STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:4321–4333
Published online 21 September 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2692

A simulation study comparing properties of heterogeneity measures in meta-analyses

M. Mittlböck[‡] and H. Heinzl[*,†]

Core Unit for Medical Statistics and Informatics, Medical University of Vienna, Austria

SUMMARY

The assessment of heterogeneity or between-study variance is an important issue in meta-analysis. It determines the statistical methods to be used and the interpretation of the results. Tests of heterogeneity may be misleading either due to low power for sparse data or to the detection of irrelevant amounts of heterogeneity when many studies are involved. In the former case, notable heterogeneity may remain unconsidered and an unsuitable model may be chosen and the latter case may lead to unnecessary complex analyses strategies. Measures of heterogeneity are better suited to determine appropriate analyses strategies. We review two measures with different scaling and compare them with the heterogeneity test. Estimates of the within-study variance are discussed and a new total information measure is introduced. Various properties of the quantities in question are assessed by a simulation study.

Heterogeneity test and measures are not directly related to the amount of between-study variance but to the relative increase of variance due to heterogeneity. It is more favourable to base the within-study variance estimate on the squared weights of individual studies than on the sum of weights. A heterogeneity measure scaled to a fixed interval needs reference values for proper interpretation. A measure defined by the relation of between- to within-study variance has a more natural interpretation but no upper limit. Both measures are quantifications of the impact of heterogeneity on the meta-analysis result as both depend on the variance of the individual study effects and thus on the number of patients in the studies.

KEY WORDS: measures of heterogeneity; between-study variance; within-study variance; fixed *versus* random effect model; total information

1. INTRODUCTION

One of the goals of a meta-analysis is to combine the results of previous research in order to arrive at summary conclusions to resolve uncertainty about the underlying medical question.

*Correspondence to: H. Heinzl, Core Unit for Medical Statistics and Informatics, Medical University of Vienna, Spitalgasse 23, A-1090 Vienna, Austria.
†E-mail: harald.heinzl@meduniwien.ac.at
‡E-mail: martina.mittlboeck@meduniwien.ac.at

Received 18 July 2006
Accepted 18 July 2006

A major problem is to decide whether the trials under review are statistically homogeneous or heterogeneous. Higgins *et al*. [1] defined statistical heterogeneity as 'The situation in which treatment effects being estimated by individual studies in a systematic review are not identical. This will manifest itself in greater variability in the estimates than would be expected by chance (sampling variation) alone'. Heterogeneity is a major issue in meta analyses, besides publication bias and outcome selection bias [2], as whenever the results of the trials are heterogeneous, the interpretation of the overall result becomes more difficult. Furthermore, the appropriateness of commonly used advanced methods may become questionable, e.g. Terrin *et al*. [3] show that the trim and fill method may inappropriately adjust for non-existing publication bias and that the funnel plot itself is inappropriate for heterogeneous meta-analyses. Meta-analysis of clinical trials with heterogeneous results provides an opportunity to learn about variations in treatment effectiveness [4–6]. Knowledge about the sources of heterogeneity can ultimately lead to better clinical understanding of the optimal way to treat patients [7].

Frequently, heterogeneity is assessed by a test, which is usually criticized to have low power. This happens especially if the meta-analysis consists of a small number of studies, whereas in the case of many studies irrelevantly small amounts of heterogeneity may be declared to be statistically significant when conventional levels of significance are used. Thus Higgins and Thompson [8, 9] have studied measures of heterogeneity which rather describe the impact of heterogeneity observed, than to test it. Their preferred measure, I^2, is scaled to values between zero (indicating no heterogeneity) and one (indicating extensive heterogeneity). Meanwhile it has been implemented in the Review Manager software of the Cochrane Collaboration [10] and it is frequently used.

Higgins and Thompson [8] distinguish between the extent and the impact of heterogeneity. A heterogeneity measure which describes the impact of heterogeneity will depend on the precisions (within-study variabilities) of the study-specific estimates, whereas a measure of the extent of heterogeneity will not, e.g. the estimate of the between-study variability itself. Consequently, the heterogeneity test and I^2 are concerned with the impact of heterogeneity.

In Section 2 the methods for meta-analyses are reviewed and their interpretation is discussed. In Section 3 a simulation study is performed to assess and compare the behaviour of the heterogeneity test, heterogeneity measures, measures of total information and the estimates of within-study variances. In Section 4 the results are summarized and further aspects of heterogeneity in meta-analyses are discussed.

2. SCIENTIFIC BACKGROUND

2.1. Estimation of summary effects

In a meta-analysis of k separate trials the overall treatment effect θ is estimated by averaging the observed treatment effects $\hat{\theta}_i$ $(i = 1, \ldots, k)$ over all trials. The effect $\hat{\theta}_i$ could be, e.g. the observed log odds ratio for a binary outcome, or the difference between two group means for a continuous outcome in the ith trial. In the fixed effect approach the true treatment effects across studies are assumed to be homogeneous ($\theta_1 = \cdots = \theta_k = \theta$) and $\hat{\theta}_i \sim \mathrm{N}(\theta, v_i)$. The precision of each study is denoted by w_i, and usually $w_i = 1/v_i$. We make the conventional assumption that the precisions are known, although in reality they are estimated from the data in each study. The overall treatment effect may then be estimated as weighted average $\bar{\theta}_F = (\sum_{i=1}^{k} w_i \hat{\theta}_i)/(\sum_{i=1}^{k} w_i)$.

