STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:2688–2699
Published online 22 December 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2481

The power of the standard test for the presence of heterogeneity in meta-analysis

Dan Jackson*,†

Centre for Operational Research and Applied Statistics, University of Salford, Manchester M5 4WT, U.K.

SUMMARY

It has been suggested that the standard test for the presence of heterogeneity in meta-analysis has low power. Although this has been investigated using simulation, there is little direct analytical evidence of the validity of this claim. Using an established approximate distribution for the test statistic, a procedure for obtaining the power of the test is described. From this, a simple formula for the power is obtained. Although this applies to a special case, the formula gives an indication of the power of the test more generally. In particular, for a given significance level, the power can be calibrated in terms of the proportion of the studies' variances that is provided by between-study variation. A consideration of this quantity confirms that the test does, in general, have low power. It is suggested that practitioners, who wish to conduct the standard test, use the ideas provided in order to investigate the operating characteristics of the test prior to performing it. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: meta-analysis; heterogeneity; standard test

1. INTRODUCTION

Meta-analysis, the statistical process of combining the results from separate studies concerned with the same treatment or issue, is becoming increasingly popular in medical applications. Despite this, a major difficulty with this methodology is determining if the studies are homogeneous, or if instead they have different underlying true treatment effects. These differences are usually described using the random effects model. This incorporates a between-study variance, denoted by τ^2, which quantifies the differences in the studies' results that cannot be explained by within-study variation alone. The random effects model assumes that the outcome of the ith study, y_i, is distributed as $y_i \sim \mathrm{N}(\mu, \sigma_i^2 + \tau^2)$, where μ denotes the overall treatment effect and σ_i^2 is the within-study variance of y_i, which is usually assumed fixed and known but is estimated in practice [1, 2]. It is further assumed that the studies are independent. The

*Correspondence to: Dan Jackson, Centre for Operational Research and Applied Statistics, University of Salford, Manchester M5 4WT, U.K.

†E-mail: d.jackson@salford.ac.uk

Received 14 July 2005
Accepted 8 November 2005

parameter τ^2 can be estimated using maximum likelihood [1, 2] or the method suggested by DerSimonian and Laird [3]. Once $\hat{\tau}^2$ has been obtained, the standard methodology described in detail by Sutton *et al.* [4, Chapter 5], can be used to make inferences concerning the treatment effect.

If τ^2 is assumed to be zero then we have that $y_i \sim \mathrm{N}(\mu, \sigma_i^2)$ and a fixed effects model is adopted. The resulting inferences are then simplified, as no between-study variation means that all studies have exactly the same underlying treatment effect. Hence some may consider the possibility of initially testing the null hypothesis that $\tau^2 = 0$ and, if this hypothesis is accepted, assuming that τ^2 is indeed zero. This 'fixed effects, if we can get away with it' approach is open to criticism and is generally discouraged. For example, Hardy and Thompson [5] argue that the standard test for heterogeneity should not provide the sole determinant of model choice. Methods that quantify the impact of heterogeneity, rather than test for the presence of this, have also been developed [6]. A truly random effects approach of estimating τ^2, which simplifies to a fixed effects model only if $\hat{\tau}^2 = 0$, may therefore be preferable.

A further issue is the alleged low power of tests that $\tau^2 = 0$, although it has also been noted that the standard test has excessive power to detect unimportant heterogeneity when there are many studies [6]. If these tests do indeed have low power for the types of meta-analyses generally encountered in practice, then failing to detect heterogeneity can hardly be considered satisfactory evidence that this is not present. Hence the power of the most frequently used, indeed standard [4], test for the presence of heterogeneity is of interest to all who adopt standard methodologies, irrespective of their fixed or random effects preferences. See Reference [4, Chapter 3], for a full discussion of this contentious issue and a wide range of references which provide often contrasting opinions concerning this unresolved debate.

