American Journal of Epidemiology

Vol. 150, No. 2
Printed in U.S.A.

Evaluation of Old and New Tests of Heterogeneity in Epidemiologic Meta-Analysis

Bahi Takkouche,[1] Carmen Cadarso-Suárez,[2] and Donna Spiegelman[3]

The identification of heterogeneity in effects between studies is a key issue in meta-analyses of observational studies, since it is critical for determining whether it is appropriate to pool the individual results into one summary measure. The result of a hypothesis test is often used as the decision criterion. In this paper, the authors use a large simulation study patterned from the key features of five published epidemiologic meta-analyses to investigate the type I error and statistical power of five previously proposed asymptotic homogeneity tests, a parametric bootstrap version of each of the tests, and τ^2-bootstrap, a test proposed by the authors. The results show that the asymptotic DerSimonian and Laird Q statistic and the bootstrap versions of the other tests give the correct type I error under the null hypothesis but that all of the tests considered have low statistical power, especially when the number of studies included in the meta-analysis is small (<20). From the point of view of validity, power, and computational ease, the Q statistic is clearly the best choice. The authors found that the performance of all of the tests considered did not depend appreciably upon the value of the pooled odds ratio, both for size and for power. Because tests for heterogeneity will often be underpowered, random effects models can be used routinely, and heterogeneity can be quantified by means of R_I, the proportion of the total variance of the pooled effect measure due to between-study variance, and CV_B, the between-study coefficient of variation. *Am J Epidemiol* 1999;150:206–15.

epidemiologic methods; heterogeneity; meta-analysis; models, statistical; significance tests

Heterogeneity between studies occurs when differences in study results for the same exposure-disease association cannot be fully accounted for by sampling variation. In meta-analysis, identifying and properly accounting for heterogeneity between studies is a critical step in the meta-analytic process. It involves a decision about whether one should pool individual results into one summary measure or present separate results for subgroups only. Heterogeneity in epidemiology originates from differences in study design, disease definition, and exposure assessment, inclusion of different covariates, and demographic variability in study populations (1, 2). Statistical criteria have been used to decide whether heterogeneity exists in a particular meta-analysis and, thus, whether it is meaningful to pool the results from individual studies into one common estimator. Hypothesis testing is the most popular of these methods, but a graphic approach has also been proposed (3). Others recommend that meta-analyses always use random effects models and thus include an estimate of the between-study variability regardless of the results of any test for heterogeneity (4). The null hypothesis for these tests assumes that there is no heterogeneity between the individual study outcomes, i.e., that the results are a random sample from one universe of results.

The limitations of hypothesis testing are well known to epidemiologists, and testing in a meta-analysis does not pose an exception to these drawbacks (5). The main problem is that the results of the test—and, by extension, the decision process following—are a function of both the magnitude of the effect that is tested *and* the sample size. In epidemiologic meta-analysis, the sample size within individual studies is typically large but the number of studies is relatively small (generally less than 30). Tests use the individual study as the statistical unit of observation. Thus, they are often based on a small sample size and may have low power against the alternative hypothesis of heterogeneity as a result.

In this paper, we compare the type I error rates and statistical power of several tests of heterogeneity that can be used in meta-analysis. We review and describe

Received for publication August 5, 1997, and accepted for publication November 20, 1998.

Abbreviations: df, degrees of freedom; LRT, likelihood ratio test; WLS, weighted least squares.

[1] Department of Preventive Medicine, School of Medicine, University of Santiago de Compostela, Santiago de Compostela, Spain.
[2] Department of Biostatistics, School of Medicine, University of Santiago de Compostela, Santiago de Compostela, Spain.
[3] Departments of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, MA.
Reprint requests to Dr. Bahi Takkouche, Facultad de Medicina, Area de Medicina Preventiva, Universidad de Santiago de Compostela, 15705 Santiago de Compostela, Spain.

the tests that have been publicized in the statistical and epidemiologic literature. We then summarize several key features of five meta-analyses of observational studies published recently in peer-reviewed journals. We conducted a large simulation study with a design based on the range of the key features of these meta-analyses. The aim of the simulation study was to compare the behavior of five tests, both in their asymptotic versions and in a bootstrap version that we developed. Since it is by far the most frequent situation, we restricted our study to the tests of heterogeneity that can be used in a multivariate setting and for which no raw data within individual studies are needed. We conclude this paper with some proposed alternatives to hypothesis testing for evaluating heterogeneity in epidemiologic meta-analyses.

THE MODELS AND THE TEST STATISTICS

The two primary models used in meta-analysis to obtain a pooled point and interval estimate of effect are the fixed effects model and the random effects model. A third approach, the mixed effects model, is an extension of the random effects model (6). The fixed effects model assumes that the S individual studies to be meta-analyzed are the universe of interest. On the contrary, the random effects model assumes that the S studies to be meta-analyzed are a random sample of a hypothetical universe of studies of the same risk factor-disease relation that will be published in the future or have already been published in unknown journals (7). The models are described below. The fixed effects model is

$$\hat{\beta}_s = \beta + e_s$$

and the random effects model is

$$\hat{\beta}_s = \beta + b_s + e_s,$$

where $E(b_s) = E(e_s) = 0$, $\text{var}(b_s) = \tau^2$, and $\text{var}(e_s) = \sigma^2_{w,s}$. $\hat{\beta}_s$ is the estimate of effect from study s, its variance is $\sigma^2_{w,s}$ under the fixed effects model or $(\sigma^2_{w,s} + \tau^2)$ under the random effects model, and $s = 1, \ldots, S$.

