in the conceptual form of $\frac{\tau^2}{\tau^2 + \sigma^2}$ that describes the proportion of total variation caused by heterogeneity, the two heterogeneity measures are accordingly calculated as follows:

$$\hat{R}_I = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \frac{n}{\sum s_i^{-2}}} \text{ and } \hat{R}_M = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \frac{1}{n}\sum s_i^2}.$$

Using Equation (2) to express $\hat{\tau}^2$ 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 $\hat{\tau}^2$.

The $\hat{R}_b$ statistic was recently proposed by Crippa et al.[22] It uses the random-effects meta-analysis result to directly impute the total variation $\tau^2+\sigma^2$. Specifically, the $\hat{R}_b$ statistic is originally defined in terms of the estimated between-study variance:

$$\hat{R}_b = \frac{1}{n}\sum \frac{\hat{\tau}^2}{\hat{\tau}^2 + s_i^2}.$$

Note that $\frac{\hat{\tau}^2}{\hat{\tau}^2 + s_i^2}$ represents the proportion of heterogeneity in study i; therefore, the $\hat{R}_b$ 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 χ^2 distribution with $n - 1$ degrees of freedom under the null hypothesis of homogeneity (ie, $H_0 : \tau^2 = 0$), 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.[19,32,33]

The steps for the parametric resampling method are as follows. First, under the null hypothesis of homogeneity, we calculate the common-effect estimate as $\bar{\mu} = \frac{\sum y_i/s_i^2}{\sum 1/s_i^2}$ and denote the heterogeneity measures and the Q statistic in the original meta-analysis as $I^{2(0)}$, $\widehat{R}_I^{(0)}$, $\widehat{R}_M^{(0)}$, $\widehat{R}_b^{(0)}$, and $Q^{(0)}$, accordingly. Second, n within-study variances are sampled with replacement from those (ie, 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 $\bar{\mu}$ 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 $I^{2(k)}$, $\widehat{R}_I^{(k)}$, $\widehat{R}_M^{(k)}$, $\widehat{R}_b^{(k)}$, 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 I^2 statistic depends only on the Q statistic as in Equation (1), and it is an increasing function of Q. Therefore, $I^{2(k)} \geq I^{2(0)}$ 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[18] 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 10) studies, while such meta-analyses are common in practice[34]; therefore, such P values may not be accurate.

Because $\widehat{R}_I$, $\widehat{R}_M$, and $\widehat{R}_b$ 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[18] 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[32] proposed two alternative heterogeneity measures I_r^2 and I_m^2, 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.[12,32] In addition to these heterogeneity measures mostly used in meta-analyses of clinical studies, Wang et al[35] 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_r^2 and I_m^2 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[36] 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[37] 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[38] 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."[39]

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 χ^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

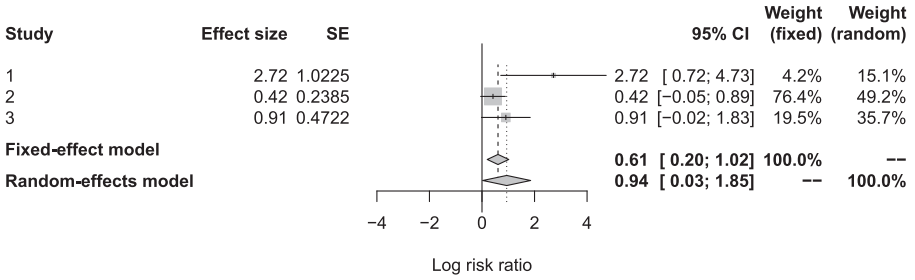

FIGURE 1 Forest plot of the meta-analysis by O'Doherty et al[36]

