Computational Statistics & Data Analysis 33 (2000) 299–313

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Testing the significance of a common risk difference in meta-analysis

Julio Sánchez-Meca*, Fulgencio Marín-Martínez

University of Murcia, Faculty of Psychology, Dept Psicología Básica y Metodología, Campus de Espinardo, Apdo 4021. 30080-Murcia, Spain

Received 1 November 1998; received in revised form 1 March 1999

Abstract

Using the Monte Carlo simulation, we estimated the statistical power and Type I error rates of five procedures for testing the significance of a common risk difference in a set of independent 2 × 2 tables. It was found that the unweighted procedure for testing the significance of a common risk difference showed Type I error rates systematically larger than the nominal significance level, and that its power was lower than that of the other procedures. The conditional weighted procedure showed the worst performance, with remarkably anomalous results under many of the conditions. Cochran's, Mantel–Haenszel's, and Yusuf's unconditional weighted procedures showed very similar results, with the best performance in both Type I error values and power values. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Meta-analysis; Effect size; Statistical power; Type I error; Risk difference

1. Introduction

Meta-analysis has become a very common research methodology in behavioural and health sciences. It can be defined as the quantitative analysis of the results of a set of studies about a given research topic (Cooper and Hedges, 1994; Greenland, 1987; Kuss and Koch, 1996). To carry out a meta-analysis an effect size index that represents the outcome of each study has to be selected and, based on this, analysis techniques are applied to achieve three main objectives: (a) To estimate the average effect of the studies; (b) to test whether the set of studies is homogeneous, and

* Corresponding author.
E-mail address: jsmeca@fcu.um.es (J. Sánchez-Meca).

PII: S 0167-9473(99)00055-9

(c) if the homogeneity hypothesis is not met, to test the influence of potential moderator variables that explain such heterogeneity.

Depending on the type of design and the nature of the outcome, there are different effect size indexes that can be used to summarise the results of an empirical study. In health sciences, as well as in behavioural sciences, the results of the study are frequently given as a contingency 2×2 table, in which a dichotomous variable (usually called 'success' vs. 'failure') is registered for the two groups of subjects (usually, treated vs. control). If the two groups of subjects have been randomly assigned to each condition, then we have a randomised clinical trial.

When, in a meta-analysis, we have a set of randomised clinical trials and their results are given as 2×2 tables, the most recommended effect size indexes are risk difference, rate ratio, and odds ratio (Haddock et al., 1998; Laird and Mosteller, 1990). Given a set of k independent studies each of them composed of two groups of subjects (treated and control) with sample sizes $n_{\mathrm{T}i}$ and $n_{\mathrm{C}i}$, respectively, and with success proportions $p_{\mathrm{T}i}$ and $p_{\mathrm{C}i}$, the risk difference is defined as $d_i = p_{\mathrm{T}i} - p_{\mathrm{C}i}$; the rate ratio is given as the ratio between the two success proportions, $rr_i = p_{\mathrm{T}i}/p_{\mathrm{C}i}$; and the odds ratio represents the relative gain of one group in respect to the other, $or_i = p_{\mathrm{T}i}(1 - p_{\mathrm{C}i})/p_{\mathrm{C}i}(1 - p_{\mathrm{T}i})$ (Ahn, 1997). The advantages and limitations of these indexes have been explored elsewhere (Fleiss, 1994; Hasselblad et al., 1995). Risk difference has the advantage of ease of interpretability, because it is the natural measure of the gain of one treatment over another, although its range depends on success proportions, $p_{\mathrm{T}i}$ and $p_{\mathrm{C}i}$.

In this paper we have focused our attention on the application of risk difference as an effect size index to summarise independent 2×2 tables derived from a set of k randomised clinical trials for testing the effectiveness of a treatment (medical, pharmacological, psychological) in comparison with a non-treated control group. Consequently, we will assume that the control group proportions in the studies, $p_{\mathrm{C}i}$, are approximately homogeneous and not particularly extreme. We will also assume a fixed-effects model in which the variability among d_i values is simply caused by within-study variance. Thus, the k independent studies come from the same population with a common population risk difference $\delta = \pi_{\mathrm{T}} - \pi_{\mathrm{C}}$, δ being the population risk difference and π_{T} and π_{C} the population success proportions in treated and control groups, respectively (Laird and Mosteller, 1990).

The global treatment effectiveness will be estimated calculating a common risk difference, $\bar{d}$, applying several weighting procedures to each single risk difference, d_i. Assuming a fixed-effects model, the weights are obtained as a function of the within-study variance of each single risk difference.

In meta-analysis one of the main objectives is to examine the statistical significance of the common risk difference to determine whether or not the effect size is null, thereby testing the hypothesis $\mathrm{H}_0: \delta = 0$. Although some interest has been shown in comparing the performance of different common risk differences under the random-effects model (Emerson et al., 1993, 1996), this has not been the case with fixed-effects models. Consequently, in this paper we will compare five statistical tests.

Under $\mathrm{H}_0: \delta = 0$, and assuming large sample sizes, $n_{\mathrm{T}i}$ and $n_{\mathrm{C}i}$, in the individual studies and a high number of studies, k, the tests are distributed as χ^2_1. However, in multiple real situations meta-analyses are applied to a few studies with low samples sizes. Moreover, the influence that such factors as the inbalance between sample sizes, $n_{\mathrm{T}i}$ and $n_{\mathrm{C}i}$, or the relationship between samples sizes and treated and control groups have on Type I and Type II errors, has not yet been tested. The objective of our study was to test by means of the Monte Carlo simulation: (a) if there were similarities between empirical and asymptotical Type I and Type II error rates in the five tests, and (b) if one test performs better than the others under the different conditions and parameters manipulated.

2. Procedures for testing the significance of a common risk difference

From the k studies, a δ estimate can be obtained by calculating a weighted mean, $\bar{d}$, of the individual risk differences, d_i:

$$\bar{d} = \sum^{k} W_i d_i \bigg/ \sum^{k} W_i. \tag{1}$$

Assuming a fixed-effects model, the optimal weight of each d_i, W_i, is obtained as the inverse of within-group variance, $\sigma^2_{d_i}$, that is, $W_i = 1/\sigma^2_{d_i}$. But the d_i variance is unknown, because it is a function of sample sizes, $n_{\mathrm{T}i}$ and $n_{\mathrm{C}i}$, and unknown population proportions, π_{T} and π_{C}. Several significance tests implying certain modifications in the weights were proposed in order to achieve a good adjustment of Type I and Type II error rates. Table 1 shows the formulas for the statistical procedures.