DOI: 10.1002/sim

This type of weighting provides the most precise estimate of the true treatment effect. The variance of $\bar{\theta}_F$ is $v_F = 1/(\sum_{i=1}^{k} w_i)$.

Heterogeneity is present if the observed variance of the treatment effects $\hat{\theta}_i$ is greater than expected by chance. In such situations the additional between-study variability should be taken into account for estimating the overall treatment effect. This approach is called random effects meta-analysis. The estimates of the ith treatment effect is assumed to follow a $N(\theta_i, v_i)$ distribution, while the true treatment effects $\theta_1, \ldots, \theta_k$ are assumed to follow a $N(\theta, \sigma_B^2)$ distribution with between-study variance σ_B^2. The weights for this random-effects approach are $w_i^* = 1/(v_i + \sigma_B^2)$. The variance of the pooled estimate $\bar{\theta}_R = (\sum_{i=1}^{k} w_i^* \hat{\theta}_i)/(\sum_{i=1}^{k} w_i^*)$ is then given by $v_R = 1/(\sum_{i=1}^{k} w_i^*)$.

The decision between a fixed or a random effects model is often based on a test of heterogeneity, or equivalently, test of homogeneity [11], of the $\hat{\theta}_i$'s.

2.2. Test of heterogeneity

A test of heterogeneity considers the null hypothesis that the between-study variance component σ_B^2 is zero [12]. The test-statistic $Q = \sum_i w_i(\hat{\theta}_i - \bar{\theta}_F)^2$ is commonly used. Under the null hypothesis of homogeneity, the statistic Q follows a χ^2-distribution with df degrees of freedom, $df = k - 1$.

It has been claimed, that such a test has low power to detect departures from homogeneity [13] and its practical importance is probably limited [14]. Thus Petitti [5] recommends to prefer a significance level of 0.1 instead of the commonly used value of 0.05; Baujat [6] emphasizes that a significant test result must be considered as strong evidence for heterogeneity. This is true if only a small number of studies are included in the meta-analysis. However, if there are many studies included, small heterogeneity negligible for meta-analysis may be found to be statistically significant. Thus it seems advisable not only to consider the statistical test but also to describe the impact of heterogeneity on meta-analysis.

2.3. Estimation of within- and between-study variance

The expectation of Q under the assumption that the weights w_i are known, is

$$E[Q] = \sigma_B^2 \left(\sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i} \right) + (k-1)$$

Equating Q with its expectation provides a moment estimate for the between-study variance σ_B^2 [12]

$$\hat{\sigma}_B^2 = \frac{Q - (k-1)}{\sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i}}$$

Since $\hat{\sigma}_B^2$ is negative for $Q < (k-1)$, $\max(0, \hat{\sigma}_B^2)$ is used in practice as estimate of the between-study variance, so that the precision of a random effects summary estimate $(1/v_R)$ will not exceed the precision of a fixed effect summary estimate $(1/v_F)$.

DOI: 10.1002/sim

A 'typical' within-study variance can be estimated by the reciprocal of the arithmetic mean weight

$$\hat{\sigma}^2_{W,1} = k v_F = k \Big/ \left(\sum_{i=1}^{k} w_i \right)$$

as suggested by Takkouche *et al*. [15]. A second possibility [8] is

$$\hat{\sigma}^2_{W,2} = \frac{(k-1)\sum_{i=1}^{k} w_i}{\left(\sum_{i=1}^{k} w_i\right)^2 - \sum_{i=1}^{k} w_i^2}$$

As $E[Q] = (k-1)\left(\left(\sigma^2_B/\sigma^2_{W,2}\right)+1\right)$, $\sigma^2_{W,2}$ is directly related to both, $E[Q]$ and the power of the heterogeneity test. Consequently, $\hat{\sigma}^2_{W,2}$ is preferable to $\hat{\sigma}^2_{W,1}$ in order to assess the impact of the within-study variance on the power of the test. Note that $\hat{\sigma}^2_{W,1} = \hat{\sigma}^2_{W,2}$ if $w_1 = \cdots = w_k$.

2.4. Measures of total information

Hardy and Thompson [13] have introduced a measure of total information

$$\mathrm{TI} = \sum_{i=1}^{k} w_i = \frac{1}{v_F} = \frac{k}{\hat{\sigma}^2_{W,1}}$$

which is related to the power of the heterogeneity test. They state that the power is low, if the total information available is low. This happens, e.g. in the case of sparse data, or when one study is much more precise than the rest. Contrary to Hardy and Thompson [13] the total information is denoted by TI instead of I in order to avoid confusion with the measure I^2 (see below).

An alternative measure of total information can be based on $\hat{\sigma}^2_{W,2}$ and is

$$\mathrm{TI}_M = \frac{k}{\hat{\sigma}^2_{W,2}}$$

As $\hat{\sigma}^2_{W,2}$ is preferable to $\hat{\sigma}^2_{W,1}$ to assess the impact of within-study variance on the power, TI_M is preferable to TI to determine the amount of total information. This is demonstrated in the simulation study below.

2.5. Measures of heterogeneity

Higgins and Thompson [8] have investigated several measures of heterogeneity to quantify the heterogeneity present. Their criteria for adequate measures are: First, such a measure should depend on the extent of heterogeneity (measures of the impact of heterogeneity usually depend on the between-study variability and thus also on the extent of heterogeneity). Secondly, it should be scale invariant, so that a linear transformation of the parameter space should result in the same measure in order to enable comparisons across meta-analyses using different scales of measurement and using different types of outcome data. Thirdly, it should be size invariant so that the measure should not depend on the number of studies in the meta-analyses, as this would resemble to the disadvantage of the heterogeneity test.