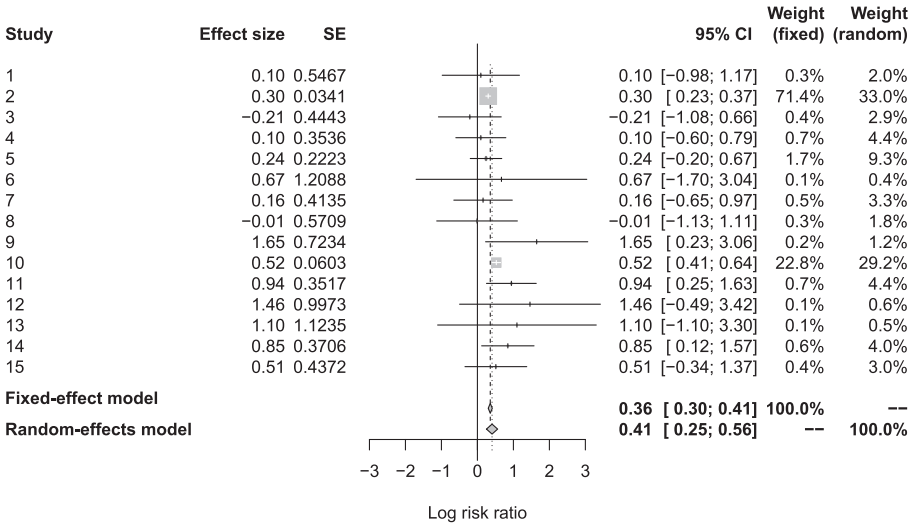

FIGURE 2 Forest plot of the meta-analysis by Makani et al[37]

the simulations were based on the meta-analysis by O'Doherty et al,[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 τ. Specifically, τ was from 0 to 2.5 by 0.5 for the simulations based on the meta-analysis by O'Doherty et al,[36] was from 0 to 0.75 by 0.15 for those based on the meta-analysis by Makani et al,[37] and was from 0 to 0.5 by 0.1 for those based on the meta-analysis by Myung et al.[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 κ coefficient to assess pairwise agreement among all four heterogeneity measures under each setting.[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 κ 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.[41(p604)]

3 | RESULTS

3.1 | Real-data examples

For the meta-analysis by O'Doherty et al,[36] the Q test led to $P_Q = 0.070$ (based on the theoretical χ^2 distribution) and 0.071 (based on the resampling method), and the heterogeneity measures were $I^2 = 62.3\%$, $\widehat{R}_I = 74.5\%$, $\widehat{R}_M = 46.3\%$, and $\widehat{R}_b = 58.9\%$ with P values $P_{I^2} = .071$, $P_{\widehat{R}_I} = .057$, $P_{\widehat{R}_M} = .105$, and $P_{\widehat{R}_b} = .079$, respectively. For the meta-analysis by Makani et al,[37] the Q test had $P_Q = 0.054$ (based on the theoretical χ^2 distribution) and 0.053 (based on the resampling method), and the heterogeneity measures were $I^2 = 40.2\%$, $\widehat{R}_I = 58.9\%$, $\widehat{R}_M = 4.4\%$, and $\widehat{R}_b = 19.0\%$ with P values $P_{I^2} = .053$, $P_{\widehat{R}_I} = .058$, $P_{\widehat{R}_M} = .138$, and $P_{\widehat{R}_b} = .120$. For the meta-analysis by Myung et al,[38] the Q test had a P value around .002 (based on both the theoretical χ^2 distribution and the resampling method), and the heterogeneity measures were $I^2 = 41.3\%$, $\widehat{R}_I = 42.8\%$, $\widehat{R}_M = 2.2\%$, and $\widehat{R}_b = 26.1\%$ with P values $P_{I^2} = .002$, $P_{\widehat{R}_I} = .002$, $P_{\widehat{R}_M} = .014$, and $P_{\widehat{R}_b} = .003$.