In the conditional weighted test, χ^2_{CW}, d_i and w_i are both a function of $p_{\mathrm{T}i}$ and $p_{\mathrm{C}i}$, and this departure of independence can affect in testing the null hypothesis $\mathrm{H}_0 : \delta = 0$

Table 1
Mathematical details of the five statistical procedures

Procedure	Weighting factor
	$w_i = (s^2_{d_i})^{-1}$
$\chi^2_{\mathrm{CW}} = \frac{\left(\sum^k w_i d_i\right)^2}{\sum^k w_i}$	$= \left[\frac{p_{\mathrm{T}i}(1-p_{\mathrm{T}i})}{n_{\mathrm{T}i}} + \frac{p_{\mathrm{C}i}(1-p_{\mathrm{C}i})}{n_{\mathrm{C}i}}\right]^{-1}$
$\chi^2_{\mathrm{C}} = \frac{\left(\sum^k w^*_i d_i\right)^2}{\sum^k w^*_i \bar{p}_i(1-\bar{p}_i)}$	$w^*_i = \frac{n_{\mathrm{T}i} n_{\mathrm{C}i}}{n_{\mathrm{T}i}+n_{\mathrm{C}i}}$
$\chi^2_{\mathrm{MH}} = \frac{\left(\lvert \sum^k w^*_i d_i \rvert - 0.5\right)^2}{\sum^k \tilde{w}_i \bar{p}_i(1-\bar{p}_i)}$	$\tilde{w}_i = \frac{n_{\mathrm{T}i} n_{\mathrm{C}i}}{n_{\mathrm{T}i}+n_{\mathrm{C}i}-1}$
$\chi^2_{Y} = \frac{\left(\sum^k w^*_i d_i\right)^2}{\sum^k \tilde{w}_i \bar{p}_i(1-\bar{p}_i)}$	$w^*_i = \frac{n_{\mathrm{T}i} n_{\mathrm{C}i}}{n_{\mathrm{T}i}+n_{\mathrm{C}i}}$
	$\tilde{w}_i = \frac{n_{\mathrm{T}i} n_{\mathrm{C}i}}{n_{\mathrm{T}i}+n_{\mathrm{C}i}-1}$
$\chi^2_{\mathrm{U}} = \frac{\bar{d}^2_{\mathrm{u}}}{S^2_{\bar{d}_{\mathrm{u}}}}$	—

Note. χ^2_{CW}=conditional weighted test; χ^2_{C}=Cochran's test; χ^2_{MH}=Mantel–Haenszel's test; χ^2_{Y}=yusuf et al.'s test; χ^2_{U}=Unweighted test.

(Fleiss, 1981, p. 163; Shadish and Haddock, 1994, p. 270). Under $H_0 : \delta = 0$, χ^2_{CW} is approximately distributed as χ^2_1; thus, the null effect hypothesis is rejected when $\chi^2_{CW} \geq_{1-\alpha} \chi^2_1$, with α as the significance level.

A practical problem arises in calculating $S^2_{d_i}$ when $p_{Ti} = p_{Ci} = 0$, because in this situation $S^2_{d_i} = 0$. To produce a non-zero value of $S^2_{d_i}$ we adapted an adjustment proposed by Tukey (1977, Chapter 15), that replaces a proportion $p = x/n$ by $p^* = (x+1/6)/(n+1/3)$, where x is the number of successes and it is distributed binomially (n, π). This adjustment was only applied in calculating $S^2_{d_i}$.

To avoid the dependence between d_i and w_i, Cochran (1954) proposed the χ^2_C test (see Table 1), where $\bar{p}_i = (x_{Ti} + x_{Ci})/(n_{Ti} + n_{Ci})$, x_{Ti} and x_{Ci} being the success numbers in treated and control groups, respectively, in the $i$esim study. Under H_0, χ^2_C is approximately distributed as χ^2_1, provided that the number of studies, k, and sample sizes, n_{Ti} and n_{Ci}, are large.

Mantel and Haenszel (1959) proposed two modifications to Cochran's test which consisted in adding the usual one-half continuity correction for χ^2 and in an adjustment of the denominator to obtain an unbiased estimate of variance (see Table 1). With a sufficiently large number of studies, k, χ^2_{MH} is still asymptotically χ^2_1, even when the sample sizes, n_{Ti} and n_{Ci}, are small. When $|\sum^k w_i^* d_i| < 0.5$, χ^2_{MH} is taken as 0. On the other hand, Yusuf et al. (1985) proposed the Mantel–Haenszel test with the 0.5 continuity correction omitted. Under H_0, χ^2_Y is also approximately distributed as χ^2_1.

The χ^2_C, χ^2_{MH}, and χ^2_Y tests are very much alike and they must yield similar results. However, under certain conditions they may offer different power and Type I error rates; in particular, when k and sample sizes n_{Ti} and n_{Ci} are small. Moreover, their performance with unbalanced sample sizes is unknown, as when the relationship between sample sizes and the extremity of proportions is manipulated.

Finally, the procedure for testing the significance of an unweighted common risk difference, χ^2_U, will be considered. Under H_0, the χ^2_U test follows an approximately χ^2_1, $S_{\bar{d}_u}^2$ being the estimated variance of the sampling distribution of the unweighted common risk difference: $S^2_{\bar{d}_u} = \hat{S}^2_{d_i}/k$, and the unweighted common risk difference is given by $\bar{d}_u = \sum^k d_i/k$.

The χ^2_U test is an inappropriate procedure for testing the significance of a common risk difference because it does not weigh the individual risk differences. We have included it in our simulation study only for the purpose of comparison, and in order to show its low statistical power in relation to the other procedures.

3. Design of the simulation study

The simulation study was carried out on an IBM-PC Pentium/133 MHz machine using GAUSS (Aptech Systems, 1992). Two binomially distributed populations were defined, $B(n_T, \pi_T)$ and $B(n_C, \pi_C)$, where π_T and π_C were the treated and control proportions, respectively. From these populations, pairs of independent random samples were generated with n_T and n_C as sample sizes. The simulated studies were carried out given a treated and a control group and a dichotomuos outcome variable. Thus, a 2 × 2 table represented the data of each simulated study.

Each 2 × 2 table simulated the data in an empirical study, in which the sample risk difference ($d_i = p_{\mathrm{T}i} - p_{\mathrm{C}i}$) was computed. A set of k 2 × 2 independent tables simulating the data of a meta-analysis was run, yielding k risk differences. All the 2 × 2 tables within the same meta-analysis estimated a common population risk difference, $\delta = \pi_{\mathrm{T}} - \pi_{\mathrm{C}}$.

The average risk difference and its statistical significance were computed in accordance with the five procedures mentioned above. To determine the Type I error rate, the common population risk difference was fixed as $\delta = 0$. To examine the statistical power, δ values other than 0 were defined.

The following parameters were manipulated: (1) the average sample size of each meta-analysis, $\overline{N}$ ($\overline{N} = \sum^k N_i/k$; being $N_i = n_{\mathrm{T}i} + n_{\mathrm{C}i}$), with values 60, 100, and 160; (2) the ratio between sample sizes of the two groups in each study, with the three conditions $n_{\mathrm{T}} = n_{\mathrm{C}}$, $n_{\mathrm{T}} = 2n_{\mathrm{C}}$, and $n_{\mathrm{T}} = 4n_{\mathrm{C}}$; (3) the number of studies, k, with values 10, 20, and 40; (4) the population risk difference, with values $\delta = \pi_{\mathrm{T}} - \pi_{\mathrm{C}} = 0$, 0.05, and 0.10; (5) the position in the range of the two population proportions, differentiating between a central condition (0.5 vs. 0.5, 0.525 vs. 0.475, and 0.55 vs. 0.45), and an extreme condition (0.1 vs. 0.1, 0.125 vs. 0.075, and 0.15 vs. 0.05), and (6) the relationship between the two population proportions and sample sizes, with the most extreme proportion assigned to the lowest sample size and vice versa (direct vs. inverse relationships).

