Scientific Research Publishing

Open Journal of Statistics, 2018, 8, 915-930
http://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Profile Likelihood Tests for Common Risk Ratios in Meta-Analysis Studies

Chukiat Viwatwongkasem[1*], Khanokporn Donjdee[2], Tantanut Poodphraw[3]

[1]Department of Biostatistics, Faculty of Public Health, Mahidol University, Bangkok, Thailand
[2]National Institute for Child and Family Development, Mahidol University, Salaya, Nakhon Pathom, Thailand
[3]Mueang Ranong District Public Health Office, Ranong, Thailand
Email: *chukiat.viw@mahidol.ac.th, khanokporn.don@mahidol.ac.th

How to cite this paper: Viwatwongkasem, C., Donjdee, K. and Poodphraw, T. (2018) Profile Likelihood Tests for Common Risk Ratios in Meta-Analysis Studies. *Open Journal of Statistics*, **8**, 915-930.
https://doi.org/10.4236/ojs.2018.86061

Received: November 6, 2018
Accepted: December 11, 2018
Published: December 14, 2018

Abstract

It is well-known that the power of Cochran's Q test to assess the presence of heterogeneity among treatment effects in a clinical meta-analysis is low due to the small number of studies combined. Two modified tests (*PL*1, *PL*2) were proposed by replacing the profile maximum likelihood estimator (PMLE) into the variance formula of logarithm of risk ratio in the standard chi-square test statistic for testing the null common risk ratios across all k studies ($i = 1, \cdots, k$). The simply naive test (*SIM*) as another comparative candidate has considerably arisen. The performance of tests in terms of type I error rate under the null hypothesis and power of test under the random effects hypothesis was done via a simulation plan with various combinations of significance levels, numbers of studies, sample sizes in treatment and control arms, and true risk ratios as effect sizes of interest. The results indicated that for moderate to large study sizes ($k \ge 16$) in combination with moderate to large sample sizes ($n_i^T, n_i^C \ge 50$), three tests (*PL*1, *PL*2, and *Q*) could control type I error rates in almost all situations. Two proposed tests (*PL*1, *PL*2) performed best with the highest power when $k \ge 16$ and moderate sample sizes ($n_i^T, n_i^C = 50, 100$); this finding was very useful to make a recommendation to use them in practical situations. Meanwhile, the standard Q test performed best when $k \ge 16$ and large sample sizes ($n_i^T, n_i^C \ge 500$). Moreover, no tests were reasonable for small sample sizes ($n_i^T, n_i^C \le 10$), regardless of study size *k*. The simply naive test (*SIM*) is recommended to be adopted with high performance when $k = 4$ in combination with ($n_i^T, n_i^C \ge 500$).

Keywords

Profile Likelihood Test, Cochran Q Test, Meta-Analysis, Heterogeneity, Risk Ratios

1. Introduction

In a clinical trial with binary outcomes, the risk ratio (*RR*) as an intervention effect is defined by the ratio of probabilities (risks) of having an adverse event between a treatment group and a control group [1] [2]. Let x^T and x^C be the number of events out of n^T and n^C, the total number of persons (or the total of times that every person exposed) in the treatment arm and the control arm, respectively. Then the maximum likelihood estimate for *RR* is obtained as $\hat{R}R = \frac{\hat{p}^T}{\hat{p}^C} = \frac{x^T/n^T}{x^C/n^C}$ [3] [4].

A meta-analysis of study size *k* is a statistical approach that combines the results from *k* studies, conducted on the same topic and with the similar methods, into a single summary result. In clinical trials, meta-analysis is an essential tool to obtain a better understanding of how well the treatment effects work. Two popularly statistical models used are the fixed effect model and the random effect model. Under the assumption of the fixed effect model, we assume that all studies share a common effect size. It means that there is no heterogeneity between the studies; all studies contain only one true effect size over all *k* independent trials, and the observed effect is determined by the common true effect plus the sampling error (within-study error). On the contrast, under the random effects model, the true effect is not the same in all studies; we allow that there is a distribution of true effect sizes. It follows that the combined estimate is not an estimate of one value, but rather it is the average of distribution values. Hence, there are two levels of errors (within-study error and between-study error). Consequently, the observed effect is determined by the mean of all true effects plus the within-study error and the between-study error. In this sense, heterogeneity may refer to various true effect sizes from studies to studies, or the difference of studies gives the difference of the effect sizes so that one can incorporate this heterogeneity into a random effect model. Alternatively, heterogeneity in the effect sizes from different studies may be explained by a set of covariates, such as characteristics of studies, type of treatment status, some average or aggregate characteristics of patients, even publication bias; therefore, a meta-regression approach may be used to account for variation from such covariates among these heterogeneous effects.

Traditionally, before combining the effects of separate studies by using either the fixed effect model as homogeneity or the random effect model as heterogeneity, the conventional Cochran's *Q* test is adopted to test whether these treatment effects are homogeneous, or not. Unfortunately, it is widely known that the standard *Q* test may be inaccurate in testing the null homogeneity of effect sizes in the sense of low power of test. Kulinskaya and Dollinger [5] and Boissel *et al.* [6] stated that Cochran's *Q* test had low power in most situations, especially, when the number of studies (*k*) was small. The work of Kulinskaya, Dollinger, and Bjørkestøl [7], Lipsitz *et al.* [1] and Lui's [2] were also confirmed the low power problem of Cochran's *Q* test. The low power of *Q* test implies the low ability to detect the effect when the effect actually exists (*i.e.* the low chance of rejecting the null homogeneous effects when the different effects exist). The simple correction for Q test to solve the problem of low power is taking a larger level of significance; Fleiss [8] recommended using a cut-off significance level of 0.1, rather than the usual 0.05. This has also been a common customary practice for the Cochran's Q homogeneity test in meta-analysis. Considerably, the way to increase the power is equivalent to the reduction of the chance of type II error. But this reduction of the chance of type II error also increases the risk or the chance of type I error. Obviously, when we make a low power problem better by using a cut-off of 10% for significance criterion, the new problem of allowance for the increase of the chance of type I error may occur. The increasing risk of type I error potentially leads to the problem of not maintaining the type I error at the conventional level of significance. Additionally, Shandish and Haddock [9] stated that when the sample sizes in each study were very large, the null hypothesis of the equal population effects might be rejected even if the individual effect estimates did not really differ much.

