Comparison of Heterogeneity Measures in Meta-Analysis

Ozlem Toluk[1,2,*], and Ilker Ercan[3]

[1]*Faculty of Medicine, Department of Biostatistics and Medical Informatics, Bezmialem Vakif University, Istanbul, Türkiye*

[2]*Institute of Health Sciences and* [3]*Faculty of Medicine, Department of Biostatistics, Bursa Uludag University, Bursa, Türkiye*

Abstract: *Background*: Heterogeneity assessment is critical in meta-analysis, as it determines the appropriateness of combining studies and affects result reliability. Cochran's Q is the traditional test, nevertheless, it has low statistical power, so many researchers resort to using heterogeneity measures to quantify the heterogeneity.

Aim: This article aims to compare the performance of the most commonly used heterogeneity measures through simulation.

Materials and Methods: We compared the performance of four heterogeneity measures (τ^2, I^2, R_b, H) across various homogeneous and heterogeneous patient-event probabilities [$\mathrm{P(P^-|E^+)}$ and $\mathrm{P(P^+|E^+)}$], various sample sizes (n) and number of studies (k), using RMSE (Root mean squared error) and BIAS values in simulation scenarios. Additionally, Cochran's Q Type-I error rate and power were evaluated using the same simulation scenarios.

Results: τ^2 and H outperformed other measures in large samples, while I^2, and R_b were preferable for small studies.

Conclusion: Researchers can use the simulation results from this study to select an appropriate heterogeneity measure for their meta-analysis work. This approach is expected to prevent time loss due to unnecessary subgroup analyses in situations where heterogeneity appears to be present but is actually absent.

Keywords: Meta Analysis, I^2 heterogeneity measure, R_b heterogeneity measure, H heterogeneity measure, Tau^2 heterogeneity measure, simulation.

1. INTRODUCTION

A science that quantitatively deals with changing observations began to emerge in the 17th century [1]. British statistician Karl Pearson was the first to apply methods to integrate observations from clinical trials. More than one study is often conducted to understand and answer important and difficult questions. In some cases, clinical decision-making becomes difficult because the results obtained vary from study to study. The need to reach decisions that affect clinical practice increases the importance of "evidence-based medicine" [2]. Evidence-based medicine can be defined as a systematic, quantitative, and preferably experimental approach to obtaining and using medical knowledge, aiming to find the best research evidence by combining clinical and patient experience [2, 3]. Systematic reviews and meta-analyses are the primary tools used to synthesize the findings needed to inform the clinical decision process, and meta-analyses are at the center because they combine the results of multiple studies and reach a general conclusion [3, 4].

Many studies have potentially different characteristics and were conducted by different research teams with different methods, so there are differences across studies and they are often expected to exhibit some degree of heterogeneity [5]. A common method for assessing whether true heterogeneity exists in a meta-analysis study is to use the Q test, a statistical test described by Cochran in 1954. The shortcoming of the Q statistic is that when the meta-analysis includes a small number of studies, the Q statistic has little power to detect true heterogeneity among studies, and when it includes a large number of studies, it has excessive power to detect negligible variability. Heterogeneity measures are suggested to overcome the shortcomings of the Q test [6].

The most commonly used measure of heterogeneity, I^2, estimates the proportion of variability in a meta-analysis that is explained by differences between included experiments rather than by sampling error. However, some studies reveal important shortcomings of the I^2 measure. Especially in meta-analyses involving a small number of samples (e.g. n<10), I^2 estimates may be unreliable. Furthermore, I^2 maybe underestimated due to time lag bias [7, 8]. Incorrect estimation of heterogeneity prevents the investigation of the causes of heterogeneity, while overestimation may lead to unnecessary examination of the causes of heterogeneity by preventing meta-analysis. Large I^2 estimates may lead authors to try all possibilities in subgroup analyses [9]. Depending on

*Address correspondence to this author at the Faculty of Medicine, Department of Biostatistics and Medical Informatics, Bezmialem Vakif University, Istanbul, Türkiye; Institute of Health Sciences, Department of Biostatistics, Bursa Uludag University, Bursa, Türkiye; Tel: +905306476966,
E-mail: erozlem12@gmail.com

the conditions, when the number of studies are small, the bias of I^2 is high [10].

Lack of comparative simulation studies of commonly used heterogeneity measures, our study aims to compare performance of them with simulation sudy. Additionally, sought to determine in which simulation scenario the heterogeneity measures are appropriate to use.

2. SCIENTIFIC BACKGROUND

Statistical heterogeneity in meta-analysis is related to the variation between studies. This variation is due to clinical or methodological differences between studies or simply randomization. The increased variance value due to heterogeneity is directly related to the heterogeneity test and heterogeneity measurements.

2.1. Heterogeneity Test with Cochran's Q Statistic

To evaluate the true heterogeneity among studies, Cochran proposed the Q statistic, also called the Chi-square heterogeneity test, which fits the χ^2 distribution with (k-1) degrees of freedom, in 1954. The Q test statistic is expressed by the following equation;

i=1, 2, 3,….k

M: weighted average of observed effect sizes

Y_i: i. the observed effect size of the study

$$Q = \sum_{i=1}^{k} w_i (Y_i - M)^2 \quad (1)$$

$$M = \frac{\sum_{i=1}^{k} w_i Y_i}{\sum_{i=1}^{k} w_i} \quad (2)$$

Since the power of Cochran's Q test is related to the number of studies included in the meta-analysis, the power of the test is low when the number of studies (k < 20) is high when the number of studies is high [11]. To eliminate this problem, heterogeneity measures should also be calculated [12].

2.2. Heterogeneity Measures

The most frequently used criteria in the literature to determine the amount of heterogeneity are H^2, R^2, τ^2, I^2 and R_b [12] in meta-analysis.

2.2.1. τ^2 Measure

The τ^2 criterion represents the variance between studies and the DerSimonian Laird method is used for its estimation. It is divided by a quantity (C) which has the effect of restoring the criterion to its original metric and turning it into an average rather than the sum of the squares of the deviations [13].

$$\tau^2 = \frac{Q-(k-1)}{C} \quad (3)$$

$$C = \sum_{i}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i} \quad (4)$$

2.2.2. H Measure

The H measure proposed by Higgins and Thomson in 2002 is given with the help of Q statistics in the following equation [14];

$$H^2 = \begin{cases} \frac{Q}{k-1}, & Q > (k-1) \\ 1, & Q \leq (k-1) \end{cases} \quad (5)$$

H^2 takes values between 1 and ∞. H=1 indicates perfect homogeneity. The H value increases depending on the number of studies [12, 15].

2.2.3. R Measure

Like the H criterion, it depends on the number of studies to be included in the meta-analysis and the τ^2 criterion is used in its calculation [14]. R^2 is calculated by considering the special case where the sampling variances of the estimates from each run are known and equal, that is, $1/\sum_{i=1}^{k} w_i = \sigma^2$ for all i [14].

$$R^2 = \frac{\tau^2+\sigma^2}{\sigma^2} \quad (6)$$

$$R = \sqrt{\frac{\sum_{i=1}^{k} w_i}{\sum_{i=1}^{k} w_i^*}} = \sqrt{\frac{\sum_{i=1}^{k} w_i}{\sum_{i=1}^{k} (w_i^{-1}+\hat{\tau}^2)^{-1}}} \quad (7)$$

If R = 1, homogeneity is perfect. When all estimates have equal precision, H and R coincide [14].

2.2.4. I^2 Measure

Using Cochran's Q and H^2 criteria, Higgins and Thomson proposed the I^2 criterion in 2002. It can be obtained with different calculations as seen in the equations below [15].

$$I^2 = \begin{cases} \frac{Q-(k-1)}{Q}, & Q > (k-1) \\ 0, & Q \leq (k-1) \end{cases} \quad (8)$$

$$I^2 = \begin{cases} \frac{H^2-1}{H^2}.100 & Q > (k-1) \\ 0, & Q \leq (k-1) \end{cases} \quad (9)$$

$$I^2 = \begin{cases} \frac{c\tau^2-1}{Q}.100 & Q > (k-1) \\ 0, & Q \leq (k-1) \end{cases} \quad (10)$$

Heterogeneity varies between 0 and 100%, and when it takes values close to 100%, it is considered that heterogeneity is high, and when it takes values close to zero, it is considered that heterogeneity is low.

2.2.5. R_b Measure

The R_b measure quantifies the contribution of τ^2 relative to the variance of the pooled random effects estimate. The R_b measure estimates the expected value of the proportion of total variance due to variation across studies [16]. R_b=1 indicates maximum heterogeneity [17].