To simulate the sample sizes, N, of k 2 × 2 tables in a meta-analysis, some properties of the sample size distribution in 30 real meta-analyses in the field of behavioural and health sciences were assessed. In particular, the Pearson skewness index of the distribution was computed throughout all the meta-analyses, obtaining a value of +1.464. In accordance with this value, three vectors of ten N,s each were selected: [24, 24, 32, 32, 36, 36, 40, 40, 168, 168], [64, 64, 72, 72, 76, 76, 80, 80, 208, 208], and [124, 124, 132, 132, 136, 136, 140, 140, 268, 268], all with the skewness= + 1.464, and averaging 60, 100, and 160, respectively. These were the sample size distributions for meta-analyses with 10 studies. To obtain meta-analyses of 20 and 40 studies, each N vector was repeated 2 and 4 times, respectively.

For each of the 198 conditions defined, 10 000 replications were run using the Monte Carlo simulation. With such a high number of replies, a conservative estimate of the maximum sampling error was ±0.0098, assuming $\pi_{\mathrm{T}} = \pi_{\mathrm{C}} = 0.5$ and a 95% confidence level. The five procedures to test the significance of the average risk difference were applied to the 10 000 replications of each condition. The criterion for the acceptance vs. rejection of the null hypothesis ($\mathrm{H}_0 : \delta = 0$) was adjusted to a nominal two-sided significance level of $\alpha = 0.05$. In conditions where $\delta = 0$, the proportion of rejections of the null hypothesis in the 10 000 replications was the estimated Type I error rate. In conditions where $\delta \neq 0$, the number of rejections of the null hypothesis was the estimated power.

In order to assess the adjustment of our empirical power values to the large sample theory, the asymptotical power was also derived for each of the four procedures: χ^2_{CW}, χ^2_{C}, χ^2_{MH}, and χ^2_{Y} tests. If z_x is the square root of any χ^2_x test (see Table 1), then the large sample distribution of the z_x statistics applied in each of the individual conditions is given by $z_x \sim N(\zeta_x, 1)$, where ζ_x is the square root of the χ_x^2 test as computed from the π_T, π_C, n_T, and n_C values. Thus the asymptotical power of the χ_x^2 test, P_x, for each of the defined conditions is given by $P_x = 1 - \phi(|z_{0.025}| - \zeta_x) + \phi(z_{0.025} - \zeta_x)$, where $\phi(x)$ is the standard normal cumulative distribution function, and $z_{0.025}$ is the 100 (0.025) per cent critical value of the standard normal distribution (assuming the two-sided $\alpha = 0.05$).

It was not possible with respect to the χ_U^2 test to derive an asymptotic power value for each condition, as had been the case with the other four procedures. In fact, when applied to the population δ values, all with a common value in the same meta-analysis, the equation for estimating the variance of the sampling distribution of the unweighted common risk difference, $S_{\bar{d}_u}^2$, gave always 0.

4. Results and discussion

Table 2 presents the empirical Type I error rates as a function of the position in the range of population proportions, π_T and π_C, the number of studies, k, the average sample size, $\overline{N}$, and the ratio between n_{Ti} and n_{Ci}. On average, the χ_Y^2 and χ_C^2 tests showed an adequate adjustment to the nominal α level ($\overline{\alpha} = 0.0500$ and 0.0512, respectively), very closely followed by χ_{MH}^2, with a more conservative empirical $\overline{\alpha}$ level of 0.0422. However, χ_{CW}^2 presented an unacceptably high Type I error rate of 0.2050 and χ_U^2 showed an empirical $\overline{\alpha} = 0.0679$ slightly higher than the nominal α.

The χ_C^2, χ_{MH}^2, and χ_Y^2 tests were scarcely affected by the manipulated factors: number of studies, average sample sizes, the ratio betwen n_{Ti} and n_{Ci}, and the position in the range of the population proportions. A good adjustment to the nominal α level was held by the χ_C^2 and χ_{MH}^2 tests throughout each of the conditions, while χ_{MH}^2 showed a slightly lower Type I error rate than $\alpha = 0.05$. The χ_U^2 test was only affected by the number of studies, decreasing its empirical rate as the number of studies increased. In contrast, the χ_{CW}^2 test was systematically affected by all the manipulated factors. Thus, the χ_{CW}^2 test dramatically increased its Type I error rate as the discrepancy between n_{Ti} and n_{Ci} grew, reaching an empirical $\overline{\alpha} = 0.4083$ when $n_{Ti}/n_{Ci} = 4$. Moreover, when the population proportions were in an extreme position ($\pi_T = \pi_C = 0.10$), the average Type I error rate was 0.3304, larger than that in a central position ($\pi_T = \pi_C = 0.50$), with an average value of $\overline{\alpha} = 0.0796$. Furthermore, in contrast to the large sample theory, χ_{CW}^2 increased its Type I error rate as the number of studies increased. The average sample size also affected the Type I error rate in χ_{CW}^2, although in this case its empirical rate decreased.

As anticipated, the problems with the χ_{CW}^2 test derived from the dependence between the individual risk differences, d_i, and the estimated weights, w_i, because the sample proportions intervene in the computation of both d_i and w_i. In particular, the poor adjustment of the empirical rates is dramatically increased with heavily unbalanced sample sizes and extreme population proportions. For example, with $k = 10$ studies, average sample size $\overline{N} = 50$, $n_{Ti}/n_{Ci} = 4$, and $\pi_T = \pi_C = 0.10$, the empirical Type I error rate reached a value of 0.5152 (see Table 2). In these conditions, biased