Rather than add to this discussion, in this paper the power of the standard test for the presence of heterogeneity will be investigated analytically. This provides a marked departure from earlier papers which use computer simulation to explore the power of this and similar tests [5, 7–9]. The analytical approach adopted here seems preferable, as it is possible to see how the various factors interact to provide the power. The rest of the paper is set out as follows. In Section 2 the standard test is described and in Section 3 an approximation of the test statistic is derived and used to obtain the power. This is applied to an example involving the use of glycerol for patients who suffer an acute stroke in Section 4. Section 5 provides a means to assess the power of the test more generally and Section 6, the discussion, concludes the paper.

2. THE STANDARD TEST FOR HETEROGENEITY

Under the assumption that $\tau^2 = 0$, so that the fixed effects model described above applies, the statistic

$$Q = \sum_{i=1}^{n} w_i (y_i - \bar{y})^2$$

is distributed as a χ^2 distribution with $n-1$ degrees of freedom [4], where n is the number of studies, $w_i = \sigma_i^{-2}$, the standard but not the only choice of weights in a fixed effects meta-analysis, and $\bar{y} = \sum_{i=1}^{n} w_i y_i / \sum_{i=1}^{n} w_i$. Hence the standard test involves computing Q and comparing this to $\chi^2_{n-1;1-\alpha}$, where α is the significance level of the test. If $Q > \chi^2_{n-1;1-\alpha}$ then the null hypothesis that $\tau^2 = 0$ is rejected. Otherwise, we accept this hypothesis. The standard test is therefore easy to conduct, which may go some way to explain its popularity.

3. THE POWER OF THE STANDARD TEST

In order to obtain confidence intervals for the DerSimonian and Laird [3] estimate of τ^2, Biggerstaff and Tweedie [1] use a gamma distribution to approximate the distribution of Q. Under the assumptions of the random effects model, Biggerstaff and Tweedie obtain

$$\mathrm{E}[Q] = (n-1) + \left(S_1 - \frac{S_2}{S_1}\right)\tau^2$$

and

$$\mathrm{Var}[Q] = 2(n-1) + 4\left(S_1 - \frac{S_2}{S_1}\right)\tau^2 + 2\left(S_2 - 2\frac{S_3}{S_1} + \frac{S_2^2}{S_1^2}\right)\tau^4$$

where $S_r = \sum_{i=1}^{n} w_i^r$. Using these results, Biggerstaff and Tweedie approximate the distribution of Q as a gamma distribution, with parameters r and λ, with these same moments. Solving $\mathrm{E}[Q] = r/\lambda$ and $\mathrm{Var}[Q] = r/\lambda^2$, and emphasizing the dependence on τ^2, provides

$$r(\tau^2) = \frac{(\mathrm{E}[Q])^2}{\mathrm{Var}[Q]}$$

and

$$\lambda(\tau^2) = \frac{\mathrm{E}[Q]}{\mathrm{Var}[Q]}$$

As Biggerstaff and Tweedie point out, if $\tau^2 = 0$ this approximation appropriately simplifies to χ^2 with $n-1$ degrees of freedom. As noted above, Q is compared to $\chi^2_{n-1;1-\alpha}$. The power of the test is therefore given by $P(Q) > \chi^2_{n-1;1-\alpha}$. Using the approximation suggested by Biggerstaff and Tweedie, the approximate power of the test, $\beta(\tau^2, \alpha)$, is given by

$$\beta(\tau^2, \alpha) = \frac{(\lambda(\tau^2))^{r(\tau^2)}}{\Gamma(r(\tau^2))} \int_{\chi^2_{n-1;1-\alpha}}^{\infty} x^{r(\tau^2)-1} \mathrm{e}^{-\lambda(\tau^2)x}\, \mathrm{d}x \qquad (1)$$

The approximate power of the standard test can therefore be evaluated numerically. Hardy and Thompson [5] investigate the power of the test using simulation (assuming that $\mu = 5$ although, as they point out, this choice is immaterial). Using (1), instead of simulation, produces results that are highly consistent with those of Hardy and Thompson, as shown in their Figures 1–4. Both methods involve some approximation to the true power of the test: those obtained from (1) involve an approximation of the true distribution of Q, while those of Hardy and Thompson are subject to Monte Carlo error. In practical terms, either approach is appropriate for evaluating the power of the test for a particular data set. An advantage of (1) is that, as an analytical result, this provides the means to produce a formula in order to give a general guide to the power of the test, as shown in Section 5.