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

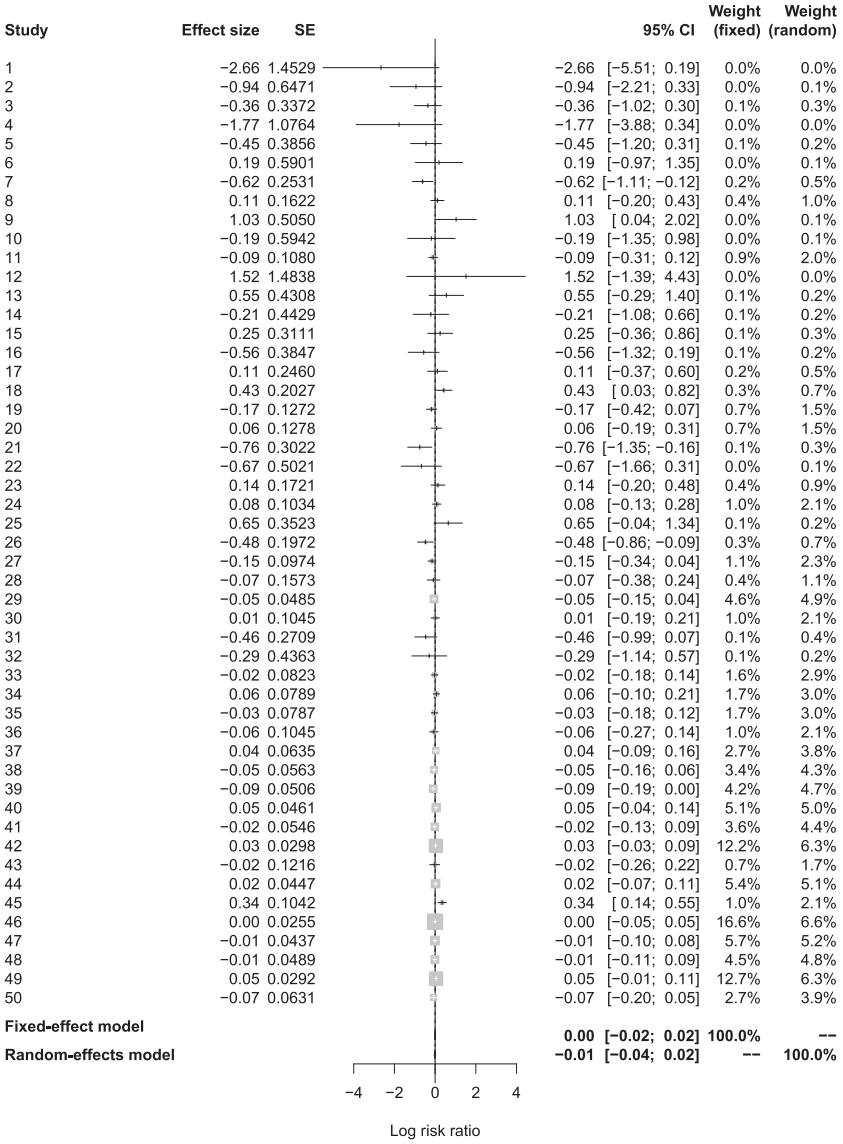

FIGURE 3 Forest plot of the meta-analysis by Myung et al[38]

nonignorable heterogeneity.[14] In the last two meta-analyses, the $\widehat{R}_M$ and $\widehat{R}_b$ statistics were noticeably smaller than I^2 and $\widehat{R}_I$. Their values (2.2% and 26.1%) seemed to seriously underestimate heterogeneity in the last meta-analysis by Myung et al,[38] 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 $\widehat{R}_I$ statistic slightly inflated (by from 0.5% to 1.4%). Moreover, the type I error rates of the $\widehat{R}_M$ and $\widehat{R}_b$ 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.[37]

As for statistical power, the I^2 statistic produced nearly the same results with the *Q* test. Generally, the $\widehat{R}_I$ 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 $\widehat{R}_M$ statistic had the lowest power among the four heterogeneity measures, and the $\widehat{R}_b$ statistic was also noticeably less powerful than the *Q* test and the I^2 and $\widehat{R}_I$ statistics. Under the setting based on the meta-analysis by Makani et al[37] with $\tau = 0.3$, the powers of the *Q* test and the I^2 statistic were around 53.6%, slightly less than the power of the $\widehat{R}_I$ statistic, 55.9%. The power of the $\widehat{R}_M$ statistic was merely 43.3%, and that of the $\widehat{R}_b$ statistic was 49.3%. The $\widehat{R}_M$ and $\widehat{R}_b$ statistics were underpowered likely because their type I error rates were too low in some cases, so these measures might be too conservative for assessing heterogeneity.

Table 3 presents Cohen's κ coefficients in the simulated meta-analyses. Under the setting based on the meta-analysis by O'Doherty et al[36] containing only three studies, all four heterogeneity measures generally had excellent agreement because nearly all κ coefficients were at least 0.75. However, for larger simulated meta-analyses

TABLE 2 Type I error rates and statistical powers (in percentage, %) of the heterogeneity measures and the Q test in the simulation studies