Table 2
Type I error rates

		$\pi_T = \pi_C = 0.50;\ \delta = 0$														
		$k = 10$					$k = 20$					$k = 40$				
$\overline{N}$	n_T/n_C	χ^2_U	χ^2_{CW}	χ^2_C	χ^2_{MH}	χ^2_Y	χ^2_U	χ^2_{CW}	χ^2_C	χ^2_{MH}	χ^2_Y	χ^2_U	χ^2_{CW}	χ^2_C	χ^2_{MH}	χ^2_Y
60	1	0.0761	0.0626	0.0541	0.0399	0.0488	0.0618	0.0622	0.0496	0.0422	0.0485	0.0560	0.0624	0.0504	0.0434	0.0477
60	2	0.0801	0.0754	0.0522	0.0400	0.0493	0.0621	0.0740	0.0513	0.0424	0.0494	0.0579	0.0742	0.0496	0.0431	0.0488
60	4	0.0802	0.1834	0.0528	0.0389	0.0501	0.0636	0.2003	0.0541	0.0450	0.0529	0.0554	0.2040	0.0500	0.0440	0.0486
100	1	0.0773	0.0572	0.0533	0.0432	0.0505	0.0674	0.0584	0.0534	0.0476	0.0534	0.0565	0.0595	0.0519	0.0469	0.0500
100	2	0.0843	0.0651	0.0553	0.0437	0.0549	0.0638	0.0601	0.0497	0.0427	0.0473	0.0583	0.0683	0.0548	0.0481	0.0548
100	4	0.0775	0.0779	0.0487	0.0386	0.0470	0.0659	0.0796	0.0511	0.0429	0.0509	0.0556	0.0835	0.0488	0.0438	0.0481
160	1	0.0846	0.0563	0.0496	0.0438	0.0489	0.0643	0.0523	0.0476	0.0438	0.0476	0.0564	0.0540	0.0483	0.0451	0.0483
160	2	0.0881	0.0622	0.0548	0.0496	0.0548	0.0647	0.0607	0.0546	0.0506	0.0540	0.0592	0.0569	0.0509	0.0456	0.0509
160	4	0.0822	0.0659	0.0501	0.0441	0.0501	0.0634	0.0662	0.0516	0.0444	0.0511	0.0576	0.0656	0.0519	0.0471	0.0519

		$\pi_T = \pi_C = 0.10;\ \delta = 0$														
		$k = 10$					$k = 20$					$k = 40$				
$\overline{N}$	n_T/n_C	χ^2_U	χ^2_{CW}	χ^2_C	χ^2_{MH}	χ^2_Y	χ^2_U	χ^2_{CW}	χ^2_C	χ^2_{MH}	χ^2_Y	χ^2_U	χ^2_{CW}	χ^2_C	χ^2_{MH}	χ^2_Y
60	1	0.0788	0.0468	0.0511	0.0343	0.0485	0.0604	0.0472	0.0488	0.0366	0.0469	0.0528	0.0491	0.0487	0.0401	0.0470
60	2	0.0783	0.1656	0.0492	0.0335	0.0475	0.0640	0.2935	0.0498	0.0369	0.0478	0.0568	0.5089	0.0529	0.0433	0.0510
60	4	0.0872	0.6175	0.0521	0.0318	0.0486	0.0685	0.8815	0.0518	0.0384	0.0504	0.0599	0.9923	0.0540	0.0421	0.0520
100	1	0.0842	0.0559	0.0492	0.0384	0.0489	0.0657	0.0555	0.0513	0.0438	0.0503	0.0580	0.0572	0.0522	0.0458	0.0516
100	2	0.0829	0.1452	0.0478	0.0368	0.0471	0.0616	0.2132	0.0490	0.0395	0.0478	0.0585	0.3515	0.0547	0.0454	0.0532
100	4	0.0890	0.5152	0.0546	0.0401	0.0545	0.0645	0.7663	0.0523	0.0413	0.0514	0.0585	0.9430	0.0508	0.0435	0.0498
160	1	0.0859	0.0541	0.0509	0.0389	0.0502	0.0660	0.0552	0.0516	0.0443	0.0507	0.0580	0.0568	0.0518	0.0466	0.0510
160	2	0.0810	0.0981	0.0508	0.0390	0.0507	0.0611	0.1392	0.0480	0.0413	0.0471	0.0571	0.2048	0.0486	0.0429	0.0478
160	4	0.0863	0.3399	0.0531	0.0387	0.0517	0.0648	0.5151	0.0502	0.0411	0.0490	0.0565	0.7521	0.0507	0.0432	0.0497

Note. π_T and π_C=population proportions; k=number of studies; $\overline{N}$=average sample size; n_T/n_C=ratio between sample sizes; δ=population risk difference.

estimates of δ receive a disproportionate weight, leading to the erroneous rejection of the null hypothesis.

Conversely the χ^2_{CW} test, applying unconditional weights allows an adequate adjustment to the nominal α level in all of the conditions. Therefore, the χ^2_{C}, χ^2_{MH}, and χ^2_{Y} tests achieve an adequate adjustment. However, the χ^2_{MH} test exhibits a slightly lower empirical rate than the nominal 0.05, implying that the one-half continuity correction makes the test more conservative.

On the other hand, the χ^2_{U} test does not adjust the Type I error rate under any of the conditions, with empirical rates systematically higher than the nominal $\alpha = 0.05$. As anticipated from the large sample theory, the performance of the χ^2_{U} test improves as the number of studies and sample size increase.

Tables 3–6 present the empirical and asymptotical statistical power of the five procedures as a function of the manipulated conditions. The χ^2_{C} and χ^2_{Y} tests showed the highest empirical power, averaging 0.8479 and 0.8461, respectively. Because of its more conservative Type I error rates, χ^2_{MH} presented a slightly lower average power of 0.8337. On the other hand, the χ^2_{CW} and χ^2_{U} tests achieved the lowest empirical power, with average values of 0.8002 and 0.8004, respectively. Nevertheless, the averages give a very simplistic picture of the results, because the factors manipulated in our simulation study greatly affected the power rates.

As anticipated, the power of all of the tests increased as the number of studies (k), average sample size ($\overline{N}$), and the population risk difference (δ) increased. Furthermore, the more extreme the population proportions, the higher the power; and as the ratio between sample sizes increased, from equal to unbalanced sample sizes, the power decreased.

Although the χ^2_{CW} test achieved the lowest average empirical power, paradoxically it achieved the largest power under many of the conditions. In particular, when the population proportions were in a central position (π_{T} and π_{C} over 0.50), the χ^2_{CW} test was systematically the most powerful of the procedures (see Tables 3 and 5). In contrast, when the population proportions were in an extreme position (π_{T} and π_{C} over 0.10) the power was not always the highest. In particular, when each of the individual 2×2 tables presented the most extreme population proportion assigned to the largest sample size, the χ^2_{CW} test suffered a drastical decrease in the power (see Tables 4 and 6), becoming the least powerful of the procedures. However, when the most extreme population proportion was linked to the lowest sample size, the power of the χ^2_{CW} test surpassed that of the other procedures.

The dependence between the sample risk differences and the estimated weights explains the irregular performance of the χ^2_{CW} test. The most problematic conditions are those of a very low statistical power, which occur when the population proportions are in an extreme position and the most extreme proportion is associated with the largest sample size. In these conditions, sample risk differences of opposite value to that of the population effect size received a disproportionate weight, leading to the erroneous acceptance of the null hypothesis.

Unlike the χ^2_{CW} test, the unweighted test, χ^2_{U}, showed a trend similar to that of the χ^2_{C}, χ^2_{MH}, and χ^2_{Y} tests in all the manipulated conditions, although exhibiting a systematically lower power.