$R_b = \frac{1}{k}\sum_{i=1}^{k}\frac{\tau^2}{\tau^2+\hat{\tau}^2}$ is calculated with equality.

3. MATERIALS AND METHODS

3.1. Simulation Scenarios

In the context of simulation studies based on the binomial distribution, the control group (P^-), hypothetical populations were generated to reflect the probability of being disease-free conditional on the occurrence of the event (E^+) with $\mathrm{P}(\mathrm{P}^-|\mathrm{E}^+)$=0.5 and $\mathrm{N_C}$=1,000,000 ($\mathrm{N_C}$=control group population size). For the patient group (P^+) hypothetical populations were generated to represent the probability of having disease when the presence of the event (E^+) with $\mathrm{P}(\mathrm{P}^+|\mathrm{E}^+)$ =0.5, 0.6, 0.7, 0.8, 0.9) and $\mathrm{N_P}$=1,000,000 (for each patient group population size). From each hypothetical population, the sample sizes $\mathrm{n_P}$=$\mathrm{n_C}$=8, 12, 25, 50, 100; the number of studies k=3, 6, 12, 24, 48 were generated. The simulation study was performed by taking 1,000 repetitions. Meta-analysis was performed for each repetition individually. The performances of the τ^2, I^2, R_b, H heterogeneity criteria obtained through the meta-analysis were examined with RMSE and BIAS values. Van Houwelingen, Zwinderman [18] stated that the Mantel Haenszel method can also be used for the random effects model when the general parameter is OR or log (OR). In our study, the Mantel Haenszel method was conducted in the simulation scenarios for homogeneous and heterogeneous studies. According to Higgins, Thompson [19] study, level of significance in the meta-analysis were derived as homogeneous and heterogeneous under Cochran's Q test and α=0.10 was taken [19].

In our study, since the Type- I error rate of the simulation results was taken as α=0.10, it was determined as robust if it was between 0.09 – 0.11 (α ± 0.1α) and as moderately robust if it was between 0.075 – 0.125 (α ± 0.25α). In both cases, their Type-I error protection performance is considered sufficient [20]. It is stated that they exhibit a conservative attitude for the tendency to estimate the Type-I error below α=0.10 and a liberal attitude for the tendency to estimate it above α=0.10 [21]. Analyses were performed with the "metafor" and "meta" packages in R-Studio 2023.12.0 (R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/).

3.1.1. Simulation Scenario of Homogeneous Studies

In the simulation scenario, hypothetical populations N_C=1,000,000 with probability $\mathrm{P}(P^-|E^+)$=0.5 for the control group and N_P=1,000,000 with the probability of having the disease when the presence of event $\mathrm{P}(\mathrm{P}^+|\mathrm{E}^+)$=(0.5, 0.6, 0.7, 0.8, 0.9) were created for homogeneous studies. By drawing random samples of $\mathrm{n_P}$=$\mathrm{n_C}$=8, 12, 25, 50, 100 from the created populations, the RMSE and BIAS values of the heterogeneity measures τ^2, I^2, $\mathrm{R_b}$ and H were calculated with the numbers of studies k=3, 6, 12, 24, 48. RMSE and BIAS values of heterogeneity measures and Cochran's Q statistics Type-I error rates and OR values of the studies are presented in the tables.

3.1.2. Simulation Scenarios of Heterogeneous Studies

High heterogeneity was achieved by taking probabilities of $\mathrm{P}(P^-|E^+)$=0.5 for the control group and $\mathrm{P}(P^+|E^+)$=0.6, 0.7, 0.8, 0.9 for the patient group. k=3, 6, 12, 24, 48 the number of studies and n_P=n_C=8, 12, 25, 50, 100 sample sizes were taken from each hypothetical population individually. For example, when the number of studies was taken as 3, high heterogeneity was achieved by selecting the control group $\mathrm{P}(P^-|E^+)$=0.5 for all three studies and the patient group taken as follows; 1st study $\mathrm{P}(P^+|E^+)$ =0.60, 2nd study $\mathrm{P}(P^+|E^+)$=0.70 and 3rd study $\mathrm{P}(P^+|E^+)$=0.80. When the number of studies was taken as k=8, high heterogeneity was provided by taking the probabilities of control group $\mathrm{P}(P^-|E^+)$=0.5 for each study, patient group as $\mathrm{P}(P^+|E^+)$=0.60 for the 1st study, $\mathrm{P}(P^+|E^+)$=0.70 for the 2nd study, $\mathrm{P}(P^+|E^+)$=0.80 for the 3rd study, $\mathrm{P}(P^+|E^+)$=0.90 for the 4th study, $\mathrm{P}(P^+|E^+)$=0.60 for the 5th study, $\mathrm{P}(P^+|E^+)$=0.70 for the 6th study, $\mathrm{P}(P^+|E^+)$=0.80 for the 7th study and $\mathrm{P}(P^+|E^+)$=0.90 for the 8th study. Moderate heterogeneity was achieved by taking probabilities of $\mathrm{P}(P^-|E^+)$=0.5 for the control group and $\mathrm{P}(P^+|E^+)$=0.6, 0.7, 0.8, 0.8 for the patient group. Low heterogeneity was achieved by taking probabilities of P$(P^{-}|E^{+})$=0.5 for the control group and P$(P^{+}|E^{+})$=0.6, 0.7, 0.7, 0.7 for the patient group. As the number of studies in the simulations increased, the probability values were increased sequentially and the number of studies were completed. RMSE and BIAS values of the heterogeneity measures τ^2, I^2, R_b and H were calculated for simulation scenarios of heterogeneous studies. RMSE and BIAS values of heterogeneity measures, OR, and Cochran's Q statistics power of the studies were presented in Tables **6-8**.

4. RESULTS

4.1. Simulation Results of Homogeneous Studies

The results of the simulations conducted for scenarios where the studies included in the meta-analysis were homogeneous were presented in Tables **1-5**. When the number of studies was held constant for P$(P^{-}|E^{+})$=0.5 and P$(P^{+}|E^{+})$=0.5, the RMSE and BIAS values of the heterogeneity measures were examined according to the sample sizes. When n increased, and k<12, I^2, R_b and H estimations converged toward each other. H criterion produced estimates closer to the population parameter than the I^2 and R_b when k≥12. When n=25, 50, 100, the τ^2 measure performed the best performance by producing the closest estimate to the parameter. Heterogeneity measures produced estimates above the population parameter as n increased when k was held constant. When the sample size was kept constant, all criteria produced values close to the parameter as k increased. While τ^2 exhibited suboptimal performance at n<25 and k≤12, I^2, R_b and H demonstrated a closely aligned trend. When n=25, I^2 yielded the worst estimation. R_b, and I^2 estimations were highly similar. When n>12 and k>6, the H demonstrated the second-highest performance, after τ^2. When n=8, and n=12, I^2, R_b and H produced

Table 1: Heterogeneity Measures' Simulation Results of P$(P^{-}|E^{+})$ = 0.5 vs $P(P^{+}|E^{+})$ = 0.5