Profile likelihood estimation, stated by Ferrari *et al.* [10] and Böhning *et al.* [11], deals with elimination of the nuisance parameters. Generally, let the log-likelihood $l(\boldsymbol{p},\boldsymbol{q})$ depend on a vector $\boldsymbol{p}$ of parameters of interest and a vector $\boldsymbol{q}$ of nuisance parameters. If $\boldsymbol{q}_p$ as a function of $\boldsymbol{p}$ is the solution such that $l(\boldsymbol{q}_p \mid \boldsymbol{p}) \geq l(\boldsymbol{q} \mid \boldsymbol{p})$ for all $\boldsymbol{q}$, then $l(\boldsymbol{q}_p \mid \boldsymbol{p}) = l^*(\boldsymbol{p})$ is called the profile log-likelihood. Profile log-likelihood $l(\boldsymbol{q}_p \mid \boldsymbol{p}) = l^*(\boldsymbol{p})$ is not an ordinary log-likelihood, but log-likelihood maximized over nuisance parameters given the values of the parameter of interest. We can observe that the profile log-likelihood $l^*(\boldsymbol{p})$ now depends only on the parameter of interest.

With the Q limitations of low power and not maintaining type I error at the conventional level of significance, many scientists have attempted to propose some new tests and/or some modified tests to be alternative candidates. To meet the gaps of limitations, our proposed tests modified from the standard χ^2 test of homogeneity as an alternative choice are based on the substitution of profile maximum likelihood estimates derived by Böhning *et al.* [11] into the variance formula of logarithm of risk ratio as the effect measure of interest over all k studies. Another comparative test was the simply naive test based on the variance estimate of the conventional Poisson likelihood. Some numerical examples are illustrated later. Then, the next contribution focuses on a comparison of the performance among these homogeneity tests via a simulation plan. The result is related to the mentioned tests through the type I error probability and the power criteria lying on the later section. The conclusion and discussion are presented finally.

2. Motivational Applications

Two examples of meta-analysis are presented to illustrate the implementation of the related Q test and the other usefulness demonstrates how to set the parameters in a simulation plan. Farquhar *et al.* [12] conducted a meta-analysis on $k = 7$ studies to assess 5 years follow up of high dose chemotherapy and autograft comparable with the conventional chemotherapy for poor prognosis breast cancer. The outcome of treatment is event free survival. The value of Cochran's *Q*-test was 4.72. Since Q is distributed as a standard chi-square statistic with *k*-1 degrees of freedom (*df*), leading to the *p*-value of 0.58 for accepting the null homogeneity of risk ratios across trials. Additionally, the I^2 statistic denoted as $I^2 = 100\% \times (Q - df)/Q$ describing the percentage of variation across studies due to heterogeneity is very low of 0%; consequently, a fixed effects model might be appropriate. The conclusion of acceptation of the null hypothesis was that there was no presence of heterogeneity (**Figure 1**). In addition, there was no difference between treatment and control groups on binary events; the pooled estimate of *RR* being of 1.01 under a fixed effects model lies on the 95% confidence interval (CI.) of [0.97, 1.06], covering the null value 1. Forest graph of meta-analysis is created by R package provided by Schwarzer *et al.* [13], http://meta-analysis-with-r.org/.

Mottillo *et al.* [14] considered the data from meta-analysis of 16 trails about the metabolic syndrome and cardiovascular risk. The value of Cochran's *Q*-test is 6.12. The Chi-square approximation with 15 degrees of freedom provides 0.0003 of the *p*-value for testing the null homogeneity. The heterogeneity value of I^2 index was 64%. The result shows evidence to conclude heterogeneity of across studies (**Figure 2**). Furthermore, there exist the treatment effects on the binary outcomes since *RR* of 2.34 under a random effect model lies away from 1; the 95% CI has the range of [2.02, 2.72], not covering 1.

3. Deriving Profile Likelihood Tests for Common Risk Ratio

The purposes of study are 1) to derive the profile likelihood tests for testing a null common risk ratio *RR* across *k* studies in which is equivalent to homogeneity of treatment effects overall *k* studies ($i = 1, \cdots, k$) by replacing the profile likelihood estimator into the formulas of the estimate of variance of logarithmic relative risk, $\hat{\text{var}}\left(\log\left(\hat{RR}_i\right)\right)$, of the standard chi-square test; 2) to compare the performance of test statistics based on the profile likelihood method regarding

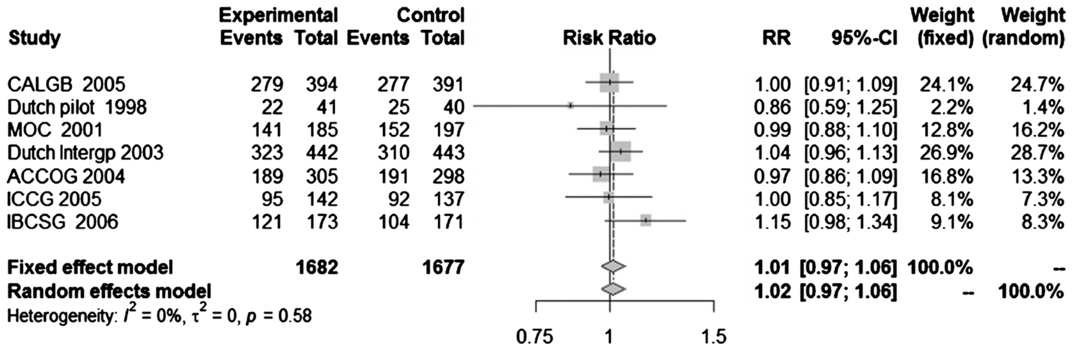
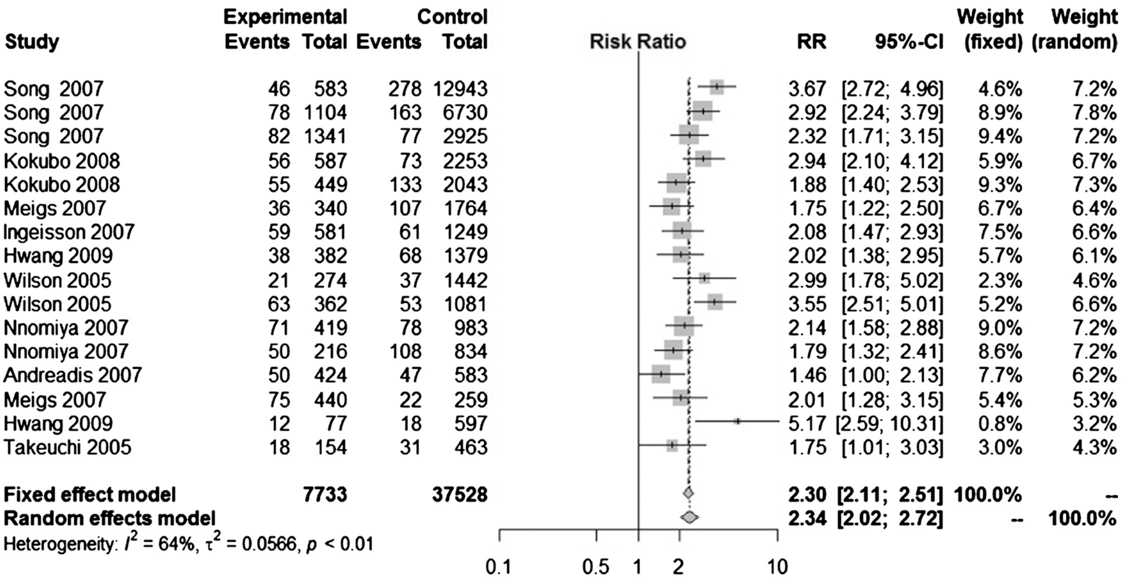