The variance of e_s is a function of the number of subjects included and other features of the individual study design. The variance of b_s is τ^2, the between-study variance. The fixed effects model is a particular case of the random effects model, where τ^2 is 0. Homogeneity tests focus on the null hypothesis, H_0: $\tau^2 = 0$, and, if this hypothesis is believed to be true, a pooled estimate is computed under the fixed effects model. The random effects model can be used regardless of the result of any hypothesis test. Under this model, the point estimate of the pooled effect measure and its confidence interval incorporate the additional variability due to between-study variance. This yields confidence intervals that are wider than those in the fixed effects model.

Several tests have been used in meta-analytic settings in epidemiology to address heterogeneity. Mantel and Haenszel (8) and Yusuf et al. (9) have proposed heterogeneity tests that require the raw cell counts and/or do not generalize to a multivariate setting. Since in meta-analysis of observational studies an adjusted estimate of the relative risk is the only valid option, these tests are generally not applicable for epidemiology and will not be considered further here. In addition, we restrict this article to the study of between-study heterogeneity in the relative risk and odds ratio. Risk differences are beyond the scope of this paper and will not be considered.

DerSimonian and Laird's *Q* test

The heterogeneity test described by DerSimonian and Laird (10) was previously proposed and discussed by Cochran (11). However, to avoid confusion and because this test is more often known by the name of the former authors, in this paper we refer to it as DerSimonian and Laird's Q test. It is by far the most popular of the heterogeneity tests used in meta-analysis in epidemiology. This noniterative test statistic consists of a weighted sum of squared deviations around the mean of the effect in each study, i.e.,

$$Q = \Sigma w_s(\hat{\beta}_s - \bar{\beta})^2,$$

where $\bar{\beta}$ is the weighted mean of the effects of each study (i.e., $\bar{\beta} = \Sigma w_s\hat{\beta}_s/\Sigma w_s$) and the weights w_s ($s = 1, \ldots, S$) are the inverse of the estimated variance of each individual study s (i.e., $w_s = [\text{var}(\hat{\beta}_s)]^{-1}$). Under the null hypothesis, Q has a χ^2 distribution with $S - 1$ degrees of freedom (df), where S is the total number of studies, provided that each individual study has a sample size that is large compared with the number of studies and provided that $\text{var}(\hat{\beta}_s)$ is independent of $\hat{\beta}_s$.

Z^2_{WLS}

Z^2_{WLS} (WLS, weighted least squares) is a test statistic proposed in a recent paper by Lipsitz et al. (12). It is a modification of DerSimonian and Laird's Q statistic. The statistic is

$$Z^2_{WLS} = (Q - S + 1)^2/2(S - 1).$$

Under the null hypothesis, Z^2_{WLS} has an F distribution with 1 and $S - 1$ df, if both the within-study sample

sizes and the number of studies are large. Further details on the derivation of this test and the two that follow are given in the paper by Lipsitz et al. (12). Since these tests have been newly proposed, meta-analysts need guidance as to when these tests may provide improved performance over Q.

$Z^2_{WLS,R}$

The $Z^2_{WLS,R}$ test was also derived from DerSimonian and Laird's Q statistic by Lipsitz et al. (12). It has the following form:

$$Z^2_{WLS,R} = (Q - S)^2/\{\Sigma[(\hat{\beta}_s - \bar{\beta})^2/(\text{var}(\hat{\beta}_s) - 1)]\}^2.$$

The "R" in the subscript is a mnemonic for "robust" variance estimate. If both S and the sample sizes within studies are large, $Z^2_{WLS,R}$ will have a χ^2 distribution with $S - 1$ df.

Z^2_K

The Z^2_K statistic was also proposed by Lipsitz et al. (12) and has the form

$$Z^2_K = \{\Sigma[(\hat{\beta}_s - \bar{\beta})^2 - \text{var}(\hat{\beta}_s)]\}^2/$$
$$\Sigma[(\hat{\beta}_s - \bar{\beta})^2 - \text{var}(\hat{\beta}_s)]^2.$$

Under the null hypothesis, this statistic will have an F distribution with 1 and $S - 1$ df, if S is large. Although it is rarely of interest in epidemiologic meta-analysis, notice that it is not necessary that the within-study sample sizes be large.

The likelihood ratio test

The application of a likelihood ratio test (LRT) statistic to heterogeneity problems has been described by Stram and Lee (13). It is the difference between two iterative statistics, comparing the log likelihood under the fixed effects model with the log likelihood under the random effects model, i.e.,

$$\text{LRT} = 2(L_{\text{fixed}} - L_{\text{random}}),$$

where L_{fixed} and L_{random} are the log likelihoods of the data under the fixed and random effects models, respectively, assuming a normal distribution for both the random effects, b_s, and the error terms, e_s, $s = 1, \ldots, S$. Under the null hypothesis, the statistic is a mixture of two χ^2 distributions with 0 df and 1 df, with a mixing parameter of ½.

τ^2-bootstrap

We use a parametric bootstrap approximation (described below) to find the cumulative distribution function of $\hat{\tau}^2$, the estimator of the between-study variance described by DerSimonian and Laird (10), where

$$\hat{\tau}^2 = \max\{0,[Q - (S - 1)]/[\Sigma w_s - (\Sigma w_s^2/\Sigma w_s)]\}.$$

We reject the null hypothesis of homogeneity if the bootstrapped empirical fifth percentile is greater than 0, the null value of τ^2, and we fail to reject the null hypothesis if the bootstrapped empirical fifth percentile is 0. Because τ^2 is bounded on the left by 0, a one-sided hypothesis test is the relevant procedure here. Note that this method does not give us a p value but rather a yes/no answer only. In addition, it assumes normality of the random effects and the error terms.