Table 3
Empirical and asymptotical (in parentheses) power rates

		$\pi_T = 0.525,\ \pi_C = 0.475;\ \delta = 0.05$														
		$k = 10$					$k = 20$					$k = 40$				
$\overline{N}$	n_T/n_C	χ^2_U	χ^2_{CW}	χ^2_C	χ^2_{MH}	χ^2_Y	χ^2_U	χ^2_{CW}	χ^2_C	χ^2_{MH}	χ^2_Y	χ^2_U	χ^2_{CW}	χ^2_C	χ^2_{MH}	χ^2_Y
60	1	0.2169	0.2706	0.2497	0.2093	0.2347	0.3009	0.4531	0.4219	0.3910	0.4148	0.4969	0.7132	0.6845	0.6588	0.6748
			(0.2323)	(0.2318)	(0.2052)	(0.2287)		(0.4108)	(0.4100)	(0.3823)	(0.4043)		(0.6889)	(0.6878)	(0.6658)	(0.6804)
60	2	0.1941	0.2619	0.2216	0.1875	0.2154	0.2805	0.4246	0.3730	0.3457	0.3650	0.4603	0.6822	0.6419	0.6167	0.6402
			(0.2117)	(0.2113)	(0.1851)	(0.2085)		(0.3728)	(0.3721)	(0.3443)	(0.3669)		(0.6377)	(0.6367)	(0.6131)	(0.6294)
60	4	0.1645	0.2989	0.1718	0.1417	0.1656	0.2160	0.4219	0.2840	0.2559	0.2799	0.3524	0.6079	0.5013	0.4782	0.4963
			(0.1654)	(0.1652)	(0.1403)	(0.1633)		(0.2839)	(0.2835)	(0.2561)	(0.2796)		(0.5009)	(0.5002)	(0.4735)	(0.4936)
100	1	0.3410	0.3710	0.3602	0.3261	0.3489	0.5484	0.6366	0.6194	0.5995	0.6194	0.8180	0.8965	0.8871	0.8777	0.8840
			(0.3533)	(0.3526)	(0.3267)	(0.3497)		(0.6099)	(0.6088)	(0.5872)	(0.6045)		(0.8861)	(0.8854)	(0.8759)	(0.8823)
100	2	0.3177	0.3475	0.3248	0.2939	0.3243	0.4926	0.5831	0.5611	0.5377	0.5541	0.7721	0.8590	0.8459	0.8353	0.8459
			(0.3209)	(0.3203)	(0.2944)	(0.3177)		(0.5608)	(0.5599)	(0.5370)	(0.5557)		(0.8481)	(0.8473)	(0.8356)	(0.8437)
100	4	0.2529	0.2917	0.2456	0.2186	0.2409	0.3851	0.4815	0.4307	0.4065	0.4296	0.6414	0.7528	0.7206	0.7043	0.7192
			(0.2446)	(0.2443)	(0.2187)	(0.2423)		(0.4330)	(0.4325)	(0.4072)	(0.4289)		(0.7167)	(0.7160)	(0.6981)	(0.7117)
160	1	0.5119	0.5320	0.5134	0.4925	0.5119	0.7727	0.8161	0.8082	0.7963	0.8082	0.9682	0.9772	0.9758	0.9745	0.9758
			(0.5170)	(0.5160)	(0.4936)	(0.5135)		(0.8084)	(0.8074)	(0.7951)	(0.8050)		(0.9796)	(0.9793)	(0.9774)	(0.9787)
160	2	0.4712	0.4858	0.4643	0.4510	0.4641	0.7245	0.7743	0.7613	0.7527	0.7584	0.9484	0.9668	0.9638	0.9613	0.9638
			(0.4704)	(0.4696)	(0.4463)	(0.4672)		(0.7601)	(0.7592)	(0.7447)	(0.7566)		(0.9649)	(0.9646)	(0.9615)	(0.9637)
160	4	0.3770	0.3929	0.3635	0.3394	0.3621	0.5955	0.6461	0.6198	0.5991	0.6196	0.8581	0.8982	0.8900	0.8827	0.8900
			(0.3604)	(0.3599)	(0.3351)	(0.3580)		(0.6201)	(0.6194)	(0.5997)	(0.6167)		(0.8933)	(0.8928)	(0.8850)	(0.8909)

Note. π_T and π_C=population proportions; k=number of studies; $\overline{N}$=average sample size; n_T/n_C=ratio between sample sizes; δ=population risk difference.