Figure 1. Forest plot of meta-analysis comparing high dose chemotherapy and autograft with the conventional chemotherapy.



Study	Experimental Events	Experimental Total	Control Events	Control Total	RR	95%-CI	Weight (fixed)	Weight (random)
Song 2007	46	583	278	12943	3.67	[2.72; 4.96]	4.6%	7.2%
Song 2007	78	1104	163	6730	2.92	[2.24; 3.79]	8.9%	7.8%
Song 2007	82	1341	77	2925	2.32	[1.71; 3.15]	9.4%	7.2%
Kokubo 2008	56	587	73	2253	2.94	[2.10; 4.12]	5.9%	6.7%
Kokubo 2008	55	449	133	2043	1.88	[1.40; 2.53]	9.3%	7.3%
Meigs 2007	36	340	107	1764	1.75	[1.22; 2.50]	6.7%	6.4%
Ingeisson 2007	59	581	61	1249	2.08	[1.47; 2.93]	7.5%	6.6%
Hwang 2009	38	382	68	1379	2.02	[1.38; 2.95]	5.7%	6.1%
Wilson 2005	21	274	37	1442	2.99	[1.78; 5.02]	2.3%	4.6%
Wilson 2005	63	362	53	1081	3.55	[2.51; 5.01]	5.2%	6.6%
Nnomiya 2007	71	419	78	983	2.14	[1.58; 2.88]	9.0%	7.2%
Nnomiya 2007	50	216	108	834	1.79	[1.32; 2.41]	8.6%	7.2%
Andreadis 2007	50	424	47	583	1.46	[1.00; 2.13]	7.7%	6.2%
Meigs 2007	75	440	22	259	2.01	[1.28; 3.15]	5.4%	5.3%
Hwang 2009	12	77	18	597	5.17	[2.59; 10.31]	0.8%	3.2%
Takeuchi 2005	18	154	31	463	1.75	[1.01; 3.03]	3.0%	4.3%
Fixed effect model		**7733**		**37528**	**2.30**	**[2.11; 2.51]**	**100.0%**	**--**
Random effects model					**2.34**	**[2.02; 2.72]**	**--**	**100.0%**

Heterogeneity: $I^2 = 64\%$, $\tau^2 = 0.0566$, $p < 0.01$

Figure 2. Forest plot of meta-analysis of 16 trials about the metabolic syndrome and cardiovascular risk.

the different formulas of the variance estimates of logarithm of risk ratio with the conventional Cochran's Q test for testing a null common risk ratio RR across k studies ($H_0 : RR_1 = RR_2 = \cdots = RR_k = RR$) against H_1 : a false H_0 , (*i.e.* $H_1 : RR_i$ has a specific distribution).

We followed the work and the notation of Böhning *et al.* [11] and further proposed some profile likelihood tests by modifying the standard χ^2 test for homogeneity through the various ways of the variance estimates of the logarithm of risk ratios at the i^{th} study.

3.1. Profile Likelihood Estimator under a Fixed Effect Point for a Common Risk Ratio across Studies

The result of the work of Böhning *et al.* [11] under profile likelihood concept provides a fixed-effect point RR for all k studies ($i = 1, \cdots, k$) as

$$RR = \sum_{i=1}^{k} \frac{x_i^T \left(n_i^T RR + n_i^C \right)}{\left(x_i^T + x_i^C \right) n_i^T} = \frac{\sum_{i=1}^{k} x_i^T}{\sum_{i=1}^{k} \frac{\left(x_i^T + x_i^C \right) n_i^T}{n_i^T RR + n_i^C}},$$

leading to the iterative processes of the profile maximum likelihood estimator (PMLE) in the following

$$\hat{R}R_{PMLE} = \sum_{i=1}^{k} \frac{x_i^T \left(n_i^T \hat{R}R_{PMLE} + n_i^C \right)}{\left(x_i^T + x_i^C \right) n_i^T} = \frac{\sum_{i=1}^{k} x_i^T}{\sum_{i=1}^{k} \frac{\left(x_i^T + x_i^C \right) n_i^T}{n_i^T \hat{R}R_{PMLE} + n_i^C}},$$

where x_i^T and x_i^C are the numbers of events in treatment and control arms for each clinical trial i and n_i^T and n_i^C are the numbers of persons at risk or person-times.