DOI: 10.1002/sim

The two preferred measures of heterogeneity [8] are:

(a) $H^2 = Q/df$: it describes the relative excess in Q over its degrees of freedom [8, 16]. Its expectation is

$$E[H^2] = \frac{E[Q]}{k-1} = \frac{\sigma_B^2}{\sigma_{W,2}^2} + 1 = \frac{\sigma_B^2 + \sigma_{W,2}^2}{\sigma_{W,2}^2}$$

H^2 has a value of 1 in the case of homogeneity and it can be interpreted as a 'variance inflation factor' due to heterogeneity. In practice, $\max(1, H^2)$ is used.

A modification of H^2 is

$$H_M^2 = H^2 - 1 = \frac{Q - df}{df}$$

which is zero in the case of homogeneity and $\max(0, H_M^2)$ can be used in practice. $E[H_M^2] = \sigma_B^2/\sigma_{W,2}^2$ and thus on average H_M^2 gives the relation of between- to within-study variance. Of course, H^2 and H_M^2 share analogous properties.

(b)

$$I^2 = 100 \times \frac{Q - df}{Q} = 100 \times \left(1 - \frac{df}{Q}\right)$$

I^2 has been incorporated in the Review Manager Software of the Cochrane Collaboration [10]. For $\sigma_B^2 = 0$, Q follows approximately a χ_{df}^2-distribution and $1/Q$ is approximately inverse-χ_{df}^2 distributed with an expectation of $1/(df - 2)$ for $df > 2$. Thus $E[I^2] = 100 \times (-2/(k-3))$ for $\sigma_B^2 = 0$ and $k > 3$. This expectation is only defined for at least 4 studies and depends on the number of studies included, although this dependence becomes smaller if the number of studies increases. For $\sigma_B^2 > 0$, the properties of I^2 are investigated in the simulation study below.

In practice, negative values of I^2 are set equal to zero so that I^2 ranges from 0 to 100 per cent. A value of 0 per cent indicates no observed heterogeneity, and larger values show increasing heterogeneity, but exactly 100 per cent can never be achieved. Alternatively, I^2 can also be represented in terms of H^2 as $I^2 = 100 \times (H^2 - 1)/H^2$. It is notable that I^2 is analogous to the shrinkage factor $\gamma = (\chi^2 - df)/\chi^2$, where χ^2 is a test statistic with a χ_{df}^2-distribution under the null hypothesis (see, e.g. Copas [17–19]).

It seems natural to scale a heterogeneity measure between 0 and 100 per cent. Scaling should simplify interpretation and assessment of heterogeneity, e.g. vast heterogeneity for values close to one and minor heterogeneity for values close to zero. But the question for the interpretation between these two extreme values remains. Higgins and Thompson [8, 9] state that mild heterogeneity can be assumed if I^2 achieves values below 30 per cent. Notable heterogeneity is achieved for values which are substantially larger than 50 per cent.

The behaviour and properties of the heterogeneity measures and the heterogeneity test are investigated and compared in the following simulation study.

DOI: 10.1002/sim

3. SIMULATION STUDY

3.1. Power of heterogeneity test, measures of heterogeneity and total information

Hardy and Thompson [13] investigated three meta-analysis characteristics which influence the expectation of Q and therefore the power of the heterogeneity test:

(a) the extent of heterogeneity present, that is, the value of σ_B^2;
(b) the number of studies k included in the meta-analysis;
(c) the weight w_i allocated to each study.

The simulations scenarios of Hardy and Thompson [13] are described in Table I. They focussed on the behaviour of the power of the heterogeneity test. Here, these simulation scenarios are repeated in order to assess and compare heterogeneity measures, measures of total information and the power of the heterogeneity test. For each scenario 1000 simulation runs were carried out. The between-study variances σ_B^2 varied in steps of 0.05 from 0 to 0.5 or until a maximum value of $\sigma_B^2/\sigma_{W,2}^2 = 5$ was reached. Values of $\hat{\theta}_i$ are randomly drawn from a $N(\theta_i, v_i)$ distribution, where $v_i = 1/w_i$ is known and θ_i is randomly drawn from a $N(\theta, \sigma_B^2)$ distribution (with $\theta = 5$, but this choice is immaterial). For each simulation scenario, the power of the test is calculated, that is, the proportion of the 1000 simulation runs in which the test statistic was significant at the 5 per cent level. Furthermore, means and medians for the heterogeneity measures I^2 and H_M^2 are calculated. In practice $\max(0, I^2)$ and $\max(0, H_M^2)$ are used but here the use of only positive I^2 and H_M^2 values would result in a positive bias, especially if the true values are close to zero. Thus unrestricted values of H_M^2 and I^2 are used for the simulation study results.