Table 4
Empirical and asymptotical (in parentheses) power rates

		$\pi_T = 0.125,\ \pi_C = 0.075;\ \delta = 0.05$														
		$k = 10$					$k = 20$					$k = 40$				
$\overline{N}$	n_T/n_C	χ^2_U	χ^2_{CW}	χ^2_C	χ^2_{MH}	χ^2_Y	χ^2_U	χ^2_{CW}	χ^2_C	χ^2_{MH}	χ^2_Y	χ^2_U	χ^2_{CW}	χ^2_C	χ^2_{MH}	χ^2_Y
60	1	0.4131	0.4921	0.5412	0.4784	0.5367	0.6415	0.7740	0.8281	0.7972	0.8247	0.8968	0.9675	0.9837	0.9801	0.9835
			(0.5353)	(0.5324)	(0.4718)	(0.5256)		(0.8256)	(0.8230)	(0.7902)	(0.8166)		(0.9837)	(0.9831)	(0.9783)	(0.9816)
60	2*d*[a]	0.4151	0.7917	0.4667	0.4013	0.4631	0.6270	0.9763	0.7767	0.7380	0.7726	0.8757	0.9996	0.9731	0.9651	0.9716
			(0.5190)	(0.4593)	(0.3990)	(0.4531)		(0.8103)	(0.7476)	(0.7079)	(0.7405)		(0.9801)	(0.9604)	(0.9510)	(0.9576)
60	2*i*[b]	0.3423	0.1524	0.5210	0.4559	0.5151	0.5647	0.2126	0.7988	0.7628	0.7936	0.8563	0.3326	0.9758	0.9704	0.9747
			(0.4614)	(0.5163)	(0.4504)	(0.5096)		(0.7501)	(0.8077)	(0.7705)	(0.8011)		(0.9613)	(0.9794)	(0.9734)	(0.9777)
60	4*d*	0.3809	0.9561	0.3338	0.2575	0.3241	0.5356	0.9997	0.6050	0.5469	0.5977	0.7771	1.0	0.9013	0.8823	0.8992
			(0.4213)	(0.3359)	(0.2761)	(0.3312)		(0.7022)	(0.5839)	(0.5324)	(0.5767)		(0.9409)	(0.8668)	(0.8428)	(0.8611)
60	4*i*	0.2187	0.1493	0.4214	0.3473	0.4116	0.3936	0.2398	0.6778	0.6262	0.6727	0.6967	0.4311	0.9162	0.9006	0.9142
			(0.3370)	(0.4195)	(0.3449)	(0.4138)		(0.5855)	(0.7000)	(0.6462)	(0.6926)		(0.8681)	(0.9398)	(0.9241)	(0.9362)
100	1	0.6807	0.7360	0.7550	0.7153	0.7519	0.9202	0.9542	0.9650	0.9577	0.9643	0.9976	0.9991	0.9993	0.9993	0.9993
			(0.7532)	(0.7502)	(0.7113)	(0.7460)		(0.9624)	(0.9614)	(0.9529)	(0.9598)		(0.9996)	(0.9995)	(0.9994)	(0.9995)
100	2*d*	0.6651	0.8864	0.6962	0.6529	0.6921	0.9039	0.9935	0.9427	0.9306	0.9417	0.9946	0.9999	0.9986	0.9982	0.9986
			(0.7371)	(0.6703)	(0.6260)	(0.6659)		(0.9563)	(0.9243)	(0.9102)	(0.9219)		(0.9994)	(0.9978)	(0.9971)	(0.9976)
100	2*i*	0.6120	0.4290	0.7302	0.6871	0.7283	0.8811	0.6432	0.9497	0.9411	0.9486	0.9934	0.8779	0.9984	0.9983	0.9984
			(0.6728)	(0.7342)	(0.6905)	(0.7300)		(0.9257)	(0.9551)	(0.9448)	(0.9534)		(0.9978)	(0.9993)	(0.9991)	(0.9993)
100	4*d*	0.5734	0.9563	0.5207	0.4613	0.5187	0.8069	0.9991	0.8343	0.8055	0.8306	0.9758	1.0	0.9885	0.9860	0.9881
			(0.6231)	(0.5090)	(0.4560)	(0.5050)		(0.8954)	(0.8006)	(0.7711)	(0.7966)		(0.9952)	(0.9775)	(0.9728)	(0.9764)
100	4*i*	0.4539	0.2004	0.6022	0.5488	0.5985	0.7355	0.1741	0.8666	0.8451	0.8648	0.9649	0.1721	0.9901	0.9883	0.9899
			(0.5106)	(0.6209)	(0.5618)	(0.6166)		(0.8021)	(0.8938)	(0.8711)	(0.8909)		(0.9779)	(0.9950)	(0.9935)	(0.9947)
160	1	0.8815	0.9146	0.9225	0.9075	0.9213	0.9933	0.9973	0.9982	0.9978	0.9981	1.0	1.0	1.0	1.0	1.0
			(0.9170)	(0.9152)	(0.8997)	(0.9135)		(0.9972)	(0.9971)	(0.9963)	(0.9969)		(1.0)	(1.0)	(1.0)	(1.0)
160	2*d*	0.8573	0.9528	0.8785	0.8573	0.8763	0.9897	0.9995	0.9943	0.9930	0.9943	1.0	1.0	1.0	1.0	1.0
			(0.9056)	(0.8576)	(0.8352)	(0.8555)		(0.9962)	(0.9900)	(0.9878)	(0.9896)		(1.0)	(1.0)	(1.0)	(1.0)
160	2*i*	0.8279	0.7327	0.8870	0.8685	0.8858	0.9845	0.9425	0.9950	0.9931	0.9947	0.9999	0.9983	1.0	1.0	1.0
			(0.8596)	(0.9038)	(0.8852)	(0.9020)		(0.9903)	(0.9960)	(0.9950)	(0.9959)		(1.0)	(1.0)	(1.0)	(1.0)
160	4*d*	0.7706	0.9575	0.7435	0.7026	0.7406	0.9576	0.9992	0.9664	0.9603	0.9659	0.9993	1.0	0.9998	0.9998	0.9998
			(0.8202)	(0.7081)	(0.6708)	(0.7053)		(0.9825)	(0.9437)	(0.9340)	(0.9424)		(0.9999)	(0.9989)	(0.9986)	(0.9988)
160	4*i*	0.6892	0.3777	0.7878	0.7561	0.7860	0.9408	0.5068	0.9716	0.9667	0.9711	0.9992	0.6608	0.9999	0.9999	0.9999
			(0.7097)	(0.8182)	(0.7846)	(0.8158)		(0.9445)	(0.9820)	(0.9775)	(0.9814)		(0.9989)	(0.9999)	(0.9999)	(0.9999)

Note. π_T and π_C=population proportions; k=number of studies; $\overline{N}$=average sample size; n_T/n_C=ratio between sample sizes; δ=population risk difference.

[a] *d* denotes a direct relationship where the most extreme proportion is associated with the lowest sample size.

[b] *i* denotes an inverse relationship where the most extreme proportion is associated with the largest sample size.

Table 5
Empirical and asymptotical (in parentheses) power rates

		$\pi_T = 0.550,\ \pi_C = 0.450;\ \delta = 0.10$														
		$k = 10$					$k = 20$					$k = 40$				
$\overline{N}$	n_T/n_C	χ^2_U	χ^2_{CW}	χ^2_C	χ^2_{MH}	χ^2_Y	χ^2_U	χ^2_{CW}	χ^2_C	χ^2_{MH}	χ^2_Y	χ^2_U	χ^2_{CW}	χ^2_C	χ^2_{MH}	χ^2_Y
60	1	0.5292	0.7241	0.7054	0.6566	0.6878	0.7877	0.9450	0.9366	0.9267	0.9354	0.9707	0.9980	0.9976	0.9971	0.9976
			(0.6921)	(0.6878)	(0.6509)	(0.6804)		(0.9359)	(0.9337)	(0.9218)	(0.9299)		(0.9985)	(0.9984)	(0.9979)	(0.9981)
60	2	0.4891	0.6932	0.6537	0.6069	0.6466	0.7407	0.9215	0.9108	0.8947	0.9071	0.9571	0.9965	0.9970	0.9964	0.9968
			(0.6410)	(0.6371)	(0.5969)	(0.6298)		(0.9070)	(0.9045)	(0.8887)	(0.8998)		(0.9963)	(0.9961)	(0.9951)	(0.9956)
60	4	0.3861	0.5918	0.5069	0.4594	0.5007	0.6118	0.8251	0.7979	0.7694	0.7947	0.8778	0.9654	0.9756	0.9716	0.9750
			(0.5038)	(0.5013)	(0.4544)	(0.4947)		(0.7954)	(0.7928)	(0.7646)	(0.7861)		(0.9761)	(0.9754)	(0.9701)	(0.9734)
100	1	0.8215	0.9014	0.8970	0.8799	0.8904	0.9802	0.9955	0.9950	0.9940	0.9950	0.9998	1.0	1.0	1.0	1.0
			(0.8884)	(0.8854)	(0.8694)	(0.8823)		(0.9944)	(0.9940)	(0.9928)	(0.9936)		(1.0)	(1.0)	(1.0)	(1.0)
100	2	0.7757	0.8622	0.8512	0.8295	0.8505	0.9650	0.9903	0.9889	0.9868	0.9880	0.9997	1.0	1.0	1.0	1.0
			(0.8507)	(0.8476)	(0.8275)	(0.8440)		(0.9888)	(0.9882)	(0.9859)	(0.9875)		(1.0)	(1.0)	(1.0)	(1.0)
100	4	0.6491	0.7487	0.7182	0.6881	0.7162	0.8964	0.9547	0.9471	0.9400	0.9459	0.9947	0.9991	0.9989	0.9989	0.9989
			(0.7199)	(0.7172)	(0.6854)	(0.7128)		(0.9491)	(0.9478)	(0.9395)	(0.9459)		(0.9991)	(0.9991)	(0.9988)	(0.9990)
160	1	0.9582	0.9811	0.9787	0.9760	0.9785	0.9994	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
			(0.9803)	(0.9793)	(0.9760)	(0.9787)		(0.9999)	(0.9999)	(0.9999)	(0.9999)		(1.0)	(1.0)	(1.0)	(1.0)
160	2	0.9356	0.9687	0.9658	0.9617	0.9657	0.9988	0.9994	0.9994	0.9994	0.9994	1.0	1.0	1.0	1.0	1.0
			(0.9660)	(0.9647)	(0.9594)	(0.9638)		(0.9997)	(0.9996)	(0.9995)	(0.9996)		(1.0)	(1.0)	(1.0)	(1.0)
160	4	0.8524	0.9080	0.9006	0.8881	0.8988	0.9901	0.9954	0.9949	0.9945	0.9949	1.0	1.0	1.0	1.0	1.0
			(0.8955)	(0.8936)	(0.8797)	(0.8917)		(0.9952)	(0.9950)	(0.9941)	(0.9948)		(1.0)	(1.0)	(1.0)	(1.0)