k	n_P=n_K	P$(P^{-}\|E^{+})$=0.5 vs $P(P^{+}\|E^{+})$=0.5 (OR=1.00)								Cochran's Q Type-I error
		RMSE				BIAS				
		τ^2	I^2	H	R_b	τ^2	I^2	H	R_b	
3	8	0.9852	0.2577	0.2523	0.2561	0.3904	0.1367	0.1192	0.1314	0.073
	12	1.2000	0.4194	0.3771	0.4219	-1.0090	-0.3494	-0.2810	-0.3517	0.092
	25	0.2901	0.2723	0.2821	0.2724	0.1199	0.1459	0.1329	0.1451	0.096
	50	0.1491	0.2749	0.3014	0.2750	0.0610	0.1456	0.1376	0.1453	0.101
	100	0.0699	0.2864	0.3027	0.2865	0.0312	0.1540	0.1454	0.1538	0.108
6	8	0.4502	0.1916	0.1495	0.1786	0.1833	0.1018	0.0732	0.0863	0.053
	12	0.4112	0.2750	0.2039	0.2251	0.2355	0.2037	0.1252	0.1319	0.086
	25	0.1795	0.2310	0.1981	0.2284	0.0807	0.1290	0.1001	0.1247	0.099
	50	0.0811	0.2323	0.1960	0.2313	0.0393	0.1303	0.1007	0.1286	0.101
	100	0.0439	0.2432	0.2117	0.2428	0.0212	0.1369	0.1086	0.1362	0.113
12	8	0.2403	0.1442	0.1001	0.1191	0.0896	0.0736	0.0479	0.0504	0.048
	12	0.2019	0.1723	0.1244	0.1534	0.0867	0.0942	0.0637	0.0741	0.073
	25	0.1100	0.1920	0.1420	0.1849	0.0549	0.1118	0.0770	0.1038	0.099
	50	0.0507	0.1828	0.1340	0.1796	0.0252	0.1027	0.0703	0.0989	0.089
	100	0.0250	0.1802	0.1328	0.1787	0.0124	0.1005	0.0688	0.0987	0.084
24	8	0.1071	0.0972	0.0609	0.0650	0.0331	0.0458	0.0275	0.0218	0.029
	12	0.1026	0.1245	0.0799	0.0973	0.0427	0.0687	0.0420	0.0437	0.063
	25	0.0644	0.1416	0.0935	0.1302	0.0314	0.0796	0.0500	0.0684	0.089
	50	0.0359	0.1514	0.1013	0.1464	0.0185	0.0869	0.0553	0.0814	0.116
	100	0.0174	0.1461	0.0978	0.1438	0.0089	0.0814	0.0517	0.0790	0.098
48	8	0.0408	0.0592	0.0345	0.0283	0.0086	0.0248	0.0139	0.0063	0.013
	12	0.0542	0.0769	0.0470	0.0574	0.0187	0.0152	0.0108	0.0209	0.050
	25	0.0423	0.1109	0.0682	0.0951	0.0205	0.0641	0.0378	0.0491	0.086
	50	0.0231	0.1105	0.0689	0.1038	0.0114	0.0618	0.0368	0.0550	0.089
	100	0.0119	0.1127	0.0699	0.1095	0.0062	0.0641	0.0381	0.0609	0.107

estimates close to the parameter in all number of studies. Heterogeneity measures produced estimates above the population parameter according to k when the sample size was held constant. In general, when k≤8 and n≥12 were taken, Cochran's Q Type- I error rates of the simulation scenarios followed a liberal course and could be preserved. When k>8 and n>12 were taken, the conservative Cochran's Q Type-I error rates were found to be moderately robust (Table **1**).

The number of studies was kept constant for P($P^-|E^+$)=0.5 and P($P^+|E^+$)=0.6, the RMSE and BIAS values of the heterogeneity measures were examined according to the sample size. However, τ^2 performed the poorest performance with small n, while sample sizes of n=25, 50, and 100 yielded the best results. When k≤12, I^2, R_b and H produced estimates that were very close to each other and tended to overestimate the population parameter. When k>6, H demonstrated the best performance after τ^2. In general, the criteria were consistently lower than the parameter value. When the n was held constant, all criteria produced increasingly similar estimates to each other, and the population parameter as k increased. When n<25, criterion H provided the most accurate estimate of the parameter. For n≥25, as the number of studies increased, τ^2 continued to yield the most accurate estimates, while H produced estimates closer to the population parameter compared to I^2 and R_b. Overall, when n was held constant and k increased, the criteria tended to overestimate the population parameter. Overall, when k=3 and n>25, the Type I error rates of

Table 2: Heterogeneity Measures' Simulation Results of P($P^-|E^+$) = 0.5 vs $P(P^+|E^+)$ = 0.6

k	n_P=n_K	P($P^-\|E^+$)=0.5 vs $P(P^+\|E^+)$=0.6 (OR=1.50)								Cochran's Q Type-I error
		RMSE				BIAS				
		τ^2	I^2	H	R_b	τ^2	I^2	H	R_b	
3	8	0.9101	0.2558	0.2434	0.2546	0.3815	0.1369	0.1173	0.1315	0.069
	12	1.3395	0.4739	0.4408	0.4805	-1.1939	-0.4189	-0.3742	-0.4259	0.076
	25	0.3945	0.3736	0.3251	0.3738	-0.2877	-0.2867	-0.1942	-0.2867	0.107
	50	0.2069	0.3438	0.3047	0.3431	-0.1544	-0.2421	-0.1433	-0.2409	0.115
	100	0.2108	0.4883	0.4714	0.4897	-0.2026	-0.4317	-0.4028	-0.4333	0.088
6	8	0.7016	0.2256	0.1986	0.2203	0.2833	0.1175	0.0938	0.1092	0.067
	12	0.4640	0.2127	0.1969	0.2101	0.0025	-0.0480	0.0052	-0.0334	0.083
	25	0.2405	0.2623	0.2519	0.2615	0.1085	0.1418	0.1229	0.1400	0.109
	50	0.1179	0.2619	0.2524	0.2615	0.0525	0.1439	0.1236	0.1428	0.110
	100	0.0538	0.2540	0.2400	0.2539	0.0241	0.1372	0.1162	0.1369	0.095
12	8	0.2076	0.1302	0.0882	0.1075	0.0759	0.0629	0.0403	0.0433	0.034
	12	0.1972	0.1679	0.1196	0.1477	0.0856	0.0954	0.0632	0.0736	0.067
	25	0.0932	0.1469	0.1111	0.1473	0.0156	-0.0102	0.0076	0.0360	0.084
	50	0.0558	0.1838	0.1400	0.1804	0.0257	0.0994	0.0696	0.0953	0.092
	100	0.0249	0.1589	0.1253	0.1586	-0.0044	0.0038	0.0191	0.0079	0.103
24	8	0.1034	0.0903	0.0566	0.0618	0.0300	0.0407	0.0243	0.0193	0.028
	12	0.1068	0.1219	0.0788	0.0955	0.0417	0.0656	0.0402	0.0409	0.062
	25	0.0691	0.1428	0.0965	0.1302	0.0319	0.0794	0.0503	0.0666	0.091
	50	0.0343	0.1301	0.0896	0.1256	-0.0074	-0.0233	-0.0041	0.0085	0.113
	100	0.2159	0.4988	0.4752	0.5040	-0.2153	-0.4833	-0.4673	-0.4889	0.096
48	8	0.0332	0.0514	0.0295	0.0224	0.0061	0.0207	0.0115	0.0044	0.008
	12	0.0613	0.0788	0.0482	0.0618	0.0214	-0.0060	-0.0002	0.0230	0.053
	25	0.0430	0.1111	0.0682	0.0951	0.0209	0.0639	0.0377	0.0490	0.089
	50	0.0258	0.1013	0.0633	0.0943	-0.0135	-0.0282	-0.0111	-0.0102	0.110
	100	0.0578	0.2389	0.1550	0.2322	-0.0570	-0.2213	-0.1444	-0.2148	0.097

Cochran's Q statistic obtained from the simulation-based meta-analysis scenarios tended to be conservative. When the number of studies was held constant, the Type I error rate of Cochran's Q statistic was maintained at a robust level when n≥16 (Table **2**).

When the RMSE and BIAS values of the heterogeneity measures were examined according to the n when the k was kept constant for P($P^-|E^+$)=0.5 and P($P^+|E^+$)=0.7, the I^2 and R_b heterogeneity measures produced estimates that were very close to each other and above the population parameter. When k=3 was taken, as n increases, τ^2 approaches the population parameter. When k≥6, all heterogeneity measures produced estimates that closely approximated the population parameter, but τ^2 showed the best performance. When n<25 and k≤12, the H showed the best performance. At k=3, 12, the heterogeneity criterion produced an underestimate of the population parameter. When k=48, I^2 and H produced higher estimates. When the RMSE and BIAS values of the heterogeneity measures were examined according to the number of studies when the sample size was held constant, all measures closely aligned the population parameter as the number of studies increased. When n=25 was taken, as k increases, τ^2 achieved the best performance to estimate the population parameter, followed by the H. When n=50 was taken, the heterogeneity criteria started to approach each other and underestimated the population parameter. When n=100, they produced estimates very close to each other and the population