4. EXAMPLE: GLYCEROL FOR ACUTE STROKE

This example is highly suitable for illustrating the results. Briefly, it involves the use of glycerol for preventing death in patients who suffer an acute stoke, and was obtained from the Cochrane Collaboration. It involves nine studies with the results summarized as two by two tables. The data are shown in Table I, where the y_i and σ_i^2 are the studies' log odds ratios and within-study variances, obtained after adding halves to all table entries in the usual way [4]. As used here, a negative log odds ratio indicates that the treatment is beneficial. The data provides $Q = 10.5$, which is a fairly large but statistically insignificant value at conventional levels of significance. Given the small number of studies and the alleged low power of the test, it may be that this is hopelessly underpowered to detect any between-study variation. As discussed in the previous section, either (1) or the simulation approach of Hardy and Thompson are suitable for determining the power of the test. Results of these procedures for the glycerol data, using a standard significance level of $\alpha = 0.05$, are shown in Figure 1 assuming that $\tau^2 = 0, 0.1, 0.2, \ldots, 0.9, 1$. For the simulations, 10 000 data sets were produced and the observed proportion of significant simulated meta-analyses provides the power. Lines are shown connecting the points to aid interpretation. Evident from this figure is that very similar results are obtained using these two methods; for many values of τ^2 these are so similar that the points shown in Figure 1 lie on top of each other and it is impossible to distinguish between the two sets of results. Since they are subject to very different types of errors, producing results in both ways is a practical way to ascertain the accuracy of the powers obtained.

Also apparent is the low power of the test. For example, Higgins and Thompson [6] suggest that heterogeneity be described as notable when this provides substantially more than 50 per cent of the studies' variances. For the median within-study variance of 0.24, $\tau^2 = 0.72$ provides sufficient heterogeneity to account for 75 per cent of the variation, which can therefore be described as really rather severe. Even with this very large value of τ^2 the power, although

Table I. The glycerol data. Tabulated values of y_i and σ_i^2 denote the log odds ratio and within-study variance of the ith study, respectively.

Study (i)	y_i	σ_i^2
1	0.31	0.54
2	−0.57	0.17
3	0.01	0.62
4	0.38	0.24
5	0.21	0.39
6	−1.11	0.16
7	1.26	2.77
8	−0.20	0.09
9	0.36	0.23

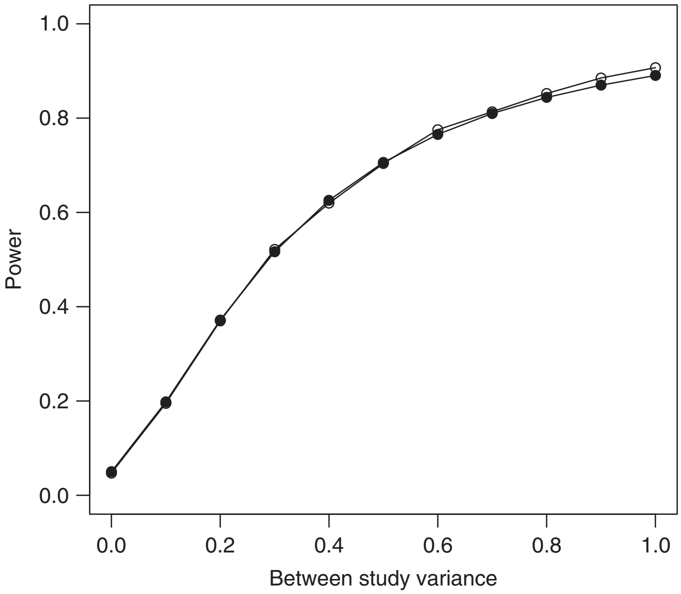

Figure 1. The approximate power of the test for the glycerol data using $\alpha = 0.05$. Solid points indicate the results obtained analytically from equation (1), hollow points denote those obtained from simulation. The results are so similar that some hollow points are not visible, hidden by the corresponding solid points.

quite high at around 0.8, is not particularly impressive. Using the sample median as a representative within-study variance seems preferable to the mean, due to the very large $\sigma_7^2 = 2.77$ and is similar to Higgins and Thompson's typical within-study variance (their equation (9)), which for the glycerol data is 0.25.