Note. π_T and π_C=population proportions; k=number of studies; $\overline{N}$=average sample size; n_T/n_C=ratio between sample sizes; δ=population risk difference.

Table 6
Empirical and asymptotical (in parentheses) power rates

		$\pi_T = 0.150,\ \pi_C = 0.050;\ \delta = 0.10$														
		$k = 10$					$k = 20$					$k = 40$				
$\overline{N}$	n_T/n_C	χ^2_U	χ^2_{CW}	χ^2_C	χ^2_{MH}	χ^2_Y	χ^2_U	χ^2_{CW}	χ^2_C	χ^2_{MH}	χ^2_Y	χ^2_U	χ^2_{CW}	χ^2_C	χ^2_{MH}	χ^2_Y
60	1	0.9056	0.9728	0.9873	0.9821	0.9869	0.9960	0.9998	0.9999	0.9999	0.9999	1.0	1.0	1.0	1.0	1.0
			(0.9854)	(0.9831)	(0.9746)	(0.9816)		(1.0)	(0.9999)	(0.9999)	(0.9999)		(1.0)	(1.0)	(1.0)	(1.0)
60	2*d*[a]	0.8933	0.9974	0.9758	0.9649	0.9751	0.9931	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
			(0.9887)	(0.9492)	(0.9296)	(0.9459)		(1.0)	(0.9991)	(0.9986)	(0.9990)		(1.0)	(1.0)	(1.0)	(1.0)
60	2*i*[b]	0.8447	0.7100	0.9809	0.9729	0.9807	0.9897	0.9278	0.9999	0.9999	0.9999	1.0	0.9976	1.0	1.0	1.0
			(0.9532)	(0.9867)	(0.9787)	(0.9855)		(0.9993)	(1.0)	(0.9999)	(1.0)		(1.0)	(1.0)	(1.0)	(1.0)
60	4*d*	0.8271	0.9998	0.9108	0.8701	0.9051	0.9703	1.0	0.9977	0.9964	0.9976	0.9999	1.0	1.0	1.0	1.0
			(0.9731)	(0.8298)	(0.7818)	(0.8235)		(1.0)	(0.9846)	(0.9783)	(0.9832)		(1.0)	(0.9999)	(0.9999)	(0.9999)
60	4*i*	0.6628	0.1406	0.9390	0.9123	0.9373	0.9393	0.1553	0.9982	0.9971	0.9980	0.9995	0.1973	1.0	1.0	1.0
			(0.8350)	(0.9700)	(0.9505)	(0.9677)		(0.9857)	(0.9997)	(0.9995)	(0.9997)		(1.0)	(1.0)	(1.0)	(1.0)
100	1	0.9959	0.9991	0.9996	0.9996	0.9996	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
			(0.9996)	(0.9995)	(0.9993)	(0.9995)		(1.0)	(1.0)	(1.0)	(1.0)		(1.0)	(1.0)	(1.0)	(1.0)
100	2*d*	0.9946	0.9998	0.9994	0.9988	0.9994	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
			(0.9998)	(0.9964)	(0.9948)	(0.9962)		(1.0)	(1.0)	(1.0)	(1.0)		(1.0)	(1.0)	(1.0)	(1.0)
100	2*i*	0.9891	0.9613	0.9994	0.9990	0.9993	1.0	0.9990	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
			(0.9969)	(0.9997)	(0.9995)	(0.9997)		(1.0)	(1.0)	(1.0)	(1.0)		(1.0)	(1.0)	(1.0)	(1.0)
100	4*d*	0.9802	0.9997	0.9912	0.9874	0.9907	0.9998	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
			(0.9989)	(0.9642)	(0.9521)	(0.9627)		(1.0)	(0.9996)	(0.9994)	(0.9996)		(1.0)	(1.0)	(1.0)	(1.0)
100	4*i*	0.9483	0.4860	0.9933	0.9901	0.9932	0.9993	0.6313	1.0	1.0	1.0	1.0	0.8332	1.0	1.0	1.0
			(0.9662)	(0.9986)	(0.9976)	(0.9985)		(0.9997)	(1.0)	(1.0)	(1.0)		(1.0)	(1.0)	(1.0)	(1.0)
160	1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
			(1.0)	(1.0)	(1.0)	(1.0)		(1.0)	(1.0)	(1.0)	(1.0)		(1.0)	(1.0)	(1.0)	(1.0)
160	2*d*	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
			(1.0)	(1.0)	(1.0)	(1.0)		(1.0)	(1.0)	(1.0)	(1.0)		(1.0)	(1.0)	(1.0)	(1.0)
160	2*i*	0.9999	0.9990	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
			(1.0)	(1.0)	(1.0)	(1.0)		(1.0)	(1.0)	(1.0)	(1.0)		(1.0)	(1.0)	(1.0)	(1.0)
160	4*d*	0.9992	1.0	0.9998	0.9997	0.9998	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
			(1.0)	(0.9974)	(0.9964)	(0.9973)		(1.0)	(1.0)	(1.0)	(1.0)		(1.0)	(1.0)	(1.0)	(1.0)
160	4*i*	0.9977	0.8958	1.0	1.0	1.0	1.0	0.9836	1.0	1.0	1.0	1.0	0.9998	1.0	1.0	1.0
			(0.9977)	(1.0)	(1.0)	(1.0)		(1.0)	(1.0)	(1.0)	(1.0)		(1.0)	(1.0)	(1.0)	(1.0)

Note. π_T and π_C=population proportions; k=number of studies; $\overline{N}$=average sample size; n_T/n_C=ratio between sample sizes; δ=population risk difference.

[a] *d* denotes a direct relationship where the most extreme proportion is associated with the lowest sample size.