Table 3: Heterogeneity Measures' Simulation Results of P($P^-|E^+$) = 0.5 vs $P(P^+|E^+)$ = 0.7

k	n_P=n_K	P($P^-\|E^+$)=0.5 vs $P(P^+\|E^+)$=0.7 (OR=2.34)								Cochran's Q Type-I error
		RMSE				BIAS				
		τ^2	I^2	H	R_b	τ^2	I^2	H	R_b	
3	8	0.7949	0.2223	0.2045	0.2213	0.3016	0.1091	0.0901	0.1057	0.043
	12	0.6912	0.2522	0.2506	0.2508	0.2618	0.1304	0.1146	0.1269	0.065
	25	0.2845	0.3242	0.2801	0.3217	0.0218	-0.2271	-0.1312	-0.2235	0.091
	50	0.1528	0.4602	0.4322	0.4601	-0.0359	-0.3921	-0.3429	-0.3917	0.108
	100	0.1733	0.5962	0.7282	0.5954	-0.1605	-0.5490	-0.6837	-0.5481	0.090
6	8	0.3746	0.1610	0.1213	0.1513	0.1426	0.0755	0.0529	0.0660	0.036
	12	0.3264	0.1912	0.1538	0.1825	0.1315	0.0963	0.0711	0.0862	0.063
	25	0.1721	0.2194	0.1821	0.2135	0.0779	0.1208	0.0915	0.1133	0.082
	50	0.0885	0.2286	0.1942	0.2263	0.0416	0.1269	0.0981	0.1232	0.095
	100	0.0416	0.2258	0.1943	0.2086	0.0135	0.1208	0.0947	0.0845	0.099
12	8	0.1682	0.1050	0.0687	0.0872	0.0558	0.0441	0.0276	0.0315	0.016
	12	0.1843	0.1499	0.1049	0.1334	0.0764	0.0785	0.0514	0.0620	0.052
	25	0.1099	0.1774	0.1287	0.1682	0.0520	0.1002	0.0677	0.0913	0.082
	50	0.0598	0.1862	0.1404	0.1821	0.0286	0.1041	0.0723	0.0993	0.096
	100	0.0291	0.1621	0.1253	0.1606	-0.0115	-0.0285	-0.0003	-0.0230	0.108
24	8	0.1282	0.1810	0.1070	0.1180	-0.0965	-0.1693	-0.0996	-0.1076	0.010
	12	0.0939	0.1051	0.0661	0.0823	0.0353	0.0519	0.0313	0.0331	0.036
	25	0.1133	0.3682	0.2827	0.3657	-0.0935	-0.3487	-0.2708	-0.3489	0.093
	50	0.0385	0.1469	0.0991	0.1402	0.0189	0.0822	0.0523	0.0752	0.095
	100	0.0202	0.1505	0.1028	0.1470	0.0101	0.0839	0.0539	0.0807	0.103
48	8	0.0319	0.0364	0.0209	0.0201	0.0048	0.0102	0.0057	0.0032	0.004
	12	0.0528	0.1505	0.0871	0.0615	-0.0184	-0.1360	-0.0782	-0.0387	0.038
	25	0.0459	0.1955	0.1204	0.1779	-0.0271	-0.1757	-0.1080	-0.1614	0.079
	50	0.0250	0.1865	0.1145	0.1918	-0.0137	-0.1631	-0.0994	-0.1719	0.093
	100	0.0547	0.1045	0.0651	0.0888	-0.0537	0.0509	0.0312	-0.0040	0.097

parameter as the number of studies increased. When n=8, 12, and 100, they overestimated the population parameter. In general, when n>12, Cochran's Q statistic Type-I error rates were preserved at a sufficient level. When k>3 and n>25 were taken, the levels of protecting Cochran's Q statistic Type-I error rates were strengthened (Table **3**).

When the RMSE and BIAS values of the heterogeneity measures were examined according to sample size when the number of studies was held constant for $\mathrm{P}(P^-|E^+)$=0.5 and $\mathrm{P}(P^+|E^+)$=0.8, I^2 and R_b yielded highly similar estimates. H provided the most accurate estimates of the parameter when k≤12 and n<25, whereas τ^2 exhibited superior performance under conditions where n≥25. Furthermore, when k=3 and n≥50, the estimates produced by τ^2 appeared to stabilize, indicating a near-constant behavior. The criteria underestimated the parameter when k=3 and k=24, whereas overestimations were observed at the remaining values of k. When RMSE and BIAS values of the heterogeneity criteria were evaluated according to k when n was held constant, as k increased, all criteria produced estimates that converged toward each other and the population parameter. When n≥25, τ^2 began to yield estimates above the population value, demonstrating the best performance. Although criterion H was followed, it produced values closer to those of criteria I^2 and R_b. At n=50, estimates were produced above the parameter value, with criterion τ^2 showing the best performance, followed by criterion H. At n=100, criterion τ^2 provided the closest estimates, with

Table 4: Heterogeneity Measures' Simulation Results of $\mathrm{P}(P^-|E^+)$ = 0.5 vs $P(P^+|E^+)$ = 0.8

k	n_p=n_K	$\mathrm{P}(P^-\mid E^+)$=0.5 vs $P(P^+\mid E^+)$=0.8 (OR=4.02)								Cochran's Q Type-I error
		RMSE				BIAS				
		τ^2	I^2	H	R_b	τ^2	I^2	H	R_b	
3	8	0.6090	0.1829	0.1611	0.1824	0.2136	0.0774	0.0618	0.0759	0.031
	12	0.5850	0.2173	0.2049	0.2164	0.2153	0.1017	0.0857	0.1003	0.052
	25	0.3905	0.2631	0.2686	0.2607	0.1554	0.1406	0.1254	0.1359	0.086
	50	0.2000	0.2823	0.2962	0.2823	0.0847	0.1537	0.1425	0.1525	0.108
	100	0.2171	0.4404	0.4034	0.4416	-0.2009	-0.3736	-0.3105	-0.3749	0.096
6	8	0.2895	0.1183	0.0873	0.1137	0.0901	0.0446	0.0303	0.0409	0.012
	12	0.2760	0.1524	0.1171	0.1464	0.1014	0.0677	0.0477	0.0635	0.032
	25	0.2514	0.4499	0.3967	0.4504	0.4504	-0.1673	-0.3651	-0.4152	0.081
	50	0.1058	0.2241	0.1916	0.2196	0.0475	0.1227	0.0948	0.1176	0.093
	100	0.0502	0.2261	0.1922	0.2237	0.0234	0.1217	0.0948	0.1190	0.095
12	8	0.1265	0.0689	0.0441	0.0615	0.0329	0.0225	0.0136	0.0178	0.004
	12	0.1503	0.1076	0.0724	0.1021	0.0543	0.0452	0.0286	0.0405	0.022
	25	0.1065	0.1382	0.1015	0.1311	0.0367	0.0147	0.0198	0.0102	0.068
	50	0.0683	0.1822	0.1370	0.1753	0.0314	0.0983	0.0685	0.0909	0.092
	100	0.0331	0.1848	0.1366	0.1810	0.0165	0.1050	0.0720	0.1014	0.094
24	8	0.0443	0.0334	0.0195	0.0261	0.0084	0.0078	0.0044	0.0052	0.001
	12	0.0722	0.0641	0.0392	0.0570	0.0214	0.0231	0.0135	0.0181	0.01
	25	0.0633	0.1173	0.0748	0.1010	0.0276	0.0611	0.0373	0.0473	0.05
	50	0.1094	0.2163	0.1383	0.1854	-0.1038	-0.1836	-0.1151	-0.1502	0.086
	100	0.0729	0.2542	0.1698	0.2317	-0.0701	-0.2215	-0.1459	-0.1971	0.118
48	8	0.0111	0.0115	0.0064	0.0072	0.0008	0.0014	0.0008	0.0005	<0.001
	12	0.0293	0.0330	0.0188	0.0260	0.0068	0.0094	0.0051	0.0064	0.001
	25	0.0357	0.0813	0.0483	0.0642	0.0143	0.0404	0.0232	0.0270	0.042
	50	0.0246	0.0947	0.0588	0.0821	0.0094	0.0397	0.0246	0.0201	0.081
	100	0.0141	0.1076	0.0668	0.1009	0.0070	0.0593	0.0351	0.0529	0.092

a clear distinction between it and the other criteria, with criterion H following closely. In general n>25 was applied, and the Type I error of Cochran’s Q statistic was adequately controlled. In large sample sizes, Type I error protection for Cochran's Q statistic, as obtained through simulation scenarios in the meta-analysis, was consolidated (Table **4**).

When the number of studies held constant for P($P^-|E^+$)=0.5 and P($P^+|E^+$)=0.9, the RMSE and BIAS values of the heterogeneity measures were examined according to the sample sizes, I^2 and R_b produced an estimate highly similar. Across all studies, when n<25, the best performance was achieved by H, whereas τ^2 exhibited the best performance when n≥25. τ^2 approached the population parameter as the sample size increased. When k=48 and n>50, the best performance was succeeded by the H criterion. Heterogeneity measures produced an overestimate of the population parameter when k=3, 6, 12, 24, and an underestimate when k=48. When n=50 was taken, τ^2 achieved the best performance, as the number of studies increased, the H followed the τ^2, but H, I^2 and R_b exhibited similar estimates. When k≥24, H, whereas k<24, τ^2 performed best performance. Heterogeneity measures produced an underestimate of the parameter when n=100, and an overestimate of the other sample sizes. For P($P^-|E^+$)=0.5 and P($P^+|E^+$)=0.9, in general, when k≤6 and n=100 were employed, Cochran's Q statistic Type-I error was preserved at a sufficient level. In the other simulation scenarios, Cochran's Q statistic Type-I error could not be preserved (Table **5**).