This suggests that a larger significance level is required in order to provide more power. A suitable level which, of course, should be selected prior to performing the test can be investigated. From (1), this is a function of τ^2 and α and a contour plot of the power for the glycerol data, obtained directly from (1), is shown in Figure 2.

Figure 2 also illustrates the low power of the test for $\alpha = 0.05$. However the steepness of the contours at $\alpha = 0.05$, through to around $\alpha = 0.15$, indicates that increasing the significance level slightly improves the power quite substantially for this example. Once α becomes around 0.15, further increases do not have the same impact on the power as when α is smaller. It is therefore harder to justify much larger values of α than this on the grounds of increased power.

It seems a reasonable recommendation that practitioners produce a plot akin to Figure 2 prior to performing the standard test so that a suitable significance level can be chosen. A difficulty is that every meta-analysis has a different distribution of within-study variances. Hence a new plot needs to be produced for every meta-analysis. In order to reduce this need, the next section provides a formula suitable for assessing the power more generally.

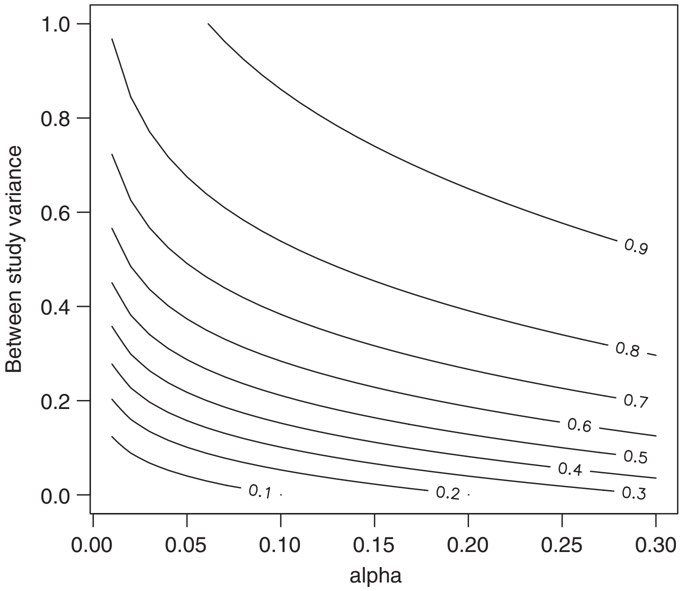

Figure 2. The power of the standard test for the glycerol data in terms of the variance τ^2 and the significance level α, obtained approximately from equation (1).

5. A SIMPLE FORMULA FOR ASSESSING THE POWER OF THE STANDARD TEST

Although the use of equation (1) in the previous sections is helpful, and supports the results obtained by running simulations, the difficulty is that each data set needs to be considered separately and it is hard to make any general statements directly from this. The intention in this section is therefore to obtain a simple formula in order to give an indication of the power of the test more generally. In addition to this motivation, there are three further reasons for developing the formula. Firstly, equation (1) is too complicated to offer any great insight, but the simplification suggested below shows more clearly how the various factors influence the power of the test, and how these interact with each other. Secondly, (1) requires numerical integration or the use of a statistical computing package in order to obtain the cumulative gamma distribution. By comparison, the formula suggested below can readily be evaluated using a handheld calculator for typical meta-analyses with small numbers of studies. Finally, by considering the special case below, the approximate power provided by (1) becomes exact, and there is no longer any concern regarding the magnitude of the error due to utilizing the approximation suggested by Biggerstaff and Tweedie.