[b] *i* denotes an inverse relationship where the most extreme proportion is associated with the largest sample size.

Tables 3–6 also included the asymptotical statistical power of the χ^2_{C}, χ^2_{MH}, χ^2_{Y}, and χ^2_{CW} tests. In general, the asymptotical power of χ^2_{CW}, which was computed using the optimal weights, $W_i = 1/\sigma^2_{d_i}$, achieved the largest values in comparison with the asymptotical power of the remaining procedures. The exception occurred when the most extreme population proportion was assigned to the largest sample size, achieving the smallest asymptotical values.

Comparing the empirical power values of the χ^2_{C}, χ^2_{MH}, and χ^2_{Y} procedures with their corresponding asymptotical powers, the three procedures showed an adequate adjustment. Conversely, the conditional χ^2_{CW} test showed the largest discrepancies between the asymptotical and empirical values. The discrepancies were more pronounced with unbalanced sample sizes and the most extreme population proportion occurring in the largest sample size.

5. Conclusions

In this paper we compared the Type I error and statistical power rates of five statistical procedures for testing the significance of a common risk difference in conditions usually found in meta-analytic research in health and behavioural sciences. We assumed a set of k 2×2 tables comparing treated vs control groups, homogeneous control proportions through the studies, and a fixed-effects model. Therefore, the results of our simulation study must be limited to the particular conditions where the use of a risk difference as the effect size index is advisable.

As anticipated, the unweighted χ^2_{U} test showed the lowest empirical power values under most of the conditions. In contrast with some meta-analytic approaches (e.g., Glass et al., 1981), this finding does not support the use of unweighted procedures in meta-analysis.

Our results demonstrate an anomalous performance by the conditional weighted χ^2_{CW} test, and for this reason we advise against its application in meta-analyses with unbalanced sample sizes, especially when the most extreme proportion occurs in the largest sample size. This adverse effect increases when the number of studies and the sample sizes are low, as is frequently the case in meta-analyses of health sciences. Therefore, although recently recommended by Shadish and Haddock (1994), we consider the χ^2_{CW} test as being far from the most appropriate procedure for testing the significance of a common risk difference.

The unconditional weighted procedures achieved the best performance, adequately adjusting both the Type I error and empirical power rates. Throughout the manipulated conditions, the χ^2_{C}, χ^2_{MH}, and χ^2_{Y} tests presented very similar power values, confirming the commentary of Laird and Mosteller (1990) concerning the equivalence of the three procedures. Nevertheless, χ^2_{MH} was slightly more conservative than the other two procedures, due to the inclusion of the one-half continuity correction. Consequently, although procedures of Mantel-Haenszel and Yusuf et al. were proposed as an improvement on Cochran's original test, our results do not confirm such higher performances under the manipulated conditions. In practice, because the detected differences among the χ^2_{C}, χ^2_{MH}, and χ^2_{Y} tests are negligible, we consider that the three tests are interchangeable. However, it must be noted that the statistical power of these procedures was not always suitable, with empirical values even lower than 0.50 in some conditions. If the criterion of 0.80 proposed by Cohen (1988) in social and behavioural sciences is taken into account as the minimum advisable power, caution is necessary under many of the conditions, especially with small sample sizes, a low number of studies, and small differences between the two population proportions.

Finally, it is important to note that the differences observed in our simulation study are limited to manipulated conditions. Real meta-analyses include more heterogeneous studies than simulated ones, for example with variable sample size ratios and different relationships between sample sizes and population proportions of treated and control groups. These, and other, factors can also affect the performance of the procedures. As a consequence, we believe that new research efforts should be devoted to assessing the performance of meta-analytic techniques in integrating 2 × 2 tables, such as homogeneity tests, the search for moderator variables and comparing different effect size indexes.

Consequently, new simulation studies are needed in order to probe deeper into the performance of the procedures advocated from different meta-analytic approaches.

References

Ahn, C., 1997. An evaluation of simple methods for the estimation of a common odds ratio in clusters with variable size. Comput. Statist. Data Anal. 24, 47–61.

Aptech Systems, 1992. The GAUSS system (Version 3.0). Maple Valley, Author.

Cochran, W.G., 1954. Some methods for strengthening the common χ^2 tests. Biometrics 10, 417–451.

Cohen, J., 1988. Statistical Power Analysis For the Behavioral Sciences, 2nd Edition. Erlbaum, Hillsdale.

Cooper, H., Hedges, L.V. (Eds.), 1994. The Handbook of Research Synthesis. Sage, New York.

Emerson, J.D., Hoaglin, D.C., Mosteller, F., 1993. A modified random-effect procedure for combining risk difference in sets of 2×2 tables from clinical trials. J. Ital. Statist. Soc. 3, 269–290.

Emerson, J.D., Hoaglin, D.C., Mosteller, F., 1996. Simple robust procedures for combining risk differences in sets of 2×2 tables. Statist. Med. 15, 1465–1488.

Fleiss, J.L., 1981. Statistical Methods For Rates and Proportions, 2nd Edition. Wiley, New York.

Fleiss, J.L., 1994. Measures of effect size for categorical data In: Cooper, H., Hedges, L.V., (Eds.), The Handbook of Research Synthesis. Sage, New York, 245–260.

Glass, G.V., McGaw, B., Smith, M.L., 1981. Meta-Analysis in Social Research. Sage, Beverly Hills.

Greenland, S., 1987. Quantitative methods in the review of epidemiologic literature. Epidemiol. Rev. 9, 1–30.

Haddock, C.K., Rindskopf, D., Shadish, W.R., 1998. Using odds ratios as effect sizes for meta-analysis of dichotomous data: a primer on methods and issues. Psychol. Meth. 3, 339–353.

Hasselblad, V., Mosteller, F., Littenberg, B., Chalmers, T.C., Hunink, M.G.M., Turner, J.A., Morton, S.C., Diehr, P., Wong, J.B., Powe, N.R., 1995. A survey of current problems in meta-analysis: discussion from the agency for Health Care Policy and research inter-PO T work group on literature review/meta-analysis. Med. Care 33, 202–220.

Kuss, O., Koch, A., 1996. Meta-analysis macros for SAS. Comput. Statist. Data Anal. 22, 325–333.

Laird, N.M., Mosteller, F., 1990. Some statistical methods for combining experimental results. Internat. J. Technol. Assessment in Health Care 6, 5–30.

Mantel, N., Haenszel, W., 1959. Statistical aspects of the analysis of data from retrospective studies of disease. J. Nat. Cancer Inst. 22, 719–748.

Shadish, W.R., Haddock, K., 1994. Combining estimates of effect size In: Cooper, H., Hedges, L.V., (Eds.), The Handbook of Research Synthesis. Sage, New York, pp. 261–281.

Tukey, J.W., 1977. Exploratory Data Analysis. Addison-Wesley, Reading, MA.

Yusuf, S., Peto, R., Lewis, J., Collins, R., Sleight, P., 1985. Beta blockade during and after myocardial infarction: an overview of the randomized trials. Prog. Cardiovas. Dis. 27, 335–371.