From $E[Q]$ we learn that not the between-study variance σ_B^2 alone determines the power but rather the ratio $\sigma_B^2/\sigma_{W,2}^2$ which is also directly reflected by the expectation of the measure H_M^2 with

Table I. Description of the four simulation scenarios, where k denotes the number of studies involved, $w_1, \ldots, w_k$ are the individual weights, $\sigma_{W,1}^2$ and $\sigma_{W,2}^2$ are measures for the within-study variance and TI and TI_M are two measures for total information.

	k	w_1	$w_2, \ldots, w_k$	$\sigma_{W,1}^2$	$\sigma_{W,2}^2$	TI	TI_M
Scenario 1 (a)	5	10	10	0.1	0.1	50	50
(b)	10	10	10	0.1	0.1	100	100
(c)	20	10	10	0.1	0.1	200	200
Scenario 2 (a)	5	20	20	0.05	0.05	100	100
(b)	10	10	10	0.1	0.1	100	100
(c)	20	5	5	0.2	0.2	100	100
Scenario 3 (a)	10	5	5	0.2	0.2	50	50
(b)	10	10	10	0.1	0.1	100	100
(c)	10	20	20	0.05	0.05	200	200
Scenario 4 (a)	10	10	10	0.1	0.1	100	100
(b)	10	50	5.56	0.1	0.12	100	80.2
(c)	10	90	1.11	0.1	0.48	100	21.0

DOI: 10.1002/sim

$E[H_M^2] = \sigma_B^2/\sigma_{W,2}^2$. Thus the behaviour of the power and measures of heterogeneity are plotted against both, the between-study variance σ_B^2 and the ratio of between- and within-study variance $\sigma_B^2/\sigma_{W,2}^2$ (Figure 1).

In scenario 1 (Figures 1(a) and (b)), the power increases with increasing number of studies. The within-study variance is constant for all numbers k of studies involved. Thus the same behaviour of the power is observed independent of the scaling of the x-axis with either σ_B^2 or $\sigma_B^2/\sigma_{W,2}^2$. Figures 1(e) and (f) reflect the linear increasing relationship between the measure H_M^2 and both, σ_B^2 and $\sigma_B^2/\sigma_{W,2}^2$, whereas the corresponding relationship of I^2 is non-linearly monotone (Figures 1(c) and (d)). I^2 should approach one as σ_B^2 increases, but there has to be a vast amount of heterogeneity before I^2 yields a value close to this limit. Here, σ_B^2 is up to five times higher than $\sigma_{W,2}^2$ and a maximum mean value of 83 per cent is observed. For $\sigma_B^2 = 0$ and $k < 10$, the mean value of I^2 (Figures 1(c) and (d)) strongly depends on the number of studies k due to the underlying inverse-χ^2-distribution. For $\sigma_B^2 > 0$, this dependence on k decreases if σ_B^2 increases. The median estimates (not shown) are less dependent on the number of studies as the distribution of I^2 is skewed to the left. For H_M^2 there is no mean bias (Figures 1(e) and (f)), but the median estimates of H_M^2 decrease with decreasing number of studies k (not shown). The distribution of H_M^2 is right-skewed like the distribution of Q.

In scenario 2 (Figures 1(g)–(l)) higher weights ($w_i = 20, 10, 5$) are given for situations with smaller number of samples ($k = 5, 10, 20$), so that the total information remains constant with $\mathrm{TI} = \mathrm{TI}_M = 100$. This results in equal power for all sample sizes (Figure 1(g)) for the same between-study variance σ_B^2. For a fixed value of σ_B^2 stronger evidence of heterogeneity is observed for meta-analyses with few studies and small within-study variances than for meta-analyses with many studies and large within-study variances (Figures 1(i) and (k)). However, if σ_B^2 is rescaled by $\sigma_{W,2}^2$, then the results of scenario 2 (Figures 1(h), (j) and (l)) will resemble those of scenario 1 (Figures 1(b), (d) and (f)), respectively.

In scenario 3 (Figures 1(m)–(r)), 10 studies are involved in every simulation set with differing within-study variances and thus differing weights $w_i = 5, 10$ and 20, so that the total information varies with $\mathrm{TI} = \mathrm{TI}_M = 50, 100$ and 200, respectively. With decreasing weights the within-study variance increases and it becomes more difficult to detect a specific between-study variance σ_B^2. Consequently, the power of the heterogeneity test and the heterogeneity measures decrease with decreasing weights (Figures 1(m), (o), and (q)). If the power and heterogeneity measures are plotted against $\sigma_B^2/\sigma_{W,2}^2$, then nearly identical results for the three different weighting situations (Figures 1(n), (p), and (r)) are observed.

In scenario 4 (Figures 1(s)–(x)), the weights are chosen so that $\sigma_{W,1}^2$ is constant and $\sigma_{W,2}^2$ varies (Table I). That is, $\sigma_{W,2}^2$ increases when the weights of the studies become more and more unbalanced. For a given value of σ_B^2, the increase in $\sigma_{W,2}^2$ leads to decreasing power and decreasing heterogeneity measures (Figures 1(s), (u), and (w)). After rescaling the x-axis to $\sigma_B^2/\sigma_{W,2}^2$, nearly identical results for the three different weighting situations (Figures 1(t), (v) and (x)) are observed.

3.2. Power of heterogeneity test and measures of heterogeneity for normally distributed outcome with additional random variation of the weights

Heterogeneity of treatment effects for normally distributed outcome was simulated as follows: Study treatment effects $\hat{\theta}_i$ were drawn from $\mathrm{N}(\theta_i, v_i)$ where v_i was drawn from $\mathrm{N}(\sigma_W^2, 0.01^2)$.