Table 5: Heterogeneity Measures’ Simulation Results of $\mathrm{P}(P^-|E^+)$ = 0.5 vs $P(P^+|E^+)$ = 0.9

k	n_p=n_K	$\mathrm{P}(P^-\|E^+)$=0.5 vs $P(P\|E^+)$=0.9 (OR=9.01)								Cochran’s Q Type-I error
		RMSE				BIAS				
		τ^2	I^2	H	R_b	τ^2	I^2	H	R_b	
3	8	0.4219	0.1282	0.1065	0.1283	0.1172	0.0418	0.0316	0.0420	0.013
	12	0.4022	0.1522	0.1300	0.1731	-0.0179	-0.0476	-0.0117	-0.0937	0.023
	25	0.3682	0.2095	0.1985	0.2483	0.0160	-0.0726	-0.0123	-0.1530	0.051
	50	0.2492	0.4767	0.4490	0.4914	-0.0905	-0.4230	-0.3901	-0.4400	0.075
	100	0.1622	0.2755	0.2928	0.2751	0.0625	0.1466	0.1362	0.1449	0.097
6	8	0.1628	0.0647	0.0447	0.0635	0.0335	0.0149	0.0097	0.0148	0.003
	12	0.1621	0.0828	0.0576	0.0886	0.0471	0.0242	0.0158	0.0283	0.004
	25	0.1897	0.1505	0.1183	0.1504	0.0715	0.0643	0.0459	0.0655	0.031
	50	0.1177	0.1894	0.1485	0.1777	0.0500	0.0986	0.0712	0.0881	0.050
	100	0.0692	0.2147	0.1780	0.2063	0.0308	0.1139	0.0867	0.1057	0.084
12	8	0.0563	0.0291	0.0176	0.0283	0.0086	0.0048	0.0028	0.0046	<0.001
	12	0.0681	0.0402	0.0240	0.0459	0.0173	0.0093	0.0054	0.0123	<0.001
	25	0.0169	0.0337	0.0207	0.0343	0.0029	0.0060	0.0035	0.0062	0.001
	50	0.0125	0.0461	0.0296	0.0467	0.0025	0.0101	0.0061	0.0104	0.002
	100	0.0071	0.0534	0.0337	0.0541	0.0016	0.0131	0.0079	0.0133	0.001
24	8	0.0156	0.0093	0.0055	0.0083	0.0009	0.0007	0.0004	0.0005	<0.001
	12	0.0228	0.0145	0.0084	0.0161	0.0034	0.0016	0.0009	0.0027	<0.001
	25	0.0024	0.0051	0.0027	0.0059	0.0002	0.0004	0.0002	0.0005	<0.001
	50	0.0064	0.0239	0.0150	0.0247	0.0008	0.0030	0.0018	0.0034	0.003
	100	0.0037	0.0263	0.0172	0.0266	0.0005	0.0038	0.0022	0.0039	0.002
48	8	0.0243	0.0000	0.0000	0.0097	-0.0243	0.0000	0.0000	-0.0097	<0.001
	12	0.0101	0.0024	0.0012	0.0080	0.0010	0.0001	0.0001	0.0008	<0.001
	25	0.0228	0.0236	0.0133	0.0307	0.0060	0.0051	0.0028	0.0084	0.001
	50	0.0264	0.0715	0.0427	0.0640	0.0101	0.0311	0.0179	0.0261	0.031
	100	0.2076	0.2422	0.1557	0.2190	-0.2070	-0.2287	-0.1478	-0.2063	0.062

4.2. Simulation Results of Heterogeneous Studies

The results of the simulation scenarios where the studies included in the meta-analysis were heterogeneous were shown in Tables **6-8**. Simulations were performed on various values of sample sizes and the number of studies according to the probability of the patient group being exposed to the event. The OR values, power values of Cochran's Q statistics, the RMSE, and BIAS values of the heterogeneity measures were presented in Tables **6-8**.

High heterogeneity was obtained from the P($P^{-}|E^{+}$)=0.5 and P($P^{+}|E^{+}$)=0.6, 0.7, 0.8, 0.9 probabilities. The RMSE and BIAS values of the heterogeneity measures were examined according to the sample sizes when the number of studies was kept constant. R_b achieved best performance when k=3 for all sample sizes. When k≤24, I^2 and R_b produced similar estimates to the parameter. When k=48, τ^2 performed best performance, as n increased. The H criterion revealed the worst performance as k was kept constant, as n increased. Heterogeneity measures predicted an underestimate of the population parameter at k=3 and an overestimate at other number of studies. When the RMSE and BIAS values of the heterogeneity measures were examined according to the number of studies when the sample size was kept constant, when n<50, H, I^2 and R_b estimations were

Table 6: Heterogeneity Measures' Simulation Results of Studies with High Heterogeneity

k	nP=nK	studies with high heterogeneity*								Cochran's Q power
		RMSE				BIAS				
		τ^2	I^2	H	R_b	τ^2	I^2	H	R_b	
3	8	1.0323	0.2654	0.2300	0.2606	-0.5067	-0.1469	-0.0635	-0.1357	0.084
	12	0.9386	0.3194	0.3436	0.3183	0.4387	0.1915	0.1822	0.1889	0.133
	25	0.6574	0.4145	0.4925	0.4155	0.3643	0.2887	0.3069	0.2867	0.265
	50	0.4217	0.5043	0.6533	0.5062	0.2671	0.3972	0.4619	0.3973	0.412
	100	0.6350	0.3750	0.8340	0.3775	-0.5856	-0.2434	-0.5749	-0.2452	0.665
6	8	0.5274	0.2027	0.1610	0.1949	0.2345	0.1084	0.0796	0.0994	0.069
	12	0.5579	0.2859	0.2442	0.2757	0.3034	0.1912	0.1501	0.1762	0.156
	25	0.4727	0.2748	0.2748	0.3013	-0.3232	-0.1348	-0.0792	-0.1653	0.373
	50	0.3970	0.1946	0.3426	0.2195	-0.2648	-0.0201	0.0694	-0.0246	0.795
	100	0.2388	0.1611	0.6182	0.1734	0.1097	0.1293	0.5074	0.1402	0.987
12	8	0.2863	0.1353	0.0951	0.1258	0.1132	0.0655	0.0427	0.0569	0.043
	12	0.3087	0.1729	0.1313	0.1786	0.1055	-0.0399	-0.0068	0.0684	0.140
	25	0.4280	0.4347	0.3866	0.4144	0.3316	0.3920	0.3284	0.3621	0.623
	50	0.2046	0.1690	0.3494	0.1999	0.0098	0.1193	0.2598	0.1490	0.966
	100	0.2016	0.0941	0.4358	0.0910	-0.1458	0.0805	0.3646	0.0762	1.000
24	8	0.1367	0.0878	0.0549	0.0746	0.0479	0.0380	0.0227	0.0279	0.031
	12	0.2436	0.1926	0.1308	0.1741	0.1450	0.1344	0.0866	0.1135	0.179
	25	0.2953	0.3534	0.3133	0.3174	0.2302	0.3260	0.2774	0.2775	0.817
	50	0.1423	0.1414	0.2910	0.1695	-0.0258	0.1198	0.2414	0.1474	0.998
	100	0.1688	0.1003	0.4209	0.0960	-0.1353	0.0941	0.3842	0.0893	1.000
48	8	0.0652	0.0628	0.0364	0.0415	0.0212	0.0273	0.0154	0.0140	0.016
	12	0.1666	0.1654	0.1043	0.1385	0.1043	0.1235	0.0750	0.0923	0.220
	25	0.1358	0.1566	0.1627	0.1643	0.0408	0.1182	0.1230	0.1093	0.950
	50	0.2010	0.2828	0.4151	0.3274	0.1793	0.2775	0.3996	0.3216	1.000
	100	0.1123	0.1713	0.5421	0.1877	0.0910	0.1695	0.5294	0.1860	1.000

*P($P^{-}|E^{+}$)=0.5 vs P($P^{+}|E^{+}$)=0.6 (OR=1.50); P($P^{-}|E^{+}$)=0.5 vs P($P^{+}|E^{+}$)=0.7 (OR=2.34); P($P^{-}|E^{+}$)=0.5 vs P($P^{+}|E^{+}$)=0.8 (OR=4.00); P($P^{-}|E^{+}$)=0.5 vs P($P^{+}|E^{+}$)=0.9 (OR=9.04).

similar to each other and the population parameter in all number of studies. When n=50 was taken, as the number of studies increased, τ^2 showed the best performance by producing estimates above the population parameter, followed by I^2 and R_b. When n=100 was taken, estimates above the population value were produced. Although τ^2 exhibited the best performance, the estimates it produced were similar to those obtained from I^2 and R_b. Heterogeneity measures overestimated the population parameter. In the simulation scenario where the studies had high heterogeneity, when the sample size was taken as n>50, the power of Cochran's Q statistic of the study was at a sufficient level (Table **6**).