3.2. Some Tests Based on Various Formulas of Variance Estimate of Logarithmic *RR_i*

For testing the null hypothesis, the true relative risks (RR_i) are the same in all k centers/studies, $H_0 : RR_1 = RR_2 = \cdots = RR_k = RR$, for $i = 1, \cdots, k$ versus the alternative that at least one of the effect sizes (RR_i) differs from the remainder. Alternatively, this is reasonable to assume that all null parameters of the centers to be combined are summarized into a single underlying population parameter, against the alternative parameters different among centers are likely to have a wholly random with a specific distribution. Our proposed tests are modified on the base of a standard χ^2 test for homogeneity in the following form:

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(\ln\left(\hat{R}R_i\right) - \ln\left(\hat{R}R_{PMLE}\right)\right)^2}{\operatorname{var}\left(\ln\left(\hat{R}R_i\right)\right)}$$

where k is the number of studies being combined, $\hat{R}R_{PMLE}$ is a PMLE of a common RR, $\hat{R}R_i = \left(x_i^T / n_i^T\right) / \left(x_i^C / n_i^C\right)$ is an estimate of RR_i at the i^{th} study, two natural logarithm transformations, such as $\ln\left(\hat{R}R_i\right)$ and $\ln\left(\hat{R}R_{PMLE}\right)$, are needed to adapt the non-symmetric distribution, and $k-1$ is degrees of freedom of χ^2 test. It is a common way that the variance of the logarithm of risk ratios at the i^{th} study, $\operatorname{var}\left(\ln\left(\hat{R}R_i\right)\right)$, is replaced by its various estimates, $\hat{\operatorname{var}}\left(\ln\left(\hat{R}R_i\right)\right)$, leading to the several candidates χ^2 tests, finally.

1) Simply naive χ^2 test (*SIM*), based on variance estimate at the i^{th} study under Poisson likelihood by Delta method, is denoted as

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(\ln\left(\hat{R}R_i\right) - \ln\left(\hat{R}R_{PMLE}\right)\right)^2}{\hat{\operatorname{var}}\left(\ln\left(\hat{R}R_i\right)\right)}$$

where $\hat{\operatorname{var}}\left(\ln\left(\hat{R}R_i\right)\right) = \frac{1}{n_i^T \hat{p}_i^T} + \frac{1}{n_i^C \hat{p}_i^C} = \frac{1}{x_i^T} + \frac{1}{x_i^C}$, $\hat{p}_i^T = x_i^T / n_i^T$, and $\hat{p}_i^C = x_i^C / n_i^C$.

2) Profile likelihood χ^2 test ($PL1$) with the same form above will be obtained but getting the different formula due to the variance estimate under the null hypothesis as

$$\hat{\operatorname{var}}\left(\ln\left(\hat{R}R_i\right)\right) = \left(\frac{1}{n_i^T \hat{R}R_{PMLE}} + \frac{1}{n_i^C}\right)\frac{1}{\hat{p}_i^C}$$

where $\hat{R}R_{PMLE} = \sum_{i=1}^{k} \frac{x_i^T \left(n_i^T \hat{R}R_{PMLE} + n_i^C\right)}{\left(x_i^T + x_i^C\right) n_i^T}$ and $\hat{p}_i^C = x_i^C / n_i^C$.

3) Profile likelihood χ^2 test ($PL2$) will also be obtained after using the different formulas of variance estimate as

$$\hat{\operatorname{var}}\left(\ln\left(\hat{R}R_i\right)\right) = \left(\frac{1}{n_i^T \hat{R}R_{PMLE}} + \frac{1}{n_i^C}\right)\frac{1}{\hat{p}_i^C} = \frac{\left(n_i^C + \hat{R}R_{PMLE} n_i^T\right)^2}{n_i^T n_i^C \hat{R}R_{PMLE}\left(x_i^C + x_i^T\right)}$$

where $\hat{R}R_{PMLE} = \sum_{i=1}^{k} \frac{x_i^T \left(n_i^T \hat{R}R_{PMLE} + n_i^C \right)}{\left(x_i^T + x_i^C \right) n_i^T}$ and $\hat{p}_i^C = \frac{x_i^T + x_i^C}{n_i^C + \hat{R}R_{PMLE} n_i^T}$ are the results of Böhning *et al.* [11] under profile likelihood concept.

4) Cochran's Q test as the weighted sum of squares is distributed as a chi-square statistic with $k-1$ degrees of freedom, under the null of homogeneity of treatment effects across k studies, denoted as

$$Q = \sum_{i=1}^{k} w_i \left(\hat{\delta}_i - \bar{\delta} \right)^2$$

where $\hat{\delta}_i = \ln \hat{R}R_i$, $w_i = \frac{1}{\hat{\text{var}}\left(\hat{\delta}_i\right)} = \frac{1}{n_i^T \hat{p}_i^T} + \frac{1}{n_i^C \hat{p}_i^C}$, $\hat{p}_i^T = x_i^T / n_i^T$, $\hat{p}_i^C = x_i^C / n_i^C$,

and $\bar{\delta} = \sum_{i=1}^{k} w_i \hat{\delta}_i \Big/ \sum w_i$.

4. Monte Carlo Simulation

We perform two simulation plans. One is conducted on type I error for testing a null common risk ratio, *RR*, over all *k* studies or in other words for testing the null homogeneity we have $H_0 : RR_1 = RR_2 = \cdots = RR_k = RR$. The other is used for comparing the performance of tests with the highest power after all test statistics could be controlled within the same limit range of the empirical type I error.

4.1. Simulation Plan for Studying Type I Error

Parameters Setting: The values of parameter setting followed two motivational examples. Let the common relative risk (*RR*) be 1, 2 and 4. Baseline risks p_i^C in the control arm for the *i*th center $(i = 1, 2, \cdots, k)$ are generated from a uniform distribution in which its range depends on the values of *RR*. For example, if $RR = 1$ then $p_i^C \sim U(0, 0.9)$ and the correspondent treatment risks have the possible values less than or equal to 0.9 as $p_i^T = p_i^C \times RR \sim U(0, 0.9)$ for the *i*th center. If $RR = 2$ then $p_i^C \sim U(0, 0.45)$ and $p_i^T = p_i^C \times RR \sim U(0, 0.9)$. The sample sizes n_i^T and n_i^C are distributed from Poisson with the mean of 5, 10, 50, 100, 500, 1000. The number of centers *k* is 4, 16, and 32.

Statistics: Poisson random variables X_i^T and X_i^C in treatment and control arms for center i $(i = 1, 2, \cdots, k)$ are generated with parameters $\left(n_i^T p_i^T \right)$ and $\left(n_i^C p_i^C \right)$, respectively. All candidate tests are then computed. The procedure is replicated 5000 times. From these replicates, the number of the null hypothesis rejections is counted for the actual (empirical) type I error.

Type I error among the tests is considered by comparing the actual (estimated) type I error ($\hat{\alpha}$) with the nominal level of significance (α). The departure of the estimated type I error from the nominal level of significance must not exceed the precise limit. In this study, the evaluation for two-sided tests in terms of the probability is based on Bradley limit [15] yielding the limiting ranges of $[0.5\alpha, 1.5\alpha]$. For an example, at $\alpha = 1\%$ level of significance, $\hat{\alpha}$ value lies between [0.5%, 1.5%], at $\alpha = 5\%$ level of significance, $\hat{\alpha}$ value lies between

[2.5%, 7.5%], at $\alpha = 10\%$ level of significance, $\hat{\alpha}$ value lies between [5%, [15%].