DOI: 10.1002/sim

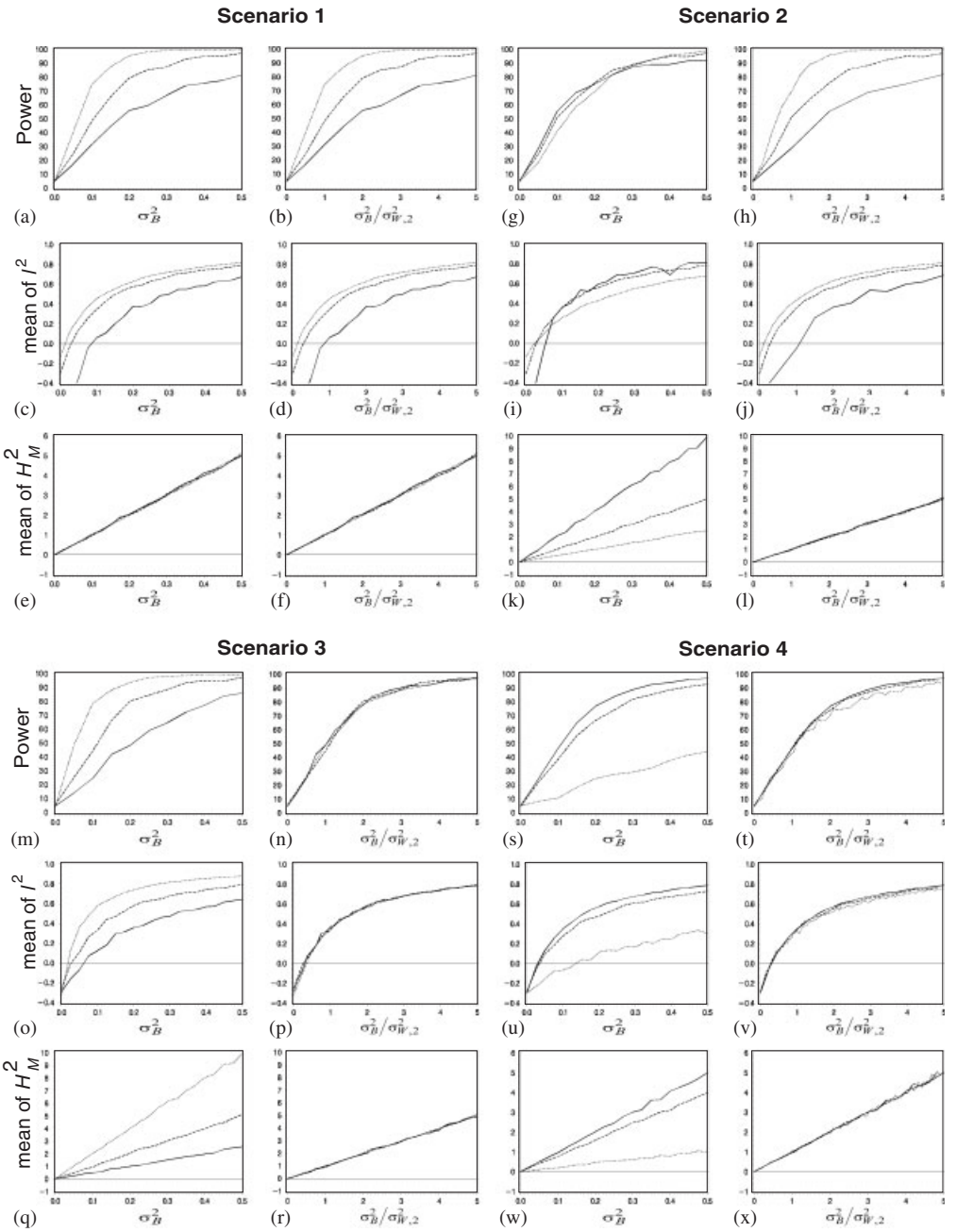

Figure 1. Simulation results of scenarios 1–4: The power of the heterogeneity test and the estimated means of I^2 and H_M^2 are plotted on the y-axes against between-study variance σ_B^2 (1st and 3rd column) and the ratio $\sigma_B^2/\sigma_{W,2}^2$ (2nd and 4th column) on the x-axes. Solid, dashed and dotted lines correspond to the sub-scenarios (a), (b) and (c), respectively, as described in Table I.