The RMSE and BIAS values of the heterogeneity measures were examined according to the sample sizes when the number of studies was kept constant while the studies had medium heterogeneity with probabilities of P($P^-|E^+$)=0.5 and P($P^+|E^+$)=0.6, 0.7, 0.8, 0.8, the I^2 and R_b had similar estimates. When k≤6, the best performance was achieved by the R_b as the sample size increased. When k>48, the best performance was performed by the τ^2 as the sample size increased. H demonstrated the poorest performance under these conditions. Heterogeneity measures underestimated the population parameter at k≤6. When the sample size was kept constant at n<50, as the number of studies increased, all measures demonstrated similar estimates to each other and the population parameter as the number of studies increased. When n≥50, as the number of studies increased, τ^2 achieved the best performance, and when n=50, H, $I^2$and R_b produced estimates close to each other. When n=100, overestimation revealed above the population parameter, τ^2 showed the best performance, followed by I^2 and R_b. Heterogeneity

Table 7: Heterogeneity Measures' Simulation Results of Studies with Moderate Heterogeneity

k	n_P=n_K	studies with moderate heterogeneity*								Cochran's Q power
		RMSE				BIAS				
		τ^2	I^2	H	R_b	τ^2	I^2	H	R_b	
3	8	1.0323	0.2654	0.2300	0.2606	-0.5067	-0.1469	-0.0635	-0.1357	0.084
	12	0.9386	0.3194	0.3436	0.3183	0.4387	0.1915	0.1822	0.1889	0.133
	25	0.6574	0.4145	0.4925	0.4155	0.3643	0.2887	0.3069	0.2867	0.265
	50	0.4217	0.5043	0.6533	0.5062	0.2671	0.3972	0.4619	0.3973	0.412
	100	0.6350	0.3750	0.8340	0.3775	-0.5856	-0.2434	-0.5749	-0.2452	0.665
6	8	0.5129	0.3149	0.2275	0.3236	-0.3033	-0.2707	-0.1887	-0.2836	0.050
	12	0.4820	0.2553	0.2171	0.2457	0.2390	0.1575	0.1220	0.1452	0.113
	25	0.9010	0.6007	1.1049	0.6420	-0.8561	-0.5495	-1.0765	-0.5933	0.283
	50	0.2865	0.4832	0.5159	0.4828	0.2065	0.4158	0.4054	0.4127	0.516
	100	0.1624	0.2161	0.4675	0.2205	0.0955	0.0828	0.2749	0.0859	0.817
12	8	0.2529	0.1384	0.0930	0.1228	0.1074	0.0714	0.0456	0.0573	0.037
	12	0.3294	0.2205	0.1648	0.2057	0.1760	0.1408	0.0986	0.1247	0.147
	25	0.2952	0.2357	0.1992	0.2463	-0.2149	-0.1145	-0.0629	-0.1309	0.358
	50	0.2179	0.2057	0.2383	0.2090	-0.1691	-0.0470	0.0024	-0.0501	0.688
	100	0.1355	0.3382	0.5635	0.3421	0.1018	0.3094	0.4960	0.3135	0.948
24	8	0.1122	0.0809	0.0500	0.0637	0.0362	0.0344	0.0203	0.0221	0.018
	12	0.1743	0.1548	0.1023	0.1333	0.0872	0.0965	0.0604	0.0738	0.103
	25	0.1986	0.3006	0.2252	0.2813	0.1453	0.2522	0.1787	0.2267	0.462
	50	0.1698	0.3706	0.3611	0.3942	0.1424	0.3420	0.3202	0.3657	0.876
	100	0.0616	0.1635	0.3342	0.1647	-0.0099	0.1395	0.2790	0.1407	0.996
48	8	0.0481	0.0502	0.0289	0.0311	0.0124	0.0183	0.0103	0.0083	0.011
	12	0.1164	0.1270	0.0783	0.1005	0.0588	0.0809	0.0480	0.0543	0.127
	25	0.1719	0.2930	0.2068	0.2704	0.1388	0.2630	0.1784	0.2336	0.675
	50	0.0656	0.1518	0.1841	0.1721	-0.0105	0.1160	0.1410	0.1395	0.973
	100	0.0459	0.1994	0.3599	0.2036	0.0199	0.1916	0.3372	0.1958	1.000

*$P(P^-|E^+)$=0.5 vs $P(P^+|E^+)$=0.6 (OR=1.50); $P(P^-|E^+)$=0.5 vs $P(P^+|E^+)$=0.7 (OR=2.34); $P(P^-|E^+)$=0.5 vs $P(P^+|E^+)$=0.8 (OR=4.00); $P(P^-|E^+)$=0.5 vs $P(P^+|E^+)$=0.8.

Table 8: Heterogeneity Measures' Simulation Results of Studies with Low Heterogeneity

k	n_P=n_K	studies with low heterogeneity*								Cochran's Q power
		RMSE				BIAS				
		τ^2	I^2	H	R_b	τ^2	I^2	H	R_b	
3	8	0.8732	0.2211	0.2273	0.2221	0.0823	0.0170	0.0542	0.0176	0.086
	12	1.4493	0.4575	0.4351	0.4662	-1.2717	-0.3926	-0.3512	-0.4021	0.097
	25	0.4152	0.3271	0.3632	0.3270	0.1961	0.1934	0.1905	0.1922	0.151
	50	0.2224	0.3146	0.3864	0.3534	0.1106	0.1511	0.1867	0.2202	0.180
	100	0.2744	0.4779	0.5857	0.4778	-0.2415	-0.3661	-0.3886	-0.3663	0.280
6	8	0.4341	0.1830	0.1406	0.1714	0.1787	0.0944	0.0672	0.0812	0.046
	12	0.3680	0.2222	0.1805	0.2107	0.1670	0.1266	0.0945	0.1122	0.087
	25	0.4096	0.3700	0.3077	0.3707	-0.3595	-0.3017	-0.2361	-0.3027	0.150
	50	0.1434	0.2761	0.2778	0.3210	0.0823	0.1485	0.1478	0.2203	0.209
	100	0.5797	0.5276	0.8393	0.5391	-0.5748	-0.4595	-0.7859	-0.4726	0.346
12	8	0.1883	0.1177	0.0786	0.0965	0.0661	0.0548	0.0345	0.0376	0.024
	12	0.2246	0.1707	0.1245	0.1573	0.1011	0.0892	0.0619	0.0803	0.079
	25	0.1552	0.2281	0.1768	0.2179	0.0819	0.1471	0.1048	0.1345	0.150
	50	0.0971	0.2005	0.1671	0.2061	-0.0622	-0.0398	0.0022	-0.0620	0.241
	100	0.0647	0.3487	0.2970	0.3480	0.0459	0.2762	0.2193	0.2753	0.407
24	8	0.0934	0.0795	0.0485	0.0565	0.0287	0.0337	0.0198	0.0184	0.012
	12	0.1194	0.1250	0.0810	0.0995	0.0492	0.0697	0.0426	0.0458	0.051
	25	0.0957	0.1821	0.1249	0.1681	0.0534	0.1188	0.0769	0.1023	0.158
	50	0.0670	0.2362	0.1682	0.2307	0.0455	0.1778	0.1196	0.1709	0.284
	100	0.0476	0.1700	0.1467	0.1713	-0.0354	-0.0196	0.0091	-0.0270	0.507
48	8	0.0315	0.0441	0.0251	0.0212	0.0058	0.0154	0.0085	0.0041	0.007
	12	0.0634	0.0861	0.0505	0.0627	0.0247	-0.0352	-0.0166	0.0256	0.058
	25	0.0598	0.1168	0.0782	0.1177	0.0178	0.0164	0.0170	0.0463	0.209
	50	0.0378	0.1550	0.1126	0.1478	0.0035	0.0846	0.0631	0.0704	0.383
	100	0.0385	0.2612	0.2010	0.2730	0.0304	0.2254	0.1677	0.2386	0.734

*$P(P^-|E^+)=0.5$ vs $P(P^+|E^+)=0.6$ (OR=1.50); $P(P^-|E^+)=0.5$ vs $P(P^+|E^+)=0.7$ (OR=2.34); $P(P^-|E^+)=0.5$ vs $P(P^+|E^+)=0.7$; $P(P^-|E^+)=0.5$ vs $P(P^+|E^+)=0.7$.

criteria overestimated the parameter when studies had medium heterogeneity. The power of Cochran's Q statistic reached a sufficient level when n>25 and k>6 in the simulation scenarios (Table **7**).