If the empirical type I error lies within the range of Bradley limit, then the statistical test can capture type I error.

4.2. Simulation Plan for Studying Power of Tests

Before comparing the power of test statistics, all test statistics could be calibrated to have the same limit range of type I error rate under the null hypothesis. It means that power comparisons of tests are reliable if all tests are previously based on the same range of type I error rate before the process of power comparisons is employed.

Under the alternative hypothesis that RR_i has been assumed a specific distribution around the mean (RR_0) of 1, 2, 4, we let $\ln RR_i = \ln RR_0 + U_m = \ln RR_0 + mm(2U - 1)$ where U_m is a uniform over (-*mm*, *mm*) for a given *mm* = 0.2, 0.4, 0.6, and U is a uniform over (0, 1). Baseline risks p_i^C are still generated from a uniform distribution over [0, 0.25]. Poisson random variables X_i^T and X_i^C are generated with parameters $\left(n_i^T p_i^T\right)$ and $\left(n_i^C p_i^C\right)$, respectively. The procedure is replicated 5000 times and the number of the null hypothesis rejections is counted for the empirical power.

5. Results

Since it is difficult to present all enormous results from the simulation study, we just have illustrated some instances, coping with 0.05 levels of significances, some common true relative risk values of 1 and 2, in both equal and unequal sample sizes.

5.1. Equal Sample Sizes ($n_i^T = n_i^C$)

5.1.1. Studying Type I Errors

- From Table 1, the results show that for small sample sizes ($n_i^T, n_i^C \le 10$), regardless of study size *k*, almost all tests cannot control type I error.
- For moderate to large study sizes ($k \ge 16$) in combination with moderate to large sample sizes ($n_i^T, n_i^C \ge 50$), two proposed tests (*PL*1, *PL*2) can maintain type I error rates in almost all situations. Meanwhile, for moderate to large study sizes ($k \ge 16$), the Q test seems to handle type I error when sample sizes are large ($n_i^T, n_i^C \ge 500$).
- For small center size (k = 4), the *SIM* test can capture type I error on some moderate and large sample sizes ($n_i^T, n_i^C \ge 100$) and the Q test can control type I error on sample size being moderate ($n_i^T, n_i^C = 50,100$).
- In summary, for study size is moderate to large ($k \ge 16$), two profile likelihood tests (*PL*1 and *PL*2) perform well with maintaining type I error rates when sample sizes are moderate to large ($n_i^T, n_i^C \ge 50$); in the meanwhile, the Q test can capture type I error on sample size being quite large ($n_i^T, n_i^C \ge 500$).

Table 1. At 5% significance level, a comparison of the empirical type I error rates among four statistical tests with the equal in the mean of sample sizes.

RR	*k*	$n_i^T = n_i^C$	*SIM*	*PL1*	*PL2*	*Q*
1	4	5	0.00	0.30	0.10	0.02
		10	0.00	2.24	1.38	0.04
		50	1.64	12.98	13.78	**2.84**
		100	**3.04**	14.74	15.40	**5.54**
		500	**5.40**	14.44	15.62	8.82
		1000	**6.06**	14.34	14.74	9.00
	16	5	0.00	0.00	0.00	0.00
		10	0.00	0.04	0.00	0.00
		50	0.10	8.24	**6.22**	0.48
		100	0.52	8.54	**6.94**	2.34
		500	1.76	**5.80**	**4.92**	**4.52**
		1000	**2.66**	**5.64**	**4.94**	**6.10**
	32	5	0.00	0.00	0.00	0.00
		10	0.00	0.00	0.00	0.00
		50	0.00	**6.54**	**4.44**	0.12
		100	0.18	**6.40**	**4.48**	0.90
		500	1.10	**4.56**	**3.96**	**4.38**
		1000	1.08	**3.42**	**2.76**	**4.92**
2	4	5	0.02	2.46	0.60	0.66
		10	0.10	**6.04**	2.48	1.24
		50	1.86	11.46	13.38	**4.66**
		100	**3.22**	12.42	14.56	**6.38**
		500	**4.44**	13.38	13.76	9.08
		1000	**5.36**	13.02	13.72	9.82
	16	5	0.00	0.24	0.00	0.00
		10	0.00	1.26	0.06	0.04
		50	0.10	**5.58**	**6.02**	1.72
		100	0.68	**5.48**	**5.62**	**2.94**
		500	1.60	**4.36**	**4.18**	**6.08**
		1000	1.74	**3.82**	**3.92**	**6.56**
	32	5	0.00	0.02	0.00	0.00
		10	0.00	0.40	0.00	0.00
		50	0.06	**4.16**	**3.62**	0.84
		100	0.14	**3.34**	**3.82**	1.82
		500	0.70	**2.58**	**2.75**	**5.24**
		1000	0.70	1.98	1.76	**5.36**

Note: Bold values indicate that the test statistics can control type I error rate.

5.1.2. Comparing Powers of Tests

- The process of power comparisons is conducted after all candidate tests can previously maintain the same limit range of type I error.
- Table 2 showed that both of the *PL*1 test and the *PL*2 test are best with the highest powers when study size is moderate to large ($k \geq 16$) and sample sizes are moderate ($n_i^T, n_i^C = 50, 100$) in every degrees of variation (*mm* = 0.2, 0.4, 0.6), coping with *RR* = 1, 2. Additionally, in more detail, *PL*2 seems better than *PL*1 with higher power.
- When study size is moderate to large ($k \geq 16$) and sample size is large ($n_i^T, n_i^C \geq 500$), the *Q* test is best with the highest power of test in every degrees of variation (*mm* = 0.2, 0.4, 0.6), coping with *RR* = 1, 2.
- For the number of studies is small (*k* = 4) in combination with large sample sizes ($n_i^T, n_i^C \geq 500$), the best performance of test is the *SIM* test since it is only one test that can formerly meet the criterion of controlling type I error.