DOI: 10.1002/sim

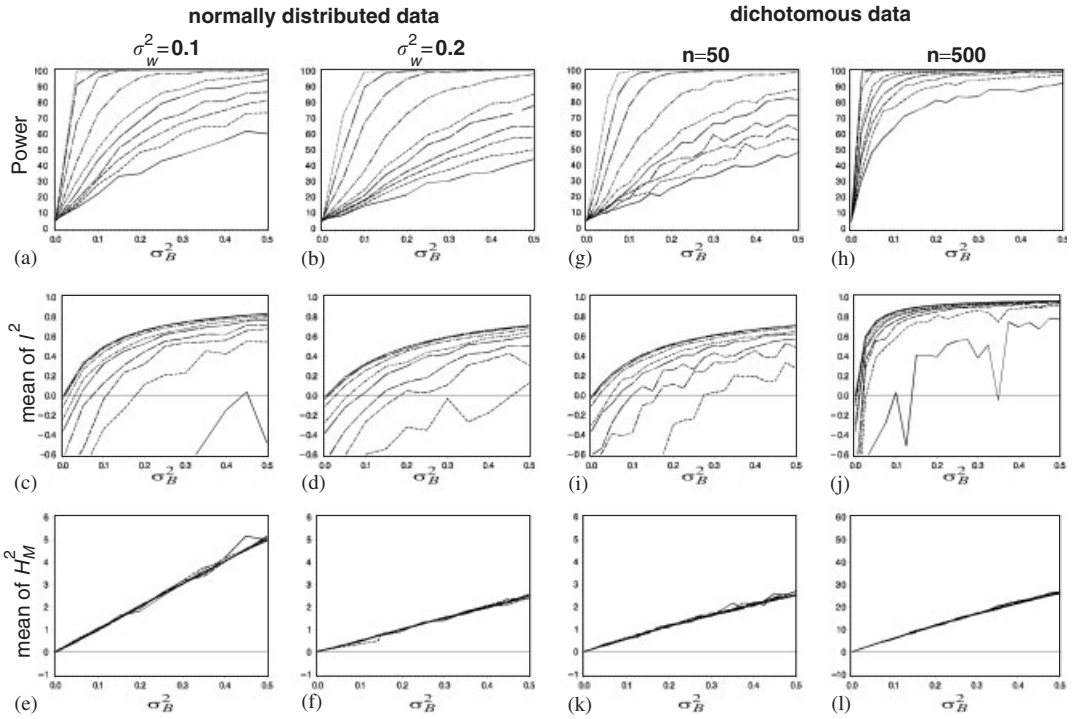

Figure 2. Simulations of normally distributed data with within-study variances of $\sigma_W^2 = 0.1, 0.2$ and dichotomous data with $n = 50, 500$. The number of studies increases from 3 studies (solid line) to 200 studies (dotted line) with $k = 3, 4, 5, 6, 8, 10, 20, 50, 100, 200$. In (d) and (i) all mean values for $k = 3$ are below -0.6 and thus not shown. The power and the means of I^2 and H_M^2 are plotted against σ_B^2.

θ_i itself was drawn from $N(\theta, \sigma_B^2)$ where the true treatment effect θ was assumed to be zero (but this choice is immaterial). Two different mean within-study variances of $\sigma_W^2 = 0.1, 0.2$ were used, σ_B^2 varied from 0 to 0.5 and the number of studies varied from 3 to 200 studies. For every situation 1000 simulation runs were performed and the power and mean and median estimates for I^2 and H_M^2 were calculated.

For normally distributed data, the power to detect heterogeneity increases with increasing number of studies and with decreasing within-study variance σ_W^2 (Figures 2(a) and (b)). Mean I^2 depends on the number of studies in the meta-analysis (Figures 2(c) and (d)). For $k = 3$ studies no expectation of I^2 exists and the computed mean of unrestricted simulated I^2 values hardly becomes positive (e.g. for $\sigma_W^2 = 0.2$ and $k = 3$ all mean estimates of I^2 are less than -0.6 and thus not shown in Figure 2(d)). The median estimates of I^2 are rather constant for constant σ_B^2, except for less than 5 studies in the meta-analysis (not shown). Mean H_M^2 shows no bias (Figures 2(e) and (f)). The median estimates of H_M^2 are smaller for smaller numbers of studies (not shown). This is due to the distribution of H_M^2 which is closely related to a χ^2-distribution. The influence of σ_W^2 on heterogeneity assessment is evident: e.g. for 200 studies included and $\sigma_B^2 = 0.25$ and

DOI: 10.1002/sim

0.5, the measure I^2 reduces to 45 per cent and 71 per cent under $\sigma_W^2 = 0.2$ (Figure 2(d)) when compared to $I^2 = 71$ and 83 per cent (Figure 2(c)) under $\sigma_W^2 = 0.1$, respectively.

3.3. Power of heterogeneity test and measures of heterogeneity for dichotomous outcome

The study effects for dichotomous outcome are usually quantified by odds-ratios ($\exp(\theta)$). In our simulations the true odds-ratio has either a value of one or two and the between study variance σ_B^2 varies from 0 to 0.5 with increments of 0.05. The log-odds ratio for each individual study θ_i was drawn from $N(\theta, \sigma_B^2)$. Assuming a response probability for group 1 (p_1) of 0.5, the response probability for the other group (p_2) was chosen so that an odds ratio of $\exp(\theta_i)$ resulted. Equal group sizes for both groups were used. Then random numbers of events for both groups (a and b) were drawn with equal group sizes ($n = n_1 = n_2$) from $B(n, p_1)$ and $B(n, p_2)$, respectively. The numbers of event-free observations are consequently $c = (n - a)$ and $d = (n - b)$, respectively. In the case of an empty cell 0.5 was added to all four cells. The odds ratio and the variance of the log-odds ratio were then estimated by $\exp(\hat{\theta}_i) = (a_i d_i)/(b_i c_i)$ and $\mathrm{var}(\hat{\theta}_i) = 1/a_i + 1/b_i + 1/c_i + 1/d_i$, respectively, and consequently the weights for the test statistic Q and the heterogeneity measures are $w_i = 1/\mathrm{var}(\hat{\theta}_i)$. The number of studies varied from 3 to 200 studies. For every situation 1000 simulation runs were performed and the power and mean and median estimates for I^2 and H_M^2 were calculated.

The power for the heterogeneity test increases for increasing number of studies and increasing sample size per study (Figures 2(g) and (h)). As no dependence on the true values of the odds ratio has been observed only the results for a true odds ratio of one are shown (Figures 2(g) and (h)). The behaviour of the heterogeneity measures is similar to the corresponding situations of the normally distributed outcome. Again not the absolute value of between-study variance σ_B^2 is important to assess heterogeneity but the between-study variance relative to the within-study variance. This can be seen, e.g. for H_M^2 where its value is multiplied by 10 if the within-study variance σ_W^2 is divided by 10 for $n = 50$ (Figure 2(k)) compared to $n = 500$ (Figure 2(l)).

4. DISCUSSION

The examination of heterogeneity usually begins with a formal statistical test for its presence. For a small number of studies the test has low power so that in the absence of a statistical significant test result the observed amount of heterogeneity should still be described. The test statistic Q itself cannot be used as heterogeneity measure as it increases with the number of studies involved. A proper heterogeneity measure should not depend on the number of studies. The two preferred heterogeneity measures, I^2 and H_M^2, also take the within-study variance into account.