Parametric bootstrap versions of the tests

Because the number of studies, S, is typically small in epidemiologic meta-analyses, asymptotic distributions may be poor approximations of the true distribution of the test statistics. A bootstrap approach may overcome this drawback. The bootstrap procedure we propose is as follows. Sample with replacement from the data $(\text{var}(\hat{\beta}_1), \ldots, \text{var}(\hat{\beta}_S))$ S points $(\text{var}(\hat{\beta})^{(1)}, \ldots, \text{var}(\hat{\beta})^{(S)})$. Given $\bar{\beta}$ and $\text{var}(\hat{\beta})^{(s)}$, $s = 1, \ldots, S$, simulate S study results from the fixed effect distributions $N(\bar{\beta}, \text{var}(\hat{\beta}_s))$, $s = 1, \ldots, S$, to obtain $(\hat{\beta}_1{}^{*(b)}, \ldots, \hat{\beta}_s{}^{*(b)})$, and calculate the test statistics, e.g., $Q^{(b)}$. Repeat this procedure B times to obtain, for example, $(Q^{(1)}, \ldots, Q^{(B)})$. Using the cumulative histogram of the B bootstrapped values of the statistic under the null hypothesis, calculate the empirical exceedance probability for the observed statistic.

In the first runs of our simulations, we carried out 5,000 bootstrap resamples. The results were virtually the same as when only 1,000 resamples were used. Thus, we conducted the rest of the simulations with 1,000 bootstrap resamples only. We also smoothed the distribution function of the $B = 1{,}000$ values of the test statistic to obtain the corresponding bootstrap p values. Because these more time-consuming results were very close to those obtained through the simple histogram, we recommend the latter approach.

A user-friendly Fortran 77 program for obtaining bootstrapped and asymptotic p values for all test statistics studied in this paper is available from the second author (C. C.-S.).

EXAMPLES OF META-ANALYSES

We chose five meta-analyses from the literature to illustrate the tests considered in this paper, and after which we patterned our simulation study. The choice of these five meta-analyses was based on several criteria. We wanted to include recent meta-analyses only, those published no earlier than 1990 in peer-reviewed journals, to reflect current practice in handling heterogeneity. We chose meta-analyses that covered a range of S's and had varying magnitudes of estimated between-study variance.

The paper by Mortimer et al. (14) is a meta-analysis of 11 case-control studies that investigated the association between head trauma and subsequent Alzheimer's disease. For comparability reasons, in their final analysis the authors focused on head injury with loss of consciousness, which limited the number of studies to seven. The article by Everhart and Wright (15) is a meta-analysis of 20 cohort and case-control studies that investigated the relation between diabetes of more than 1 year's duration and subsequent occurrence of pancreatic cancer. The Gerstein (16) paper is a meta-analysis of 14 case-control studies that related early intake of cow's milk to subsequent type I diabetes mellitus. Spector and Hochberg's (17) article is a meta-analysis of six case-control and three cohort studies that looked for a possible protective effect of oral contraceptives on the occurrence of rheumatoid arthritis. As an alternative to hypothesis testing for heterogeneity, the authors used the graphic "odd-man-out" approach (3). The paper by Romieu et al. (18) is the largest meta-analysis in our series. It contains five cohort studies and 28 case-control studies on the relation between oral contraceptives and breast cancer. The authors presented separate results for each study design. In order to investigate the ability of the test statistics to detect obvious between-study heterogeneity, we computed an overall estimate.

Table 1 displays key features of each of the five meta-analyses. The five meta-analyses chosen cover a wide range of numbers of individual studies (from 7 to 33). In an informal review of the literature, we found that meta-analyses in epidemiology are generally based on 15–20 individual studies. When authors suspect or find heterogeneity, pooled analyses are then based on subgroups (e.g., by type of study) of further reduced sample size.

The average number of cases of disease in the individual studies included in these meta-analyses ranged between 151 and 621. Although this number of cases is rarely available in meta-analyses, the number of cases is a good indication of the amount of information on which the individual studies were based. Using the number of subjects included in each individual study is misleading, as cohort studies generally have larger sample sizes than case-control studies but

TABLE 1. Key features of five published meta-analyses used to pattern a simulation study of the performance of heterogeneity tests

Feature	Meta-analysis				
	Mortimer et al. (14)	Everhart and Wright (15)	Gerstein (16)	Spector and Hochberg (17)	Romieu et al. (18)
No. of studies included (S)	7	20	14	9	33
Mean no. of cases	151	176	205	156	621
Method used for pooling	Not stated	Random effects	Random effects	Fixed effects	Random effects
Pooled odds ratio	1.70	1.61	1.23	0.76	1.10
p value for H_0: $\beta = 0$	0.017	0.000	0.000	0.000	0.000
SE†$(\bar{\beta})/\lvert\bar{\beta}\rvert$	0.42	0.13	0.26	0.28	0.20
τ^2‡	0.00	0.1412	0.0718	0.1095	0.0148
CV_B§	0.00	0.788	1.307	1.222	1.256
R_I¶	0.00	0.65	0.65	0.68	0.53
Method used for test of heterogeneity	"Interaction terms not significant in logistic regression"	Q	Q	χ^2 (not specified)	Q

† SE, standard error.
‡ τ^2, variance between studies (DerSimonian and Laird's formula (10)).
§ CV_B, between-study coefficient of variation ($\sqrt{\tau^2}/\lvert\bar{\beta}\rvert$).
¶ R_I, proportion of the total variance due to between-study variance: $\tau^2/(\tau^2 + (S \times \text{var}(\bar{\beta})))$.

do not always provide more precise estimates of effect.

The most popular method of pooling results was the use of DerSimonian and Laird's random effects model (10). This method was used in three of the meta-analyses, each of which also tested for homogeneity using DerSimonian and Laird's method. In one meta-analysis, the inverse variance weights were used without further details; an unspecified χ^2 test was used to test for homogeneity in another. In Mortimer et al.'s meta-analysis (14), no clear explanation was given on how heterogeneity was handled, although one may guess that a 6 df LRT was used.