5.2. Unequal Cases ($n_i^T \neq n_i^C$)

5.2.1. Studying Type I Errors

- Table 3 indicates that for *RR* = 2 and moderate to large study size ($k \geq 16$), three tests (*PL*1, *PL*2, *Q*) can capture type I error when both of sample sizes in treatment and control groups are moderate to large ($n_i^T \geq 50$, $n_i^C \geq 100$). The *SIM* cannot control type I error in every case of sample sizes.
- Table 4 is considered to highlight only for small study sizes (*k* = 4). For small study sizes (*k* = 4), the *SIM* seems to control type I error at least when one sample size of treatment groups is large. Both of *PL*1 and *PL*2 tests can control type I error when one sample size of treatment groups is small. The *Q* test can rarely control type I error in every sample size for small study sizes.

5.2.2. Studying Power of Tests

- Table 5 indicates that for moderate to large study sizes ($k \geq 16$) in combination with moderate sample sizes ($n_i^T = 50, n_i^C = 100$), two proposed tests (*PL*1, *PL*2) perform best and quite close together.
- For moderate to large study sizes ($k \geq 16$) in combination of at least one treatment arm being large sample sizes ($n_i^T \geq 50, n_i^C \geq 500$), *Q* test seems to have best performance with the highest power, followed by *PL*2 and *PL*1 tests.
- Additionally, when the sample sizes of both treatment and control arms are quite small ($n_i^T, n_i^C = 5, 10$), regardless of study size (*k*), no tests among four tests (*SIM*, *PL*1, *PL*2, *Q*) are reasonable since almost all tests cannot control type I error rates and they give too low powers.

6. Discussion

In this study, we further focus on a comparison of the performance among four statistical tests including the simply naive test approach (*SIM*), the conventionally null approach of profile likelihood (*PL*1), the full profile likelihood approach

Table 2. Comparisons of the power of tests after capturing type I error at 0.05 significance level when means of sample sizes in treatment groups are equal ($n_i^T = n_i^C = n$).

mm	k	n	$RR = 1$				$RR = 2$			
			SIM	$PL1$	$PL2$	Q	SIM	$PL1$	$PL2$	Q
0.2	4	50	1.80	14.14	15.54	**3.48**	2.32	13.08	15.68	**5.74**
		100	**4.50**	18.26	18.78	**7.84**	**4.98**	15.94	18.22	**9.46**
		500	**17.66**	33.14	33.88	24.64	**21.18**	36.32	37.38	31.06
		1000	**32.78**	48.18	48.86	40.68	**39.72**	54.92	56.06	50.90
	16	50	0.16	10.38	**7.66**	0.68	0.38	**7.34**	**7.68**	2.52
		100	1.18	11.50	**10.38**	4.18	2.14	**9.36**	**10.80**	**8.18**
		500	20.78	**30.96**	**31.34**	**35.46**	28.56	**37.36**	**38.92**	**51.72**
		1000	**51.28**	**59.90**	**60.74**	**68.30**	67.32	**73.38**	**74.12**	**83.70**
	32	50	0.02	**8.52**	**6.58**	0.22	0.04	**5.32**	**6.42**	1.64
		100	0.50	**11.60**	**10.04**	3.24	1.00	**8.20**	**9.52**	7.62
		500	27.24	**40.32**	**40.34**	**50.70**	39.28	**48.48**	**50.28**	**70.96**
		1000	72.04	**78.52**	**79.16**	**87.56**	86.30	**89.34**	**90.22**	**96.74**
0.4	4	50	3.18	19.22	21.20	**6.04**	5.46	19.68	22.52	**11.22**
		100	**10.42**	27.68	29.74	**15.98**	**13.82**	30.06	33.66	**22.92**
		500	**54.58**	68.90	69.08	61.94	**64.94**	76.50	76.70	72.74
		1000	**77.72**	85.92	85.94	81.82	**84.96**	90.46	90.82	88.72
	16	50	0.70	16.06	**15.44**	2.94	2.10	**15.30**	**17.50**	8.90
		100	7.68	29.48	**28.54**	18.18	14.70	**30.82**	**34.54**	**33.86**
		500	86.10	**89.84**	**90.84**	**92.24**	94.84	**96.34**	**96.88**	**98.10**
		1000	**99.02**	**99.32**	**99.44**	**99.56**	99.84	**99.96**	**99.92**	**99.96**
	32	50	0.20	**17.70**	**15.24**	2.02	1.42	**17.24**	**19.08**	10.48
		100	8.20	**34.50**	**33.76**	22.70	15.72	**36.70**	**41.66**	45.10
		500	98.18	**99.02**	**99.02**	**99.36**	99.60	**99.76**	**99.78**	**99.96**
		1000	100.0	**100.0**	**100.0**	**100.0**	100.0	**100.0**	**100.0**	**100.0**
0.6	4	50	7.36	28.50	31.06	**12.62**	11.96	31.24	34.22	**21.18**
		100	**22.40**	43.82	46.36	**30.48**	32.30	51.14	53.48	**43.82**
		500	**80.78**	87.62	87.80	84.28	**86.58**	91.86	92.08	90.00
		1000	**92.58**	95.42	95.48	93.86	**95.36**	97.04	97.02	96.22
	16	50	4.00	30.32	**31.50**	11.44	10.36	**34.56**	**38.26**	27.96
		100	32.52	57.08	**60.66**	51.06	51.52	**68.12**	**72.20**	**74.22**
		500	99.42	**99.60**	**99.72**	**99.70**	99.84	**99.94**	**99.94**	**99.96**
		1000	**99.96**	**99.98**	**99.98**	**99.98**	99.98	**100.0**	**100.0**	**100.0**
	32	50	3.62	**39.78**	**40.16**	14.12	12.26	**44.48**	**49.34**	39.50
		100	46.78	**73.78**	**76.76**	69.80	71.48	**85.24**	**87.64**	91.08
		500	99.98	**100.0**	**100.0**	**100.0**	100.0	**100.0**	**100.0**	**100.0**
		1000	100.0	**100.0**	**100.0**	**100.0**	100.0	**100.0**	**100.0**	**100.0**

Note: Bold values indicate that the statistic tests can previously control type I error rates.

Table 3. At 5% significance level, a comparison of the estimated type I error rates for moderate to large study sizes ($k \geq 16$) with the unequal sample sizes ($n_i^T \neq n_i^C$).