The results of Hardy and Thompson [13] investigated three characteristics of meta-analyses which influence the power of the heterogeneity test: between-study variance, number of included studies and the weights of the studies. These criteria have been reinvestigated with a main focus on the behaviour of heterogeneity measures. The main conclusions are:

- The estimates of the within-study variance $\hat{\sigma}_{W,1}^2$ and $\hat{\sigma}_{W,2}^2$ may differ considerably whenever the weights of the included studies are unequal within meta-analyses. The use of $\hat{\sigma}_{W,2}^2$ better reflects the impact of heterogeneity, either assessed by the heterogeneity test or quantified by heterogeneity measures.

DOI: 10.1002/sim

- The number of studies affects both the power of the heterogeneity test and the heterogeneity measures I^2, but not H_M^2.
- The power of heterogeneity tests and heterogeneity measures increase if the between-study variance increases and/or the within-study variance decreases. Thus the impact of heterogeneity in meta-analyses is mainly determined by the ratio of between- and within-study variances, and not by the between-study variance alone.
- The expectation of H_M^2 is the ratio of between- and within-study variance. Thus H_M^2 is linearly related on average to the increase in the between-study variance for constant within-study variance. H_M^2 has a lower limit of minus one and no finite upper limit, where values higher than zero indicate heterogeneity. Thus $\max(0, H_M^2)$ is used in practice.

  H_M^2 is unbiased, but the median estimates of H_M^2 depend on the number of studies; which is due to the increased skewness of H_M^2 when the number of studies is decreased.
- I^2 is non-linearly monotone related to the ratio of between- and within-study variances. I^2 has no lower limit and an upper limit of 100 per cent. Values higher than zero indicate heterogeneity. In practice $\max(0, I^2)$ is used. The upper limit of 100 per cent can never be achieved. In order to achieve values close to 100 per cent the between-study variance has to be much higher than the within-study variance.

  The expected value of I^2 depends on the number of studies in the meta-analyses but for more than 10 studies this dependence is of minor importance. The median of I^2 is rather independent on the number of studies.
- The definition of total information TI_M is preferred to TI since the within-study variance is estimated by $\sigma_{W,2}^2$. TI_M depends on the number and the precisions (weights) of the studies. Few studies with high weights may have the same total information TI_M as many studies with small weights. TI_M has a direct relationship to the power of the heterogeneity test for a given between-study variance, e.g. the higher TI_M the higher is the power, but it has no direct relationship to the more meaningful heterogeneity measures as both, TI_M and TI depend on the number of studies. Furthermore, the total information measure does not take the ratio of between- and within-study variances into account, which is crucial when assessing the impact of heterogeneity on meta-analysis results.

The two heterogeneity measures I^2 and H_M^2 have a non-linear but monotone relationship. H_M^2 is zero if between- and within-study variances are equal and can be interpreted how many times the between-study variance exceeds the within-study variance. I^2 is scaled between zero and 100 per cent and can be interpreted as the percentage of the total variability due to between-study variance. Both measures have similar properties and are suitable for practical use, except that I^2 may have a considerable dependence on the number of studies k, in particular if k is small. H_M^2 may be more intuitive for some researchers, as it directly gives the additional variance due to between-study variability relative to the within-study variability. I^2 may be preferred due to its scaling to a fixed interval. However, as for any fixed-interval scaled measure, one has to be aware about the interpretation of the scale. If $I^2 = 80$ per cent the between-study variance is 4 times as high as the within-study variance and if $I^2 = 90$ per cent the between-study variance is 9 times as high as the within-study variance, that is, I^2 increases non-linearly with the between-study variance.

Higgins and Thompson [8, 9] classified $I^2 \geqslant 50$ per cent as notable heterogeneity, which would correspond to $H_M^2 \geqslant 1$. In this case the relationship of between- to within-study variance is at least

DOI: 10.1002/sim

fifty–fifty ($\sigma_B^2 \geqslant \sigma_{W,2}^2$). This seems to be a reasonable choice for practical use. Furthermore, they classify mild heterogeneity by $I^2 \leqslant 30$ per cent which corresponds to $H_M^2 \leqslant 0.43$ or $\sigma_B^2 \leqslant 0.43\sigma_{W,2}^2$. However some researchers may even want to use lower limits.

In practice $\max(0, I^2)$ and $\max(0, H_M^2)$ are used, where zero indicates no heterogeneity. However, values of $I^2 < 0$ and $H_M^2 < 0$ may still have a sensible interpretation as they may indicate that studies in the meta-analysis may not be independent [20].

I^2 and H_M^2 measure the impact of heterogeneity [8] as they depend on the number of patients in the individual studies. These patient numbers affect the variance of the estimates and thus the weights of the studies (precisions). Since both I^2 and H_M^2 measure a meta-analysis specific feature (impact) and not a population specific feature (extent) of heterogeneity, their comparison across meta-analyses is questionable.

As a consequence of our results we suggest that meta-analysis software should give by default the estimates for the between- and within-study variances, σ_B^2 and $\sigma_{W,2}^2$, respectively. Both heterogeneity measures, I^2 and H_M^2, should be given either by default or optionally. The total information measure TI_M should be optionally available.