The pooled estimate of the relative risk was moderate in all five meta-analyses, ranging from 0.76 to 1.70. However, all pooled estimates were statistically different from the null value. The proportion of total variance due to between-study variance (R_I), where $R_I = \tau^2/(\tau^2 + S\ \text{var}(\bar{\beta}))$, was 0 in the smallest meta-analysis and was substantial and of similar magnitude in the other four studies, whose R_I's were approximately 60 percent. Another way to quantify heterogeneity is through the between-study coefficient of variation, CV_B, which is $\sqrt{\tau^2}/|\bar{\beta}|$. This measure has the disadvantage that it increases rapidly as β nears 0 and is undefined when β equals 0. R_I does not have this drawback, but it has the disadvantage that it increases as the within-study variances decrease for the same τ^2.

As table 2 shows, within each meta-analysis, p values from the different tests differed considerably, leading to contradictions in the decision process about rejection or acceptance of the null hypothesis of homogeneity. The asymptotic DerSimonian and Laird's Q test yielded p values that were close to those of its bootstrap version. Except for Mortimer et al.'s small meta-analysis (14), the null hypothesis of homogeneity was rejected by Q in all cases. In general, the concordance in the acceptance/rejection decision at the type I error rate of 0.05 between an asymptotic test and its bootstrap version held for every test except $Z^2_{WLS,R}$, which was designed to be adequate when there are both a large number of studies and large sample sizes within each study. This test led to conflicting decisions for two of the five meta-analyses, including Romieu et al.'s (18), which was based on a relatively large number of studies. Z^2_K in its asymptotic version as well as in its bootstrap version did not reject the null hypothesis of homogeneity in any of the five cases. The LRT yielded similar results in its asymptotic and bootstrap versions, which were close to the result obtained by the Q test. The τ^2-bootstrap test results were in accordance with those obtained by Q test and LRT. Because of these conflicting results, we undertook a simulation study to better understand the properties of these tests.

SIMULATION STUDY

The objective of the simulation study was to compare the behavior of the statistics Q, Z^2_{WLS}, $Z^2_{WLS,R}$, Z^2_K, and LRT, the bootstrap version of each of these statistics, and τ^2-bootstrap. We compared the type I error rates and statistical power of the tests. All tests were performed at the nominal level of 5 percent. Whenever required, the range of parameters needed by the simulation study was restricted to that observed in table 1.

To study both the type I error and power, we needed to select values for the within-study variance σ^2_w that were representative of those found in practice. In order to obtain plausible values for this parameter for each

TABLE 2. Resulting *p* values from heterogeneity tests (asymptotic and bootstrap versions) of five published meta-analyses that were used to pattern a simulation study of the performance of heterogeneity tests

Test	Meta-analysis				
	Mortimer et al. (14)	Everhart and Wright (15)	Gerstein (16)	Spector and Hochberg (17)	Romieu et al. (18)
Q	0.885	0.000	0.001	0.002	0.000
Q*	0.885	0.001	0.001	0.002	0.001
Z^2_{WLS}†	0.330	0.000	0.001	0.004	0.000
Z^2_{WLS}*	0.241	0.000	0.001	0.016	0.000
Z^2_K	0.180	0.130	0.084	0.144	0.304
Z^2_K*	0.084	0.220	0.110	0.278	0.371
$Z^2_{WLS,R}$	0.010	0.016	0.082	0.071	0.057
$Z^2_{WLS,R}$*	0.048	0.089	0.203	0.239	0.013
LRT†	0.500	0.001	0.002	0.004	0.000
LRT*	0.631	0.000	0.000	0.000	0.000
τ^2-bootstrap	Nonsignificant	Significant	Significant	Significant	Significant

* Parametric bootstrap version of the test.
† WLS, weighted least squares; LRT, likelihood ratio test.

value of the effect measure studied, we plotted the upper bound of the 95 percent confidence interval of the log relative risk reported by every study included in the five meta-analyses against the estimate of the log relative risk. We thus obtained a scatterplot of 83 studies. Through least squares regression, we fitted the quadratic curve $\ln UB = \gamma_0 + \gamma_1(\ln OR)_s + \gamma_2 j(\ln OR)_s^2$, where ln UB is the logarithm of the upper bound of the 95 percent confidence interval of the relative risk estimate for each individual study and OR (odds ratio) is the corresponding relative risk estimate. We used the upper and lower bounds of the 95 percent confidence interval for this curve to obtain a plausible range of values for σ^2_w for each value of the effect measure considered, by solving for the variance of ln OR as $\text{var}(\hat{\beta}_s) = [(\ln UB - \beta)/1.96]^2$.

Type I error

The null hypothesis of homogeneity of effects is

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_s = \beta \Leftrightarrow H_0: \tau^2 = 0.$$

The model used was $\hat{\beta}_s = \beta + e_s$, where $e_s \sim N(0, \sigma^2_{w,s})$, $s = 1, \ldots, S$, and the parameters needed were β, the measure of effect S, and the within-study variance $\sigma^2_{w,s}$, $s = 1, \ldots, S$. Four different values of β were chosen, corresponding to odds ratios of 1, 1.5, 2, and 5. Three values of S were used, 7, 20, and 40, corresponding to the range of values in table 1. To simulate N meta-analytic data sets of the form $(\hat{\beta}_s^{(i)}, \text{var}(\hat{\beta}_s)^{(i)})$, $s = 1, \ldots, S$ and $i = 1, \ldots, N$, where N was fixed at 5,000, we sampled S values of UB from the uniform distribution bounded by the upper and lower confidence limits of the quadratic regression curve (described above) of ln UB on $\hat{\beta}$ at the given value of β, and solved for $\text{var}(\hat{\beta}_s)^{(i)}$, $s = 1, \ldots, S$. Under the fixed effects model, $\text{var}(\hat{\beta}_s) = \sigma^2_{w,s}$. Thus, $\beta_1^{(i)}, \ldots, \beta_S^{(i)}$ were generated from the S normal distributions with mean β and $\text{var}(\hat{\beta}_s)^{(i)}$, and the test statistics from the simulated data $(\beta_1^{(i)}, \ldots, \beta_S^{(i)}, \text{var}(\hat{\beta}_1)^{(i)}, \ldots, \text{var}(\hat{\beta}_S)^{(i)})$ were calculated. The proportion of times in the N simulations in which the null hypothesis was rejected is the type I error rate.