RR	*k*	n_i^T	n_i^C	*SIM*	*PL*1	*PL*2	*Q*
2	16	50	100	0.16	**6.06**	**5.00**	1.72
			500	0.60	**6.16**	**5.02**	**5.64**
			1000	0.72	**6.74**	**6.14**	**7.10**
		100	500	0.78	**5.68**	**4.36**	**4.80**
			1000	0.70	**5.18**	**4.52**	**6.22**
		500	1000	1.34	**3.82**	**3.54**	**6.02**
2	32	50	100	0.02	**4.26**	**2.82**	0.96
			500	0.18	**4.64**	**3.50**	**4.42**
			1000	0.16	**4.12**	**3.62**	**6.24**
		100	500	0.14	**3.66**	2.38	**3.88**
			1000	0.22	**3.24**	2.52	**4.96**
		500	1000	0.60	1.78	1.46	**5.34**

Note: Bold words denoting the statistic test can control type I error.

Table 4. At 5% significance level, a comparison of the actual type I error rates for small study size (k = 4) with the unequal sample sizes ($n_i^T \neq n_i^C$).

k	*RR*	n_i^T	n_i^C	*SIM*	*PL*1	*PL*2	*Q*
4	1	5	10	0.02	0.46	0.74	0.40
			50	**5.70**	**4.48**	**6.82**	15.02
			100	12.92	**4.88**	7.82	25.12
			500	26.04	**5.26**	**6.14**	39.40
			1000	27.68	**5.60**	**5.94**	41.38
		10	50	1.50	**2.72**	**4.46**	**5.26**
			100	**4.26**	**3.34**	**4.74**	10.54
			500	14.60	**4.02**	**4.28**	22.70
			1000	16.18	**4.16**	**4.38**	24.70
		50	100	2.18	14.34	14.60	**4.50**
			500	**4.66**	16.06	16.04	8.02
			1000	**5.20**	15.78	15.64	8.98
		100	500	**4.46**	15.72	15.66	**7.42**
			1000	**5.50**	17.54	17.42	9.22
		500	1000	**5.82**	16.36	15.76	9.44
4	2	5	10	0.00	2.24	1.02	0.18
			50	0.76	1.96	**3.32**	**6.48**
			100	2.00	2.20	**3.22**	12.82
			500	9.78	2.14	2.46	25.94
			1000	12.56	2.34	**2.56**	29.68
		10	50	0.58	**5.82**	**6.66**	**3.54**

Continued

			100	1.62	**6.82**	7.58	7.68
			500	**4.70**	**6.18**	**6.40**	16.08
			1000	**6.54**	**6.10**	**6.04**	18.30
		50	100	2.46	12.20	13.68	**5.52**
			500	**2.68**	11.72	11.72	8.14
			1000	**2.92**	11.82	11.58	8.44
		100	500	**2.80**	12.00	11.82	**7.50**
			1000	**3.66**	12.46	12.16	9.04
		500	1000	**4.62**	12.64	12.78	9.36
4	4	5	10	0.12	**4.34**	1.26	**3.40**
			50	0.02	**2.96**	**3.30**	1.78
			100	0.26	1.86	2.08	**3.94**
			500	1.90	1.44	1.50	11.68
			1000	**3.78**	1.44	1.38	15.54
		10	50	0.12	**5.18**	**5.82**	**3.08**
			100	0.26	**4.18**	**4.42**	**4.20**
			500	1.24	**3.66**	**3.68**	10.82
			1000	1.46	**3.66**	**3.50**	13.04
		50	100	1.92	8.86	10.92	**6.80**
			500	1.14	**6.22**	**6.26**	8.36
			1000	1.10	**6.40**	**6.16**	9.24
		100	500	1.86	**7.48**	7.68	7.80
			1000	1.84	**6.92**	**6.46**	9.44
		500	1000	**3.24**	8.88	9.26	8.48

Note: Bold words denoting the statistic test can control type I error.

Table 5. Comparison of the power of tests at 0.05 significance level for moderate to large study sizes ($k \geq 16$) with the unequal sample sizes ($n_i^T \neq n_i^C$) at $mm = 0.2$.

RR	*k*	n_i^T	n_i^C	*SIM*	*PL*1	*PL*2	*Q*
2	16	50	100	0.66	**9.64**	**8.06**	4.92
			500	1.88	**12.50**	**10.90**	**14.74**
			1000	2.66	**12.98**	**11.92**	**20.00**
		100	500	5.54	**16.98**	**15.28**	**24.08**
			1000	6.50	**17.86**	**16.88**	**30.72**
		500	1000	48.96	**57.50**	**57.88**	**73.18**
2	32	50	100	0.06	**7.38**	**5.90**	3.44
			500	1.56	**11.32**	**9.88**	**19.84**
			1000	1.50	**11.80**	**10.82**	**25.12**
		100	500	4.00	**16.24**	**14.08**	**30.96**
			1000	6.38	**19.26**	**17.70**	**41.68**
		500	1000	67.72	74.88	74.90	**90.86**

Note: Bold words denoting the statistic test can formerly control type I error.

(*PL*2), and the conventionally weighted sum of square approach (*Q*). The main results found in the followings.

- No tests could not capture type I error rates for small sample sizes ($n_i^T, n_i^C \le 10$), regardless of study size *k*. This same result happened to the study of Mathes and Kuss [16]; they stated that estimating between-study heterogeneity in meta-analysis of a small number of sample sizes ($n_i^T, n_i^C \le 5$) is difficult in this situation.
- The work of Willis and Riley [17] was also confirmed the properties of Q test to be a good test when there are large study sizes (50 studies or more), but for fewer studies the Q test has the low power.
- We are scientist group that have attempted to propose some new/modified tests to bridge the gaps of limitation of the Q test. The idea of this paper shows how to use two proposed tests (*PL*1, *PL*2) based on substituting profile maximum likelihood estimates into the different variance formulas for obtaining the modified standard chi-square tests of heterogeneity.
- Our profile likelihood tests (*PL*1 and *PL*2) for moderate to large study sizes ($k \ge 16$) in combination with moderate sample sizes ($n_i^T, n_i^C = 50,100$) can defeat the Q test with the higher power after capturing the same range of type I error limits.
- The work of Bagheri, Ayatollahi and Jafari [18] and Viechtbauer [19] which also could evaluate the influence of the size of centers (*k*) and sample sizes (n_i^T, n_i^C) on the type I error and the power for the null homogeneity testing in some situations. It means that the investigators should pursue their attempts to find some new/modified tests further.
- In contrast, although two proposed tests (*PL*1, *PL*2) perform well in above situations, they cannot defeat the Q test when the number of studies is moderate to large ($k \ge 16$) in combination with large sample sizes ($n_i^T, n_i^C \ge 500$). Additionally, in unbalanced cases, for moderate to large study sizes ($k \ge 16$) and combination of moderate sample size and large sample sizes ($n_i^T \ge 50, n_i^C \ge 500$), the Q test performs best with the highest power, followed by *PL*2 and *PL*1 tests.