Low heterogeneity achieved with probabilities of P($P^-|E^+$)=0.5 and P($P^+|E^+$)=0.6, 0.7, 0.7, 0.7. The RMSE and BIAS values of the heterogeneity criteria were evaluated when the number of studies was kept constant, as the sample size increased. When k=3, the H, R_b and I^2 produced close values to each other. At k=6, 12, 24, and 48 the H criterion outperformed I^2 and R_b, the τ^2 obtained the best performance. At k=3 and 6, all criteria underestimated the parameter, while at k=12 and 48, the H overestimated the population parameter and the others underestimated. The RMSE and BIAS values of the heterogeneity criteria were examined as the number of studies increased when the sample size was kept constant. All criteria produced values close to each other and to the population parameter when n≤25. When n≥50 was taken, τ^2 showed the best performance. The criteria overestimated the parameter. The power of Cochran's Q statistics of the studies did not achieve a satisfactory level (Table **8**).

5. DISCUSSION

The continuity of scientific knowledge is founded on the findings of prior research within a given discipline; the building blocks of this knowledge are constituted by individual studies [21]. Meta-analysis is relevant when studies on the same or similar topics or problems yield contradictory findings, making the results difficult to interpret [22]. Heterogeneity refers to the variability in effect sizes across different studies included in the meta-analysis [5]. Patsopoulos, Evangelou [9] developed algorithms that they applied to meta-analysis databases to assess the change in heterogeneity across studies. These algorithms aimed to remove one or more studies to obtain the maximum or minimum I^2 according to a predetermined threshold value [23]. Higgins [5] criticized the authors in his study, stating that if a clear outlier is excluded, another study may appear to be an outlier when the remaining studies are evaluated and will be excluded in turn. Therefore, a predetermined stopping rule, which the authors call a “desired heterogeneity threshold,” may seem like a useful way to go. However, as Patsopoulos, Evangelou [9] have pointed out, Higgins [5] has expressed concerns about whether excluding studies is useful for assessing the sensitivity of heterogeneity measures and, in particular, whether it makes sense to set a desired threshold value for the I^2 statistic as these authors do. Higgins [5] argues that I^2 is not appropriate for measuring the magnitude of between-study heterogeneity or for using it as a point estimate of between-study heterogeneity, but only represents an approximation of how much of the total variability in the point estimates can be attributed to heterogeneity. Since the total variation depends significantly on within-study precision and mainly on the sample sizes of the studies, I^2 is affected by sample size. Higgins [5] mentions that Patsopoulos, Evangelou [9] neglect to specify the magnitude of heterogeneity, τ^2, which is the point estimate of the between-study variance.

A diversity of opinions in the literature concerning the assessment of heterogeneity in meta-analysis. However, there is a lack of sufficient simulation studies that compare the four commonly used heterogeneity measures. Our study aims to compare the performances of the I^2, R_b, τ^2 and H heterogeneity criteria in simulation scenarios, which are commonly used in meta-analysis of binary data. They evaluated by the RMSE and BIAS values, in terms of homogeneous and heterogeneous studies with low, medium, and high heterogeneity with various studies, sample sizes, and effect sizes. Heterogeneity levels were determined by the literature. Additionally, the studies were examined in terms of Cochran’s Q Type-I rate and the power of Cochran’s Q statistic.

In cases where the effect sizes of the studies were homogeneous and the event did not pose a disease risk [OR=1.00; P($P^-|E^+$)=0.5 and P($P^+|E^+$)=0.5], the performances of the heterogeneity measures were compared according to the sample sizes when the number of studies was kept constant. While the number of studies was low, I^2, R_b and H were similar and overestimated the population parameter. Under small sample conditions, I^2, R_b, τ^2 and H did not maintain the Type I error rate of Cochran’s Q statistic at an acceptable level. The τ^2 produced estimates close to the population parameter as the sample size increased in each number of studies. As the sample size increased, the criteria overestimated the parameter, and their performance in protecting against Cochran's Q statistic Type-I error reached a sufficient level. When the sample size was kept constant, the four heterogeneity measures produced estimations close to each other and the parameter as the number of studies increased in small samples. As the number of studies increased in large samples, the criteria approached the parameter. The best performance was achieved by τ^2, followed by the Crippa, Khudyakov [16] suggested the R_b, stating that I^2 was derived under the assumption that within-study variances were homogeneous and were not sufficient to determine heterogeneity. However, in the simulation scenarios we designed, the R_b produced similar estimates to I^2. In general, as the number of studies and sample size increased, it was observed that the heterogeneity criteria followed a liberal attitude towards preserving Cochran's Q statistic Type-I error. When the risk factor had a low effect on the disease in homogeneous scenarios, the number of studies was held constant, increasing the sample size in a small number of studies led to estimates of τ^2 that approached the true population parameter. I^2, H and R_b yielded similar values, and their estimations exhibited minimal variation regardless of whether the sample size was small or large. In general, heterogeneity measures produced an overestimation of the population parameter. Since we reached a similar conclusion with Huedo-Medina [6], who stated that τ^2 is the parameter representing the true heterogeneity between the true effects of the studies, we can say that τ^2 is the criterion that shows the best performance. τ^2 was followed by the H criterion. Taking the sample size too high in the high number of studies of homogeneous studies with low effect sizes caused the criteria to deviate from the parameter. While the number of studies was fixed, the performance of Cochran's Q statistic in protecting against Type I errors increased as the sample size increased. In cases where the effect sizes of the studies were homogeneous and the factor posed a low risk to the disease when the sample size was held constant, the criteria estimated the parameter better as the number of studies increased. When the sample size was taken as high, τ^2 began to differentiate from other criteria and produce estimates closer to the parameter.

In cases where homogeneous studies have high effect sizes when the number of studies is kept constant, as the sample size increases, τ^2 approaches the population parameter. Huedo-Medina [6] emphasized that, I^2 should be interpreted very carefully in the small number of studies. In our study, it was observed that the performance of I^2 decreases even when the sample size is increased in the small number of studies. When I^2, H and R_b were taken as k=3, they move away from the parameter when the sample size increases a lot, while τ^2 continues to approach. A similar situation occurred when the number of studies is taken as high, I^2, H and R_b move away from the parameter as the sample size increases. In all cases, τ^2 was the heterogeneity criterion that produced the closest estimates. Except for a small number of studies, the H criterion produced the best estimates after τ^2 as the sample size increased. The estimates of the heterogeneity criterion tended to be above the parameter. In homogeneous studies where the risk factor has a high effect when the sample size is taken as high, Cochran's Q statistic Type-I error can be preserved. When the sample size is taken as constant, as the number of studies increases, the estimates of the criteria approach the parameter. When the sample size is taken as high, τ^2 showed the best performance in every number of studies simulated from the smallest number of studies to the largest number of studies, followed by the H criterion. Heterogeneity criteria generally produce estimates above the population parameter, and in very high sample sizes, as the number of studies increases, they tend to produce values below the population.

In cases where heterogeneous studies have high effect sizes when the number of studies is kept constant, I^2 and R_b generally perform better than other criteria as the sample size increases, while when the number of studies is taken as high, the sample size increases and τ^2 produces estimates closer to the population parameter. In the small number of studies, the criteria tended to produce estimates below the parameter, while in other number of studies, they produced estimates above the parameter. When the sample size is kept constant, in a small number of studies, H, I^2 and R_b produce estimates close to the population parameter according to τ^2, while as the number of studies increases, τ^2 approaches the other criteria and the parameter. When the sample size is taken as n=100, while the number of studies increases, the estimates of other criteria approach the population parameter, while the H criterion moves away. They generally produce estimates above the parameter. In heterogeneous studies with high effect sizes, even if k≥4 is taken as stated by Patsopoulos, Evangelou [9], we conclude that I^2 is insufficient to determine heterogeneity, in line with Higgins [5] in our thesis study. The power of Cochran's Q statistic of simulation scenarios increased as the number of studies and sample size increased.