τ	Q (Based on χ^2)	Q (Based on resampling)	I^2	$\widehat{R}_I$	$\widehat{R}_M$	$\widehat{R}_b$
Simulations based on the meta-analysis by O'Doherty et al[36] with three studies:						
0	5.2	5.1	5.1	5.8	4.5	5.0
0.5	18.0	17.9	17.9	19.4	16.7	17.6
1.0	45.3	45.2	45.2	46.7	43.7	44.9
1.5	64.9	64.7	64.7	66.1	63.8	64.5
2.0	76.7	76.6	76.6	77.6	75.7	76.4
2.5	83.8	83.8	83.8	84.4	83.3	83.8
Simulations based on the meta-analysis by Makani et al[37] with 15 studies:						
0	5.1	5.1	5.1	6.4	1.9	3.2
0.15	22.3	22.1	22.1	23.8	12.2	17.2
0.30	53.7	53.6	53.6	55.9	43.3	49.3
0.45	76.5	76.6	76.6	78.1	69.5	73.3
0.60	89.6	89.5	89.5	90.3	85.0	87.4
0.75	95.2	95.2	95.2	95.7	93.0	94.3
Simulations based on the meta-analysis by Myung et al[38] with 50 studies:						
0	4.9	4.9	4.9	5.5	2.1	3.3
0.1	21.6	21.5	21.5	22.6	13.8	17.7
0.2	63.7	63.4	63.4	64.8	54.1	59.1
0.3	88.1	88.2	88.2	88.8	83.9	86.1
0.4	96.7	96.7	96.7	97.0	95.1	96.0
0.5	99.1	99.1	99.1	99.2	98.5	98.8

Note. τ, the between-study standard deviation. The results corresponding to $\tau = 0$ are type I error rates, and those corresponding to $\tau > 0$ are statistical powers.

containing 15 studies based on that by Makani et al,[37] all κ coefficients dramatically decreased. Only the pair of I^2 and $\widehat{R}_I$ had good agreement. All remaining five pairs generally had fairly poor agreement; many κ coefficients were close to zero. When the simulated meta-analyses became even larger with 50 studies as in the one by Myung et al,[38] the agreement between I^2 and $\widehat{R}_I$ remained good with κ coefficients around 0.7, while κ coefficients of other five pairs of heterogeneity measures became closer to zero, indicating poor agreement.

4 | DISCUSSION

4.1 | Main findings

This article has reviewed four measures that can be used to quantify heterogeneity in meta-analyses of clinical studies. The $\widehat{R}_I$ 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 $\widehat{R}_b$ statistic was noticeably lower than that of the foregoing

TABLE 3 Cohen's κ coefficient of each pair of the four heterogeneity measures based on the simulated meta-analyses

τ	I^2 vs $\widehat{R}_I$	I^2 vs $\widehat{R}_M$	I^2 vs $\widehat{R}_b$	$\widehat{R}_I$ vs $\widehat{R}_M$	$\widehat{R}_I$ vs $\widehat{R}_b$	$\widehat{R}_M$ vs $\widehat{R}_b$
Simulations based on the meta-analysis by O'Doherty et al[36] with three studies:						
0	0.857	0.893	0.961	0.751	0.876	0.875
0.5	0.862	0.878	0.949	0.742	0.859	0.883
1.0	0.870	0.886	0.953	0.757	0.850	0.907
1.5	0.885	0.891	0.954	0.777	0.859	0.918
2.0	0.886	0.908	0.963	0.796	0.861	0.935
2.5	0.903	0.893	0.956	0.797	0.876	0.922
Simulations based on the meta-analysis by Makani et al[37] with 15 studies:						
0	0.636	0.061	0.477	0.034	0.265	0.210
0.15	0.618	0.033	0.412	0.022	0.231	0.132
0.30	0.597	0.030	0.264	0.023	0.138	0.220
0.45	0.625	−0.028	0.113	−0.013	0.043	0.244
0.60	0.632	−0.059	0.039	−0.038	−0.008	0.321
0.75	0.654	−0.060	0.040	−0.034	0.010	0.398
Simulations based on the meta-analysis by Myung et al[38] with 50 studies:						
0	0.664	0.000	0.010	0.000	0.005	0.000
0.1	0.735	0.000	0.016	0.000	0.008	0.000
0.2	0.703	0.000	0.010	0.000	−0.010	0.012
0.3	0.679	−0.008	−0.085	−0.006	−0.072	0.058
0.4	0.687	−0.028	−0.092	−0.018	−0.067	0.079
0.5	0.700	−0.033	−0.088	−0.022	−0.065	0.027

Note. τ, the between-study standard deviation.

measures, and the $\widehat{R}_M$ 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 $\widehat{R}_I$ 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 variances, so their true values are different, 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.