Statistical power

Simulations assuming normal distributions. For power, we used a procedure similar to that used in the simulation study of type I error. However, under the alternative hypothesis, to generate β_s under the model

$$\beta_s = \beta + b_s + e_s,$$

where $e_s \sim N(0, \sigma^2_{w,s})$ and $b_S \sim N(0, \tau^2)$, τ^2 was needed. The variance between studies was used by fixing R_I, the proportion of total variance due to variance between studies, and by using the $\text{var}(\bar{\beta})$ corresponding to that observed in the meta-analysis of size S. This variance between studies is as follows:

$$\tau^2 = \{[R_I \times S \times \text{var}(\bar{\beta})]\}/(1 - R_I).$$

The values of R_I chosen were 0.1, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.75, 0.8, and 0.9.

Simulations assuming nonnormal distributions. In general, in epidemiologic research, the sample size within studies is large and the number of studies S is relatively small. Hence, under the random effects model,

$$\beta_s = \beta + b_s + e_s.$$

It is reasonable to assume that $e_s \sim N(0, \sigma^2_{w,s})$, $s = 1, \ldots, S$. However, the assumption that $b_s \sim N(0, \tau^2)$ is more likely to be violated. To investigate the robustness of the results obtained to a nonnormality of this nature, we conducted an additional set of simulations to investigate the power of the tests. First, we assumed that b_s followed a parametric but asymmetric distribution, the exponential distribution. Next, we assumed that b_s followed an empirical distribution estimated from the data given by the five meta-analyses considered in this paper.

Exponential case. The exponential case simulation follows a procedure similar to the one used in the power simulation explained above. All of the steps are the same, except that b_s follows a mean-centered exponential distribution with parameter $1/\tau$. Thus, b_s has mean 0 and standard deviation τ. As was explained above, τ is obtained by fixing R_I.

Empirical distribution. In addition to the fact that no parametric distribution is assumed for b_s, the simulation procedure in the case of an empirical distribution differs from the normal and exponential case described above in that R_I, τ^2, and S are not chosen before the simulation but are intrinsically defined by each of the five meta-analyses in table 1. The nonparametric simulation consists of the following steps:

a. From the meta-analytic data set at hand $(\hat{\beta}_s, \text{var}(\hat{\beta}_s))$, $s = 1, \ldots, S$, compute $\bar{\beta}_{RE} = \Sigma w_s^{RE} \hat{\beta}_s / \Sigma w_s^{RE}$, the weighted mean of the effects under the random effects (RE) model, where $w_s^{RE} = [\text{var}(\hat{\beta}_s) + \tau^2]^{-1}$.
b. For each $\hat{\beta}_s$, $s = 1, \ldots, S$ compute $b_s = \hat{\beta}_s - \bar{\beta}_{RE}$.
c. Create the smoothed cumulative distribution function $\hat{F}_S$ of b_s.

d. For each simulation i, $i = 1, \ldots, N$, do the following:
  d.1. For each study s, $s = 1, .., S$, generate $u_s^{(i)}$ from $U[0,1]$.
  d.2. Calculate $\hat{F}_S^{-1}(u_s^{(i)}) = b_s^{(i)}$.
  d.3. Generate $e_s^{(i)}$ from $N(0, \text{var}(\hat{\beta}_s))$.
  d.4. Compute $\hat{\beta}_s^{(i)} = \bar{\beta}_{RE} + b_s^{(i)} + e_s^{(i)}$.

Finally, the test statistics from the simulated data $(\hat{\beta}_1^{(i)}, ..., \hat{\beta}_s^{(i)}, \text{var}(\hat{\beta}_1), ..., \text{var}(\hat{\beta}_s))$ are calculated. The proportion of times in the N simulations in which the null hypothesis is rejected is the power.

RESULTS

In the studies of both type I error and of power, the results were virtually identical for every odds ratio considered. We thus present and discuss only the case in which the odds ratio is equal to 2.

Type I error

The type I error results are presented in table 3. The 95 percent confidence interval for the nominal level (5 percent) was 4.40 percent–5.60 percent for these 5,000 simulations. Test results that fell outside of this range are printed in boldface type. Except for the Q statistic, which performed well in every setting considered, the asymptotic tests did not give the correct size under the null hypothesis. This behavior was consistent for all values of S considered. The worst results were given by $Z^2_{WLS,R}$, where the type I error rate was always anticonservative. The Z^2_{WLS} and LRT statistics, on the contrary, were overly conservative, although both improved as the number of studies included in the meta-analysis increased. The behavior of the Z^2_K statistic was unpredictable and unacceptable.

The parametric bootstrap approximation of the same tests yielded good results for Z^2_{WLS}, $Z^2_{WLS,R}$, Z^2_K, and LRT. The τ^2-bootstrap method also gave results that were in agreement with the correct size. The results of the bootstrap approximation of Q were worse than those obtained by the asymptotic version: All of the sizes exceeded the correct one, and results worsened with increasing number of studies.