If there is no heterogeneity present the choice between a fixed-effect and a random-effects model is not so crucial, as both models yield results that are rather comparable [4, 21, 22], although the random-effects model is generally conservative. Whatever method is used the analysis of heterogeneity will often be a '*post hoc*' exploratory analysis for a better understanding of the data [4, 21, 23–25]. Thompson [26] describes the distinction between statistical heterogeneity and clinical heterogeneity. He uses the term clinical heterogeneity to refer to differences in the characteristics of the studies, such as their designs and the rates of loss to follow-up, differences in the characteristics of study subjects, such as their mean age and the severity of illness, and differences in the intervention, such as the dose or duration of treatment. Analyses which investigate whether particular variables may explain some of the heterogeneity of results in a meta-analysis are becoming more common [1, 26–32].

The absence of formal explorations of the reasons for heterogeneity, which is not uncommon [5], represents a lost opportunity for enhancing the contribution of the evidence as a whole to scientific understanding [26, 33]. Meta-analysis should not be exclusively used to arrive at an average or 'typical' value for effect size [25] and it should not be seen as a pure statistical method but rather as a multi-component approach for making sense of information [4].

REFERENCES

1. Higgins J, Thompson S, Deeks J, Altmann DG. Statistical heterogeneity in systematic reviews of clinical trials: a critical appraisal of guidelines and practice. *Journal of Health Services Research and Policy* 2002; **7**(1):51–61.
2. Williamson PR, Gamble C. Identification and impact of outcome selection bias. *Statistics in Medicine* 2005; **24**:1547–1561.
3. Terrin N, Schmid CH, Lau J, Olkin I. Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine* 2003; **22**:2113–2126.
4. Petitti DB. *Meta-analysis, Decision Analysis and Cost-Effectiveness Analysis*. Oxford University Press: New York, 2000.
5. Petitti DB. Approaches to heterogeneity in meta-analysis. *Statistics in Medicine* 2001; **20**:3625–3633.
6. Baujat B, Mah C, Pignon J, Hill C. A graphical method for exploring heterogeneity in meta-analyses: application to a meta-analysis of 65 trials. *Statistics in Medicine* 2002; **21**:2641–2652.
7. Schmid CH. Exploring Heterogeneity in randomized trials via meta-analysis. *Drug Information Journal* 1999; **33**:211–224.
8. Higgins J, Thompson S. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* 2002; **21**:1539–1558.

DOI: 10.1002/sim

9. Higgins J, Thompson S, Deeks J, Altmann DG. Measuring inconsistency in meta-analyses. *British Medical Journal* 2003; **327**:557–560.
10. RevMan Analyses [Computer program]. *Version 1.0 for Windows*. Review Manager (RevMan) 4.2. Oxford, England: The Cochrane Collaboration, 2002.
11. Cochrane WG. The combination of estimates from different experiments. *Biometrics* 1954; **10**:101–129.
12. Der Simonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**:177–188.
13. Hardy R, Thompson S. Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine* 1998; **17**:841–856.
14. Everitt BS. *Modern Medical Statistics—A Practical Guide*. Arnold: London, 2003.
15. Takkouche B, CadarsoSurez C, Spiegelman D. Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis. *American Journal of Epidemiology* 1999; **150**:206–215.
16. National Health and Medical Research Council (Australia). *How to Review Evidence*: *Systematic Identification and Review of the Scientific Literature*, Canberra, 2000.
17. Copas JB. Regression, prediction and shrinkage (with Discussion). *Journal of the Royal Statistical Society*, *Series B* 1983; **45**:311–354.
18. Copas JB. Cross-validation shrinkage of regression predictors. *Journal of the Royal Statistical Society*, *Series B* 1987; **49**:175–183.
19. Copas JB. Using regression models for prediction: shrinkage and regression to the mean. *Statistical Methods in Medical Research* 1997; **6**:167–183.
20. Hamer R, Simpson P. SAS® tools for meta-analysis. *SAS SUGI Proceedings*: *Statistics*, *Data Analysis and Data Mining*, SUGI 27. 2002; Paper 250.
21. Greenland S, Salvan A. Bias in the one-step method for pooling study results. *Statistics in Medicine* 1990; **9**:247–252.
22. Thompson S, Pocock S. Can meta-analysis be trusted. *Lancet* 1991; **338**:1127–1130.
23. Greenland S. Quantitative methods in the review of epidemiologic literature. *Epidemiology Reviews* 1987; **9**:1–30.
24. Greenland S, Longnecker M. Methods for trend estimation from summarized dose–response data, with applications to meta-analysis. *American Journal of Epidemiology* 1992; **135**:1301–1309.
25. Jenicek M. Meta-analysis in medicine: where we are and where we want to go. *Journal of Clinical Epidemiology* 1989; **42**:35–44.
26. Thompson S. Why sources of heterogeneity in meta-analysis should be investigated. *British Medical Journal* 1994; **309**:1351–1355.
27. Thompson S, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in Medicine* 1999; **18**:2693–2708.
28. Berlin JA. Benefits of heterogeneity in meta-analysis of data from epidemiologic studies. *American Journal of Epidemiology* 2004; **142**:383–387.
29. Knapp G, Hartung J. Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine* 2003; **22**:2693–2710.
30. Thompson S, Higgins J. How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine* 2002; **21**:1559–1573.
31. Higgins J, Thompson S. Controlling the risk of spurious findings from meta-regression. *Statistics in Medicine* 2004; **23**:1663–1682.
32. Böhning D. Meta-analysis: a unifying meta-likelihood approach framing unobserved heterogeneity, study covariates, publication bias and study quality. *Methods of Information in Medicine* 2005; **44**(1):127–135.
33. Lau J, Ioannidis JPA, Schmid CH. Summing up evidence: one answer is not always enough. *Lancet* 1998; **351**:123–127.

DOI: 10.1002/sim