Our study had several limitations. First, we restricted our comparisons to be among the four heterogeneity measures with the same interpretation for generic meta-analyses, while several other measures briefly reviewed in Section 2.3 are also good options for different purposes. The heterogeneity measures may be selected on a case-by-case basis with considerations of many factors (eg, the presence of outliers). Second, the different forms of the four heterogeneity measures in terms of $\hat{\tau}^2$ and Q in Table 1 were based on the method-of-moments estimate of the between-study variance $\hat{\tau}^2$. Originally, I^2 was derived based on Q, while the other three measures were based directly on $\hat{\tau}^2$. This article used the method-of-moments estimate mainly because it yielded the one-to-one relationship between the two forms in terms of $\hat{\tau}^2$ and Q for each measure; otherwise, such relationship may not hold for other estimates of τ^2. However, although the method-of-moments estimate has been popular so far, it has been found to be inferior in some cases.[42] Alternative estimates, such as those based on the restricted maximum likelihood or Bayesian analysis, may be better options to be used in the heterogeneity measures.[43,44] Third, as in conventional meta-analysis methods, the within-study variances were treated as true values in our analysis and the four heterogeneity measures. However, they were actually estimates and were subject to sampling error; these variance estimates could be poor if the corresponding studies had small sample sizes.[16,27,45] Future studies are highly needed to effectively account for such sampling errors in within-study variances and to quantify heterogeneity more accurately.

4.3 | Recommendations

In summary, based on their statistical powers and agreement, both the I^2 and $\widehat{R}_I$ statistics may be preferred measures of heterogeneity, while the $\widehat{R}_M$ and $\widehat{R}_b$ 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.[46-49]

Despite that the heterogeneity measures may not greatly outperform the traditional Q test with respect to statistical power, the I^2 and $\widehat{R}_I$ 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.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

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

REFERENCES

1. Berlin JA, Golub RM. Meta-analysis as evidence: building a better pyramid. *JAMA*. 2014;312(6):603-606.
2. Gurevitch J, Koricheva J, Nakagawa S, Stewart G. Meta-analysis and the science of research synthesis. *Nature*. 2018;555(7695):175-182.
3. Thompson SG, Pocock SJ. Can meta-analyses be trusted? *The Lancet*. 1991;338(8775):1127-1130.
4. Eysenck HJ. Meta-analysis and its problems. *BMJ*. 1994;309(6957):789-792.
5. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. *BMJ*. 1994;309(6965):1351-1355.
6. Hardy RJ, Thompson SG. Detecting and describing heterogeneity in meta-analysis. *Stat Med*. 1998;17(8):841-856.
7. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. *Stat Med*. 1999;18(20):2693-2708.
8. Engels EA, Schmid CH, Terrin N, Olkin I, Lau J. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. *Stat Med*. 2000;19(13):1707-1728.
9. Higgins JPT. Commentary: heterogeneity in meta-analysis should be expected and appropriately quantified. *Int J Epidemiol*. 2008;37(5):1158-1160.
10. Ioannidis JPA. Interpretation of tests of heterogeneity and bias in meta-analysis. *J Eval Clin Pract*. 2008;14(5):951-957.
11. Ioannidis JPA, Patsopoulos NA, Rothstein HR. Reasons or excuses for avoiding meta-analysis in forest plots. *BMJ*. 2008;336(7658):1413-1415.
12. Ma X, Lin L, Qu Z, Zhu M, Chu H. Performance of between-study heterogeneity measures in the Cochrane library. *Epidemiology*. 2018;29(6):821-824.
13. Dutilleul PRL. *Spatio-Temporal Heterogeneity: Concepts and Analyses*. Cambridge, UK: Cambridge University Press; 2011.
14. Higgins JPT, Green S. *Cochrane Handbook for Systematic Reviews of Interventions*. Chichester, UK: John Wiley & Sons; 2008.
15. Cochran WG. The combination of estimates from different experiments. *Biometrics*. 1954;10(1):101-129.
16. Hoaglin DC. Misunderstandings about Q and 'Cochran's Q test' in meta-analysis. *Stat Med*. 2016;35(4):485-495.
17. Jackson D. The power of the standard test for the presence of heterogeneity in meta-analysis. *Stat Med*. 2006;25(15):2688-2699.
18. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med*. 2002;21(11):1539-1558.
19. Takkouche B, Cadarso-Suárez C, Spiegelman D. Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis. *Am J Epidemiol*. 1999;150(2):206-215.
20. Takkouche B, Khudyakov P, Costa-Bouzas J, Spiegelman D. Confidence intervals for heterogeneity measures in meta-analysis. *Am J Epidemiol*. 2013;178(6):993-1004.
21. Xiong C, Miller JP, Morris JC. Measuring study-specific heterogeneity in meta-analysis: application to an antecedent biomarker study of Alzheimer's disease. *Stat Biopharm Res*. 2010;2(3):300-309.