Statistical power

For power, the simulations performed in the exponential setting yielded results that were similar to those obtained in the normal setting. The simulations carried out in a nonparametric setting were in accordance with those obtained in the parametric cases. The departure from normality (due to a small number of studies) of relative risk estimates apparently had little effect on the behavior of the tests. Thus, only the results of power simulations performed under a normal setting are shown in table 4 and figure 1.

Although the validity of the new tests suggested by Lipsitz et al. (12) (Z^2_{WLS}, $Z^2_{WLS,R}$, and Z^2_K) is questionable, at least in the settings considered in our simulation study, which mirror the meta-analytic environment of epidemiology, we proceeded nonetheless with an evaluation of the power of all of the tests. As expected, the power increased dramatically as R_I increased to 1. The power also depended upon S, for the same R_I, increasing with increasing S, but not as strongly as the dependency upon R_I. The Q statistic performed better than the other asymptotic test statistics. The LRT and the Z^2_{WLS} statistic exhibited similar behavior, with results that were slightly worse than those of the Q test. The performance of these two statistics was improved by the use of the parametric bootstrap approach.

Except with large values of R_I, the Z^2_K statistic had low power that was even lower in the bootstrap version. The behavior of $Z^2_{WLS,R}$ was not consistent and was not improved by the use of the bootstrap approach. Except for the highest values of R_I and S, this test had poor power.

The performance of the τ^2-bootstrap test was as good as that of the Q statistic. The results of the simulations performed in the exponential and distribution-free settings were consistent with the conclusions obtained

TABLE 3. Type I error rates (%) of heterogeneity tests (asymptotic and bootstrap versions) obtained in a simulation study, according to the number of studies (S) included in the meta-analysis†

	Test										
	Q	Z^2_{WLS}‡	$Z^2_{WLS,R}$	Z^2_K	LRT‡	Q*	Z^2_{WLS}*	$Z^2_{WLS,R}$*	Z^2_K*	LRT*	τ^2-bootstrap
$S = 7$	4.96	**2.20**§	**22.16**	**2.62**	**1.74**	**6.24**	5.08	4.74	4.80	4.80	4.66
$S = 20$	5.16	**3.90**	**13.64**	**12.06**	**2.10**	**7.96**	5.12	4.84	4.72	4.70	4.82
$S = 40$	5.46	**4.18**	**10.10**	**9.64**	**3.20**	**9.24**	5.22	5.26	4.78	5.00	5.18

* Parametric bootstrap version of the test.
† Odds ratio = 2.
‡ WLS, weighted least squares; LRT, likelihood ratio test.
§ Values printed in boldface are those for which the error rate was outside of the 95% confidence region.

TABLE 4. Statistical power (%) of heterogeneity tests (asymptotic and bootstrap versions) obtained in a simulation study, according to the number of studies (S) included in the meta-analysis and the proportion of total variance due to between-study variance†

	Test										
	Q	Z^2_{WLS}‡	$Z^2_{WLS,R}$	Z^2_K	LRT‡	Q*	Z^2_{WLS}*	$Z^2_{WLS,R}$*	Z^2_K*	LRT*	τ^2-bootstrap
R_I§ = 0.25											
S = 7	14.38	8.84	13.66	1.40	7.85	12.18	14.12	2.40	2.44	14.68	15.32
S = 20	25.90	20.68	6.18	4.58	17.90	24.18	24.12	1.22	1.24	25.97	24.92
S = 40	38.12	31.20	9.48	5.44	30.64	35.10	33.42	1.38	0.94	37.81	37.94
R_I = 0.5											
S = 7	37.54	28.42	6.68	0.50	27.14	27.60	37.32	1.02	1.14	39.10	38.50
S = 20	70.56	64.28	17.82	9.06	64.11	58.54	68.10	0.32	0.26	69.21	68.62
S = 40	90.88	88.20	56.46	41.36	88.61	79.76	89.02	20.34	13.28	90.21	91.14
R_I = 0.75											
S = 7	77.66	70.76	5.50	0.08	70.21	58.84	77.68	0.16	0.14	78.31	75.02
S = 20	98.66	98.24	67.54	53.60	97.51	93.64	98.46	4.76	6.06	99.12	98.34
S = 40	100.0	100.0	98.12	97.00	99.93	99.64	100.0	83.16	80.14	99.94	99.94

* Parametric bootstrap version of the test.
† Odds ratio = 2.
‡ WLS, weighted least squares; LRT, likelihood ratio test.
§ R_I, proportion of the total variance due to between-study variance: $\tau^2/((\tau^2 + (S \times \text{var}(\bar{\beta})))$.

from the normal simulations. However, the exponential simulations yielded some differences. For Q, the power figures were 1–19 percent lower than those in the normal setting. For Q*, on the contrary, the figures were always higher. For LRT, they were up to 30 percent lower. These differences may seem substantial in proportion, but since the power even in the normal simulations is low in most cases considered, this variation is not meaningful.

DISCUSSION

The results of our simulation study, patterned on the features of real meta-analyses, indicate that no matter what the underlying relative risk and number of individual studies included in the meta-analysis are, four out of five asymptotic tests do not give the correct size under the null hypothesis. The exception was the Q statistic. Although the Q statistic is strictly appropriate only when $\text{cov}(\hat{\beta}_s, \text{var}(\hat{\beta}_s)) = 0$, an assumption which is violated by binomial data as in these epidemiologic meta-analyses (19), violation of this assumption did not have a detectable adverse impact on the validity of the test in the settings considered. This issue may be more of a concern for estimation than for testing. On the contrary, errors in the decision of whether or not to reject the null hypothesis of homogeneity of effects through studies are highly likely for the asymptotic tests, and these tests should probably not be used in epidemiologic meta-analyses.