In order to simplify the situation, let us assume that all studies are the same size, i.e. $\sigma_i^2 = \sigma^2$ for all i. Let $w = \sigma^{-2}$. Using the gamma approximation suggested by Biggerstaff and Tweedie under this assumption provides

$$r(\tau^2) = r = \frac{n-1}{2}$$

and

$$\lambda(\tau^2) = \frac{1}{2(1 + w\tau^2)}$$

As alluded to above, when all the studies are the same size, this approximation is exact. This is because the simplification means that the y_i are assumed to be i.i.d. from a normal distribution. Using a standard result concerning the sample variance under this assumption (see Reference [10, Theorem 5.3.1, part c], for example) provides $Q \sim (1+w\tau^2)\chi^2_{n-1}$, or equivalently that Q is distributed as a gamma distribution with the same parameter values provided by the approximation. Hence the approximation suggested by Biggerstaff and Tweedie appropriately simplifies to the correct distribution if all the studies are the same size.

This simplification also means that r is no longer a function of τ^2 and, furthermore, if n is odd then the first parameter of the gamma approximation is an integer. There is an interesting relationship between the gamma and Poisson distributions (see Reference [10, p. 100], for example). As parameterized here, if $X \sim \text{gamma}(r, \lambda)$, where r is an integer, then $P(X \leqslant x) = P(Y \geqslant r)$, where $Y \sim \text{Poisson}(\lambda x)$. This means that, under the assumption that the odd number of studies are all the same size, the approximate power in (1) is exact and is also given by

$$\beta(\tau^2, \alpha) = \exp\left(\frac{-\chi^2_{n-1;1-\alpha}}{2(1 + w\tau^2)}\right) \sum_{i=0}^{(n-3)/2} \left(\frac{\chi^2_{n-1;1-\alpha}}{2(1 + w\tau^2)}\right)^i \bigg/ i! \qquad (2)$$

Formula (2) has many desirable features. In particular, it shows how the power depends directly on the quantity $(\chi^2_{n-1;1-\alpha}/(1 + w\tau^2))$, double the parameter value of the resulting Poisson distribution. This type of dependence is not at all evident from (1), even after making the simplifying assumptions. It is interesting that the power depends on the product of w and τ^2, rather than on these two variables separately. Hence the power of the test to detect a between-study variance depends on the magnitude of τ^2 relative to the size of the studies.

For any meta-analysis, with a given odd number of studies of a particular size, (2) can be used to obtain the power of the test. There are two issues raised by the simplification. Firstly, a meta-analysis may well have an even number of studies. If this is the case, then (2) can be calculated for the neighbouring odd numbers and linear interpolation may be used; alternatively (1) could be used directly. Secondly, real meta-analyses do not have studies of exactly the same size. To give an indication of the power of the test for meta-analyses more generally, (2) can be used with $w = \bar{w}$, or some other representative value of w. Although this is a further approximation to any data set in question, this gives an indication of the power of the test for meta-analyses involving the same number of studies, and which are similar in size, as the observed meta-analysis. Hence this gives pertinent information concerning the power of the test and gives an indication of this more generally. It should be noted however that Hardy and Thompson [5] find that tests involving meta-analyses with a single very large study have low power compared to those with a more even distribution of study sizes. This should therefore only be used as a rough guide for any particular example. Other possibilities for a representative value of w include the reciprocal of Higgins and Thompson's [6] typical within-study variance and the sample median. Both of these alternatives produce very similar results to those obtained using $w = \bar{w} = 4.23$ for the glycerol data. Using this value of

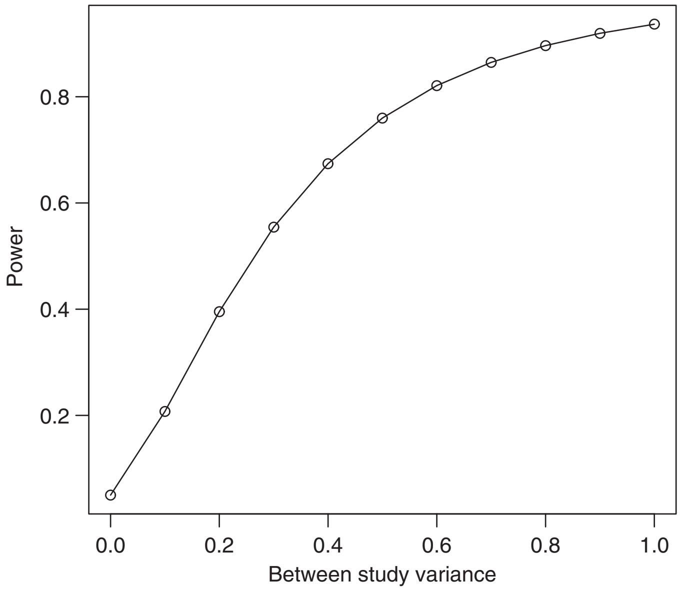

Figure 3. The power of the test for the glycerol data, using (2) with the approximation that $w_i = \bar{w}$ for all i, and $\alpha = 0.05$.