22. Crippa A, Khudyakov P, Wang M, Orsini N, Spiegelman D. A new measure of between-studies heterogeneity in meta-analysis. *Stat Med*. 2016;35(21):3661-3675.
23. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 7. Rating the quality of evidence—inconsistency. *J Clin Epidemiol*. 2011;64(12):1294-1302.
24. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ*. 2003;327(7414):557-560.
25. Mittlböck M, Heinzl H. A simulation study comparing properties of heterogeneity measures in meta-analyses. *Stat Med*. 2006;25(24):4321-4333.
26. von Hippel PT. The heterogeneity statistic I^2 can be biased in small meta-analyses. *BMC Med Res Methodol*. 2015;15(1):35.
27. Hoaglin DC. Practical challenges of I^2 as a measure of heterogeneity. *Res Synth Methods*. 2017;8(3):254.
28. Böhning D, Lerdsuwansri R, Holling H. Some general points on the I^2-measure of heterogeneity in meta-analysis. *Metrika*. 2017;80:1-11.
29. Kvålseth TO. Cautionary note about R^2. *Am Stat*. 1985;39(4):279-285.
30. Biggerstaff BJ, Tweedie RL. Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Stat Med*. 1997;16(7):753-768.
31. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials*. 1986;7(3):177-188.
32. 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.
33. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Boca Raton, FL: CRC Press; 1998.
34. 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(1):160.
35. Wang J-F, Zhang T-L, Fu B-J. A measure of spatial stratified heterogeneity. *Ecol Indic*. 2016;67:250-256.
36. O'Doherty LJ, Taft A, Hegarty K, Ramsay J, Davidson LL, Feder G. Screening women for intimate partner violence in healthcare settings: abridged Cochrane systematic review and meta-analysis. *BMJ*. 2014;348(may12 1):g2913.
37. Makani H, Bangalore S, Desouza KA, Shah A, Messerli FH. Efficacy and safety of dual blockade of the renin-angiotensin system: meta-analysis of randomised trials. *BMJ*. 2013;346(jan28 1):f360.
38. Myung S-K, Ju W, Cho B, et al. Efficacy of vitamin and antioxidant supplements in prevention of cardiovascular disease: systematic review and meta-analysis of randomised controlled trials. *BMJ*. 2013;346(jan18 1):f10.
39. Schwarzer G. Meta: an R package for meta-analysis. *R News*. 2007;7(3):40-45.
40. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;20(1):37-46.
41. Fleiss JL, Levin B, Paik MC. *Statistical Methods for Rates and Proportions*. 3rd ed. Hoboken, NJ: John Wiley & Sons; 2003.
42. Cornell JE, Mulrow CD, Localio R, et al. Random-effects meta-analysis of inconsistent effects: a time for change. *Ann Intern Med*. 2014;160(4):267-270.
43. 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.
44. Turner RM, Davey J, Clarke MJ, Thompson SG, Higgins JPT. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane database of systematic reviews. *Int J Epidemiol*. 2012;41(3):818-827.
45. Lin L. Bias caused by sampling error in meta-analysis with small sample sizes. *PLoS One*. 2018;13(9):e0204056.
46. Rücker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I^2 in assessing heterogeneity may mislead. *BMC Med Res Methodol*. 2008;8(1):79.
47. Borenstein M, Higgins JPT, Hedges LV, Rothstein HR. Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. *Res Synth Methods*. 2017;8(1):5-18.
48. Thorlund K, Imberger G, Johnston BC, et al. Evolution of heterogeneity (I^2) estimates and their 95% confidence intervals in large meta-analyses. *PLoS ONE*. 2012;7(7):e39471.
49. Ioannidis JPA, Patsopoulos NA, Evangelou E. Uncertainty in heterogeneity estimates in meta-analyses. *BMJ*. 2007;335(7626):914-916.

SUPPORTING INFORMATION

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

How to cite this article: Lin L. Comparison of four heterogeneity measures for meta-analysis. *J Eval Clin Pract*. 2019;1–9. https://doi.org/10.1111/jep.13159