The parametric bootstrap corrects the poor type I error rate for all tests for which asymptotic results were deficient. The surprisingly poor result for the bootstrapped Q, Q*, and the good result obtained by Z^2_{WLS} may be explained by the fact that Z^2_{WLS}* is close to a studentized version of Q*. In a bootstrap context, a studentized variable is one in which we subtract the original statistic from the bootstrap statistic and divide by the standard error of the bootstrap statistic (20). A "true" studentized Q would be Q* minus Q divided by the standard error of Q*. Bootstrap experts recommend the use of studentized quantities to improve results (20, p. 324). To confirm this, we recalculated the type I error rates in the simulations for the "true" studentized Q and obtained values for S = 7, S = 20, and S = 40 of 4.89 percent, 4.92 percent, and 5.27 percent, respectively, as expected. However, since the type I error rate is correct for the asymptotic version of Q, the bootstrap version is unnecessary.

The results of the power simulations show that for small values of R_I, no test has a satisfactory power of, for example, 80 percent. It may thus be deceptive to use any homogeneity test when the proportion of between-study variance is lower than 0.4, as long as $S \leq 40$. For values of R_I that range between 0.4 and 0.75, one needs to consider the value of S in deciding whether the test will perform satisfactorily. Small meta-analyses (of less than 10 studies, approximately) will still have unsatisfactory power in this range, but performing a test in a larger meta-analysis may be reasonable. The same careful attention is needed in situations where R_I is large (>0.75) and the number of studies is moderate.

In summary, this exercise generated both good news and bad news. The good news is that the Q test is

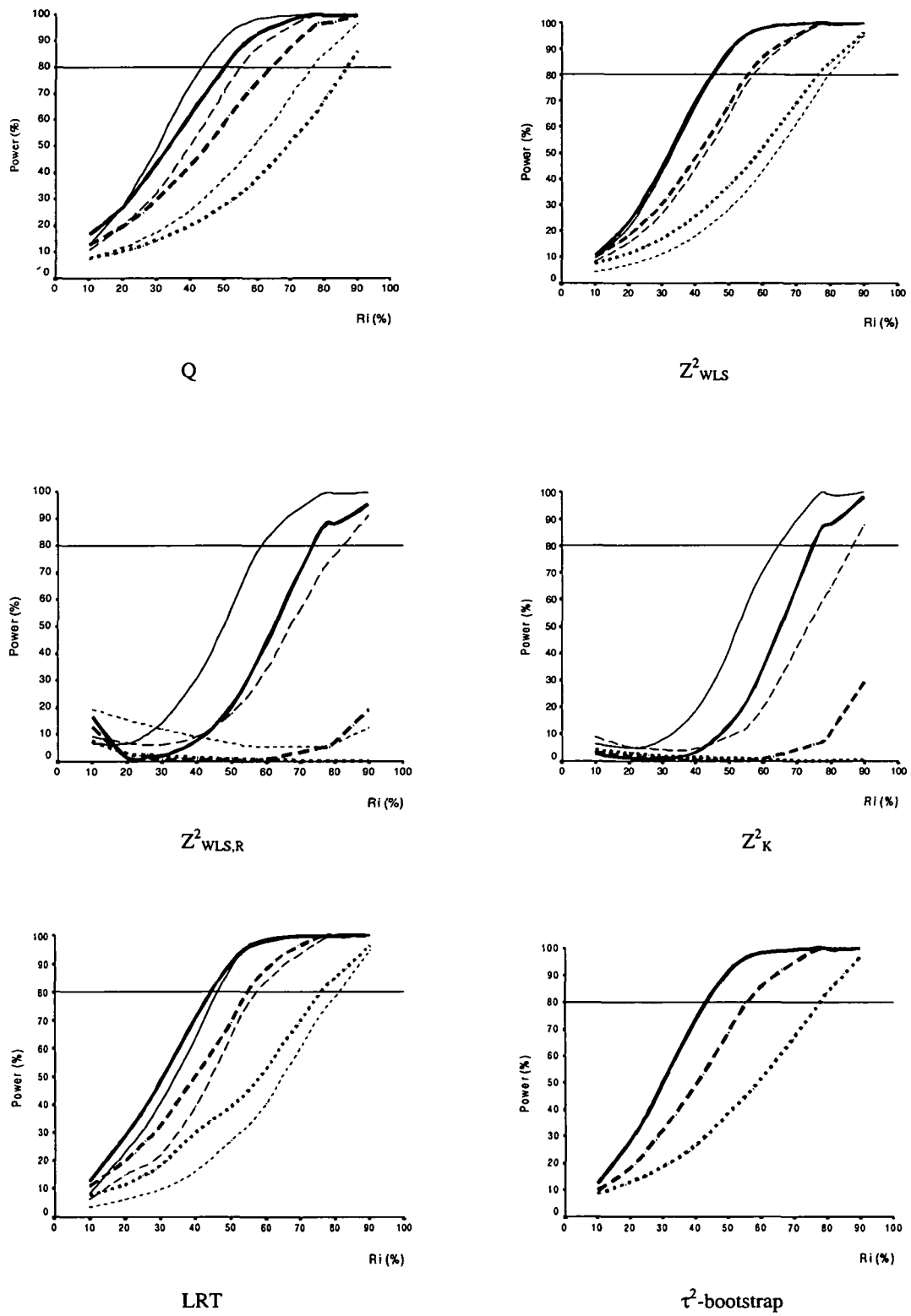

FIGURE 1. The statistical power of heterogeneity tests (asymptotic and bootstrap versions) in meta-analysis as a function of the proportion of total variance due to between-study variance (R_i) and the number of studies (S) included in the meta-analysis (odds ratio = 2). (WLS, weighted least squares; LRT, likelihood ratio test). Key: —, $S = 40$ (bootstrap); —, $S = 40$ (asymptotic); – – –, $S = 20$ (bootstrap); – – –, $S = 20$ (asymptotic); - - - -, $S = 7$ (bootstrap); - - - -, $S = 7$ (asymptotic).

clearly the best from the point of view of size, power, and computational simplicity. Although the LRT and the τ^2-bootstrap test have statistical properties that are nearly as good, both are relatively complex computationally. In addition, τ^2-bootstrap has the drawback that it does not provide a numerical p value, and both of these tests impose normality assumptions that are empirically difficult to verify. The new tests proposed by Lipsitz et al. (12) are not useful in epidemiologic meta-analyses.