In scenarios where heterogeneous studies have medium effect sizes when the number of studies is kept constant, the H criterion generally moves away from the parameter as the sample size increases, and in the small number of studies, the high sample size also causes the τ^2 criterion to move away from the parameter. In general, in a small number of studies, as the sample size increases, the I^2 and R_b criteria produce the closest estimates of the population. The criteria tend to produce estimates above the parameter as the sample size increases in the high number of studies. When the sample size is kept constant, they tend to produce estimates close to the population value as the number of studies increases. When the sample size is high, the τ^2 shows the best performance and the H criterion shows the worst performance as the number of studies increases. It has been observed that I^2 and R_b produce values very close to each other. Crippa, Khudyakov [16] suggested the R_b in their study because it is easier to interpret and estimate the population better. In our study, it has been observed that the I^2 and R_b act together. For this reason, we can say that the R_b can be used instead of I^2. In general, heterogeneity criteria produced estimates above the population value when the sample size was kept constant. The power of Cochran's Q statistic in simulation scenarios where the heterogeneity of the studies was at a moderate level increased as the number of studies and the sample size increased.

Under circumstances where heterogeneous studies have low effect sizes when the number of studies is kept constant, as the sample size increases, τ^2 produces the closest estimate to the population parameter. As Higgins [5] stated in his study, using I^2 as a descriptive statistic instead of a heterogeneity criterion is also consistent with our results. In a small number of studies, H, I^2 and R_b produced values very close to each other. In a large number of studies, as the sample size increases, the H criterion showed the best performance after τ^2. In general, when the number of studies is small, the sample size increases and the criteria tend to produce values lower than the population parameter. When the sample size is kept constant, as the number of studies increases, the criteria produced estimates close to each other and the population value. When the sample size increases, as the number of studies increases, all criteria approach the population parameter, and τ^2 showed the best performance. As the number of studies increases, the criteria tend to produce estimates above the population value. Although the power of Cochran's Q statistic for heterogeneous studies with low effect size increases as the sample size and number of studies increase, it reached the highest level at k=48 n=100; but it was not sufficient. When the studies had low heterogeneity, the worst performance was shown by R_b and I^2 measures. τ^2 performed better in large sample sizes.

Although τ^2 did not perform well in small samples and many studies when the studies were homogeneous, it estimated the population parameter better than the H, I^2 and R_b when the sample size increased. The heterogeneity measure most affected by the sample size was τ^2. When the number of studies was taken as high, the performance of the heterogeneity measures decreased as the sample size increased.

In our study where we examined the performance of heterogeneity measures, we also evaluated the performance of Cochran's Q statistic, which is widely used in examining heterogeneity. In cases where heterogeneous studies have high and medium effect sizes, the power of Cochran's Q statistic of the simulation scenarios increased as the sample size and number of studies increased. In heterogeneous studies with low effect sizes, the power of Cochran's Q statistic of the simulation could not reach a sufficient level even at the highest number of studies and sample sizes we included in the simulation scenarios.

6. CONCLUSION

As a result, the I^2 heterogeneity criterion, which is widely used in the literature, estimated the parameter well in the small number of studies and small sample sizes. The H criterion moved away from the parameter in cases where the studies had medium and high heterogeneity, while I^2 and R_b approached. It was observed that the I^2 and R_b acted together in all scenarios and produced very close estimates. When examining heterogeneity in meta-analysis, we recommend that, I^2 and R_b should be examined first in small sample sizes and a small number of studies and the H heterogeneity criterion should be examined first after τ^2 in high sample sizes and a high number of studies.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

No funds, grants, or other support was received.

FINANCIAL INTERESTS

The authors declare they have no financial interests.

DATA AVAILABILITY STATEMENT

The data/code generated and/or analyzed during the current study is available from the corresponding author on reasonable request.

REFERENCES

[1] O'rourke K. An historical perspective on meta-analysis: Dealing quantitatively with varying study results. Journal of the Royal Society of Medicine 2007; 100(12): 579-82. https://doi.org/10.1177/0141076807100012020

[2] Haidich AB. Meta-analysis in medical research. Hippokratia 2010; 14(Suppl 1): 29-37.

[3] Mikolajewicz N, Komarova SV. Meta-Analytic Methodology for Basic Research: A Practical Guide. Frontiers in Physiology 2019; Volume 10 - 2019. https://doi.org/10.3389/fphys.2019.00203

[4] Smith T. Proposing Alternative Methods for Testing Heterogeneity of Studies in a Meta-analysis: Illinois State University; 2021.

[5] Higgins JPT. Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. International Journal of Epidemiology 2008; 37(5): 1158-60. https://doi.org/10.1093/ije/dyn204

[6] Huedo-Medina TB, Sánchez-Meca J, Marín-Martínez F, Botella J. Assessing heterogeneity in meta-analysis: Q statistic or I² index? Psychological Methods 2006; 11(2): 193-206. https://doi.org/10.1037/1082-989X.11.2.193

[7] Jackson D. Assessing the Implications of Publication Bias for Two Popular Estimates of Between-Study Variance in Meta-Analysis. Biometrics 2007; 63(1): 187-93. https://doi.org/10.1111/j.1541-0420.2006.00663.x

[8] Thorlund K, Imberger G, Johnston BC, Walsh M, Awad T, Thabane L, *et al*. Evolution of Heterogeneity (I2) Estimates and Their 95% Confidence Intervals in Large Meta-Analyses. PLOS ONE 2012; 7(7): e39471. https://doi.org/10.1371/journal.pone.0039471

[9] Patsopoulos NA, Evangelou E, Ioannidis JP. Sensitivity of between-study heterogeneity in meta-analysis: proposed metrics and empirical evaluation. International Journal of Epidemiology 2008; 37(5): 1148-57. https://doi.org/10.1093/ije/dyn065

[10] von Hippel PT. The heterogeneity statistic I2 can be biased in small meta-analyses. BMC Medical Research Methodology 2015; 15(1): 35. https://doi.org/10.1186/s12874-015-0024-z

[11] Baujat B, Mahé C, Pignon J-P, Hill C. A graphical method for exploring heterogeneity in meta-analyses: application to a meta-analysis of 65 trials. Statistics in Medicine 2002; 21(18): 2641-52. https://doi.org/10.1002/sim.1221

[12] Mittlböck M, Heinzl H. A simulation study comparing properties of heterogeneity measures in meta-analyses. Statistics in Medicine 2006; 25(24): 4321-33. https://doi.org/10.1002/sim.2692

[13] Borenstein M LVH, Julian P. T. Higgins, Hannah R. Rothstein. Introduction to Meta-Analysis: John Wiley & Sons, Ltd; 2009. 450 p. https://doi.org/10.1002/9780470743386

[14] Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Statistics in Medicine 2002; 21(11): 1539-58. https://doi.org/10.1002/sim.1186

[15] Rücker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I2 in assessing heterogeneity may mislead. BMC Medical Research Methodology 2008; 8(1): 79. https://doi.org/10.1186/1471-2288-8-79

[16] Crippa A, Khudyakov P, Wang M, Orsini N, Spiegelman D. A new measure of between-studies heterogeneity in meta-analysis. Statistics in Medicine 2016; 35(21): 3661-75. https://doi.org/10.1002/sim.6980

[17] Crippa A, editor A new measure of between-studies heterogeneity in meta-analysis. XXVIIIth International Biometric Conference; 2016; Victoria. https://doi.org/10.1002/sim.6980

[18] Van Houwelingen HC, Zwinderman KH, Stijnen T. A bivariate approach to meta-analysis. Statistics in Medicine. 1993; 12(24): 2273-84. https://doi.org/10.1002/sim.4780122405

[19] Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003; 327(7414): 557-60. https://doi.org/10.1136/bmj.327.7414.557

[20] Peterson K. Six Modifications Of The Aligned Rank Transform Test For Interaction. The Journal of Modern Applied Statistical Methods 2002; 1(1). https://doi.org/10.22237/jmasm/1020255240

[21] Hsiung T-H, Olejnik S. Type I Error Rates and Statistical Power for the James Second-Order Test and the Univariate F Test in Two-Way Fixed-Effects ANOVA Models under Heteroscedasticity and/or Nonnormality. The Journal of Experimental Education. 1996; 65(1): 57-71. https://doi.org/10.1080/00220973.1996.9943463

[22] Paul J, Barari M. Meta-analysis and traditional systematic literature reviews—What, why, when, where, and how? Psychology & Marketing 2022; 39(6): 1099-115. https://doi.org/10.1002/mar.21657

[23] Stroup DF, Thacker SB. Meta analysis. In: Britannica TEoE, editor. Encyclopedia Britannica 2024.

Received on 10-05-2025 Accepted on 01-06-2025 Published on 04-07-2025

https://doi.org/10.6000/1929-6029.2025.14.30