w and $n = 9$ in (2), with $\alpha = 0.05$, produces the powers shown in Figure 3, which are similar to those in Figure 1.

Examining the power of the test for the glycerol data further, producing a contour plot in the same way as in Figure 2, using equation (2) with $w = \bar{w}$, results in Figure 4. This and other examples indicate that equation (2) can be used as an approximation, with a suitable value of w, provided the study sizes are not too disparate.

As noted above, a main advantage of (2) is that this clearly demonstrates the dependence of the power on the value $\theta = (\chi^2_{n-1;1-\alpha}/(1 + w\tau^2))$. Hence this value and n alone can be used to give an indication of the power of the test more generally. The power in terms of these values is shown in Figure 5. Here (2) was evaluated for a grid of odd n and values of θ in order to produce a matrix of powers from which the contour plot was produced. Contours of 0.1, 0.2 (low power), 0.5 (moderate power), 0.8 and 0.9 (high power) have been used to show a wide range of possibilities. This plot can be used to quickly ascertain a rough guide to the power of the standard test for a particular meta-analysis.

As an example of using Figure 5 in practice, let us continue to use the glycerol data. For this example, $w = \bar{w} = 4.23$, suggesting that $\tau^2 = 1/4.32$ is broadly in line with the size of the within-study variances. We might reasonably require that the power of the test is at least 0.5 for such a value of τ^2. The contour providing a power of 0.5 in Figure 5 indicates that $\theta = 7$ is small enough to provide such a power for $n = 9$. Substituting $\theta = 7$ and $w\tau^2 = 1$ into the definition of θ provides $\chi^2_{8;1-\alpha} = 14$, and therefore $\alpha = 0.08$. Similarly, we might require a higher power, around 0.8 say, for a larger τ^2 such as 0.5. Again from Figure 5, this requires θ of around 4, providing $\chi^2_{8;1-\alpha} = 12.46$ and therefore that $\alpha = 0.14$. These arguments support the use of $\alpha = 0.1$ for meta-analyses similar to this one, as sometimes suggested [4].

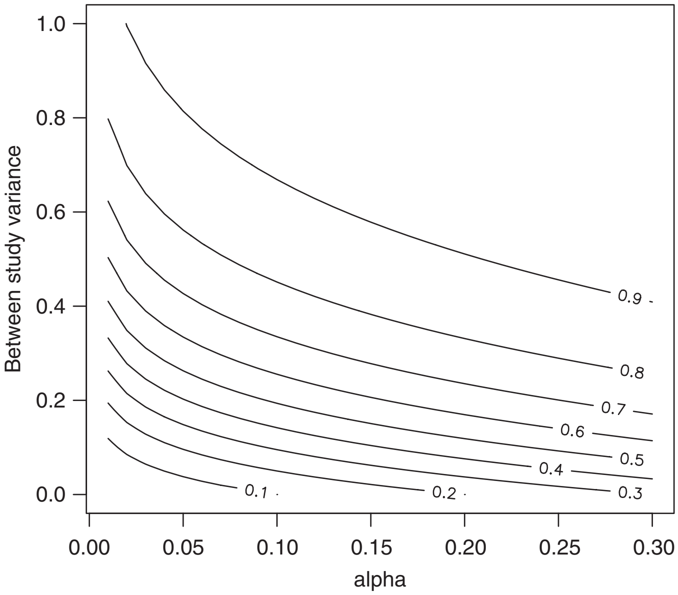

Figure 4. The power of the standard test for the glycerol data in terms of the variance τ^2 and the significance level α, obtained approximately from equation (2).