The bad news is that for the typical "sample sizes" seen in epidemiologic meta-analysis, no available test has acceptable power, unless heterogeneity is quite pronounced ($R_I \geq 0.75$). Because these tests often falsely fail to detect true heterogeneity, it may be advisable to use random effects models routinely, as suggested by the National Research Council (4). Since the results of hypothesis tests are a function of sample size and other arbitrary design features, in other instances they may reject the null hypothesis when the magnitude of the differences is very small and substantively inconsequential. In this case, when there are many studies included in the meta-analysis (e.g., $S \geq 100$), these tests may reject the null hypothesis with a very small degree of between-study heterogeneity. In either the possibly underpowered case or the overpowered case, heterogeneity can be jointly quantified by R_I and CV_B, and mixed effects regression models can be used to investigate and adjust for identifiable sources of heterogeneity when R_I and CV_B indicate that the magnitude of the between-study heterogeneity is sufficiently large (21).

These results show that the asymptotic DerSimonian and Laird Q statistic and the bootstrap versions of the other tests give the correct type I error under the null hypothesis but that all of the tests considered have low statistical power, especially when the number of studies included in the meta-analysis is small (<20). From the point of view of validity, power, and computational ease, the Q statistic is the best choice. However, in addition to the application of statistical techniques, common sense and a priori biologic knowledge, to the extent that it exists, must be vigilantly utilized when synthesizing the results of many studies.

ACKNOWLEDGMENTS

This work was partially completed during a scientific stay of Dr. Bahi Takkouche at the Harvard School of Public Health (Boston, Massachusetts) that was funded by a grant from the University of Santiago de Compostela (Santiago de Compostela, Spain). Dr. Carmen Cadarso-Suárez's work on this project was funded by grant XUGA 20701B96 from Xunta de Galicia (Santiago de Compostela, Spain). Dr. Donna Spiegelman's work was funded by grant CA55075 from the National Institutes of Health (Bethesda, Maryland).

The authors are grateful to Dr. Bernard Rosner for his helpful comments and to Julian Costa-Bouzas for his computing assistance.

REFERENCES

1. Colditz GA, Burdick E, Mosteller F. Heterogeneity in meta-analysis of data from epidemiologic studies: a commentary. Am J Epidemiol 1995;142:371–82.
2. Berlin JA. Invited commentary: benefits of heterogeneity in meta-analysis of data from epidemiologic studies. Am J Epidemiol 1995;142:383–7.
3. Walker AM, Martin-Moreno JM, Artalejo FR. Odd man out: a graphical approach to meta-analysis. Am J Public Health 1988; 78:961–6.
4. National Research Council. Combining information: statistical issues and opportunities for research. Washington, DC: National Academy Press, 1992:52.
5. Rothman KJ. A show of confidence. (Editorial). N Engl J Med 1978;299:1362–3.
6. Stram DO. Meta-analysis of published data using a linear mixed-effects model. Biometrics 1996;52:536–44.
7. Fleiss JL. The statistical basis of meta-analysis. Stat Methods Med Res 1993;2:121–45.
8. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 1959;22:719–48.
9. Yusuf S, Peto R, Lewis J, et al. Beta blockade during and after myocardial infarction: an overview of the randomized trials. Prog Cardiovasc Dis 1985;27:335–71.
10. DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled Clin Trials 1986;7:177–88.
11. Cochran WG. The combination of estimates from different experiments. Biometrics 1954;10:101–29.
12. Lipsitz SR, Dear KB, Laird NM, et al. Tests for homogeneity of the risk difference when data are sparse. Biometrics 1998; 54:148–60.
13. Stram DO, Lee JW. Variance components testing in the longitudinal mixed effects model. Biometrics 1994;50:1171–7.
14. Mortimer JA, van Duijn CM, Chandra V, et al. Head trauma as a risk factor for Alzheimer's disease: a collaborative reanalysis of case-control studies. EURODEM Risk Factors Research Group. Int J Epidemiol 1991;20(suppl 2):S28–35.
15. Everhart J, Wright D. Diabetes mellitus as a risk factor for pancreatic cancer: a meta-analysis. JAMA 1995;273:1605–9.
16. Gerstein HC. Cow's milk exposure and type I diabetes mellitus: a critical overview of the clinical literature. Diabetes Care 1994;17:13–19.
17. Spector TD, Hochberg MC. The protective effect of the oral contraceptive pill on rheumatoid arthritis: an overview of the analytic epidemiological studies using meta-analysis. J Clin Epidemiol 1990;43:1221–30.
18. Romieu I, Berlin JA, Colditz G. Oral contraceptives and breast cancer: review and meta-analysis. Cancer 1990;66:2253–63.
19. Emerson JD, Hoaglin DC, Mosteller F. A modified random-effect procedure for combining risk differences in sets of 2×2 tables from clinical trials. J Ital Stat Soc 1993;2:169–90.
20. Efron B, Tibshirani RJ. An introduction to the bootstrap. New York, NY: Chapman and Hall, 1993.
21. Berkey CS, Hoaglin DC, Mosteller F, et al. A random effects regression model for meta-analysis. Stat Med 1995;14:395–411.