7. Conclusion

In summary, the idea of replacement of profile likelihood estimates into the variance formulas of logarithm of relative risks works well when $k \ge 16$ in combination with ($n_i^T, n_i^C = 50,100$).

8. Recommendation

Two proposed tests (*PL*1, *PL*2) based on substituting profile maximum likelihood estimates into the different variance formulas, perform best with the highest power (under formerly within the same range of type I error limits) in some situations, for examples, when the number of studies is moderate to large sizes ($k \ge 16$) in combination with moderate sample sizes ($n_i^T, n_i^C = 50,100$). This result leads to the suggestion of the use of two proposed tests in such practical situations.

In contrast, although two proposed tests (*PL*1, *PL*2) perform well with the high powers in above situations, they cannot defeat the Q test when numbers of studies are moderate to large ($k \geq 16$) in combination with large sample sizes ($n_i^T, n_i^C \geq 500$) in both balanced and unbalanced cases. This result leads to the suggestion to use the Q test in these situations. It means that it should be further investigated to find the new appropriate test to fill the gaps of low power of Q test in such situations.

Acknowledgements

The study was partially supported for publication by Faculty of Public Health, Mahidol University, Bangkok, Thailand.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Lipsitz, S.R., Dear, K.B.G., Laird, N.M. and Molenberghs, G. (1998) Tests for Homogeneity of the Risk Difference When Data Are Sparse. *Biometrics*, **54**, 148-160. https://doi.org/10.2307/2534003

[2] Lui, K.J. (2007) Testing Homogeneity of the Risk Ratio in Stratified Noncompliance Randomized Trials. *Contemporary Clinical Trials*, **28**, 614-625. https://doi.org/10.1016/j.cct.2007.02.010

[3] Smolinsky, L. and Marx, B.D. (2018) Odds Ratios, Risk Ratios, and Bornmann and Haunschild's New Indicators. *Journal of Informetrics*, **12**, 732-735. https://doi.org/10.1016/j.joi.2018.06.011

[4] Améndola, C., *et al.* (2019) The Maximum Likelihood Degree of Historic Varieties. *Journal of Symbolic Computation*, **92**, 222-242. https://doi.org/10.1016/j.jsc.2018.04.016

[5] Kulinskaya, E. and Dollinger, M.B. (2015) An Accurate Test for Homogeneity of Odds Ratios Based on Cochran's Q-Statistic. *BMC Medical Research Methodology*, **15**, 49. https://doi.org/10.1186/s12874-015-0034-x

[6] Boissel, J.-P., Blanchard, J., Panak, E., Peyrieux, J.-C. and Sacks, H. (1989) Considerations for the Meta-Analysis of Randomized Clinical Trials: Summary of a Panel Discussion. *Controlled Clinical Trials*, **10**, 254-281. https://doi.org/10.1016/0197-2456(89)90067-6

[7] Kulinskaya, E., Dollinger, M.B. and Bjørkestøl, K. (2011) On the Moments of Cochran's Q Statistic under the Null Hypothesis with Application to the Meta-Analysis of Risk Difference. *Research Synthesis Methods*, **2**, 254-270. https://doi.org/10.1002/jrsm.54

[8] Fleiss, J.L. (1986) Analysis of Data from Multiclinic Trials. *Controlled Clinical Trials*, **7**, 267-275. https://doi.org/10.1016/0197-2456(86)90034-6

[9] Shadish, W.R. and Haddock, C.K. (1994) The Handbook of Research Synthesis. Russell Sage Foundation.

[10] Ferrari, S.L., Lucambio, F. and Cribari-Neto, F. (2005) Improved Profile Likelihood Inference. *Journal of Statistical Planning and Inference*, **134**, 373-391. https://doi.org/10.1016/j.jspi.2004.05.001

[11] Böhnimg, D., Kuhnert, R. and Rattanasiri, S. (2008) Meta-Analysis of Binary Data Using Profile Likelihood. Chapman & Hall/CRC Press, Boca Raton.

[12] Farquhar, C.M., Marjoribanks, J., Lethaby, A. and Basser, R. (2007) High Dose Chemotherapy for Poor Prognosis Breast Cancer: Systematic Review and Meta-Analysis. *Cancer Treatment Reviews*, **33**, 325-337. https://doi.org/10.1016/j.ctrv.2007.01.007

[13] Schwarzer, G. (2007) Meta: An R Package for Meta-Analysis. *R News*, **7**, 40–45. https://cran.r-project.org/doc/Rnews/Rnews_2007-3.pdf

[14] Mottillo, S., Filion, K.B., Genest, J., Joseph, L., Pilote, L., Poirier, P., *et al.* (2010) The Metabolic Syndrome and Cardiovascular Risk: A Systematic Review and Meta-Analysis. *Journal of the American College of Cardiology*, **56**, 1113-1132. https://doi.org/10.1016/j.jacc.2010.05.034

[15] Bradley, J. (1978) Robustness? *British Journal of Mathematical and Statistical Psychology*, **31**, 144-152. https://doi.org/10.1111/j.2044-8317.1978.tb00581.x

[16] Mathes, T. and Kuss, O. (2018) A Comparison of Methods for Meta-Analysis of a Small Number of Studies with Binary Outcomes. *Research Synthesis Methods*, **9**. https://doi.org/10.1002/jrsm.1296

[17] Willis, B.H. and Riley, R.D. (2017) Measuring the Statistical Validity of Summary Meta-Analysis and Meta-Regression Results for Use in Clinical Practice. *Statistics in Medicine*, **36**, 3283-3301. https://doi.org/10.1002/sim.7372

[18] Bagheri, Z., Ayatollahi, S.M.T. and Jafari, P. (2011) Comparison of Three Tests of Homogeneity of Odds Ratios in Multicenter Trials with Unequal Sample Sizes within and among Centers. *BMC Medical Research Methodology*, **11**, 58. https://doi.org/10.1186/1471-2288-11-58

[19] Viechtbauer, W. (2007) Hypothesis Tests for Population Heterogeneity in Meta-Analysis. *British Journal of Mathematical and Statistical Psychology*, **60**, 29-60. https://doi.org/10.1186/1471-2288-11-58