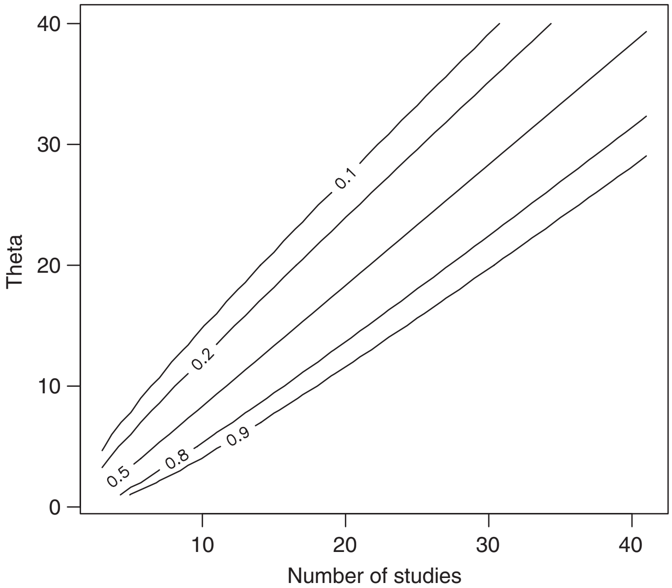

Figure 5. The power of the standard test in terms of $\theta = (\chi^2_{n-1;1-\alpha}/(1 + w\tau^2))$ and the number of studies, assuming all studies are the same size.

Interesting to note is that the curves in Figure 5 rise sharply as n increases. This suggests that for larger data sets the standard test may possess reasonable power and the more conventional $\alpha = 0.05$ might be appropriate. This is investigated further in the next subsection. Using equation (2) or Figure 5, coupled with the types of argument above, can indicate suitable values of α enabling practitioners to adopt an appropriate significance level. If those using this approach are concerned that (2) provides a poor approximation as the studies differ greatly in size then (1) could also be used. Hence the paper provides sufficient ideas to enable those conducting the standard test to adopt a suitable value of α in all circumstances.

5.1. Investigating the power of the test using $\alpha = 0.05$ and $\alpha = 0.1$

Although Figure 5 gives an indication of the values of θ needed to provide a variety of powers for a range of sizes of meta-analyses, this does not afford much insight into the power of the test when using conventional levels of significance. For a meta-analysis with a given odd number of studies of the same size, the fact that the power depends only on θ is very convenient for the purposes of producing this figure and providing a relatively simple result, but this does not distinguish between the roles played by τ^2 and α, respectively.

It is therefore of interest to obtain the value of θ required to give a particular power, for an odd value of n, numerically from (2). Assuming the use of a particular significance level, the required value of $w\tau^2$ can then be obtained directly from the definition of θ. From this, the proportion of the studies' variances that is provided by between-study variation can be calculated as $w\tau^2/(1 + w\tau^2)$. This is the quantity measured by the I^2 statistic, suggested by Higgins and Thompson [6], and provides a suitably interpretable value. Hence we will define $I^2 = w\tau^2/(1 + w\tau^2)$. The values of I^2 required to provide the same range of powers as in Figure 5, for a wide range of odd numbers of studies, are shown in Tables II and III for $\alpha = 0.05$ and 0.1, respectively. In these tables, I^2_β denotes the value of I^2 required to provide a power of β. For example, the fourth column of Table II shows the values of I^2 that provide a moderate power of 0.5 when using $\alpha = 0.05$. For Table III, and therefore $\alpha = 0.1$, the column $I^2_{0.1}$ is omitted as all entries are zero.

From Table II it can be seen that, using $\alpha = 0.05$, the power of the test is indeed quite low unless there is a large number of studies. For example, if there are 15 studies, $I^2 = 0.6$ is required to achieve the reasonably high power of 0.8. As noted in Section 4, Higgins and Thompson suggest that notable heterogeneity be described as providing considerably more

Table II. The power of the standard test for $\alpha = 0.05$, assuming all studies are the same size. The tabulated I^2_β denote the values of I^2 required to provide a power of β.

Number of studies	$I^2_{0.1}$	$I^2_{0.2}$	$I^2_{0.5}$	$I^2_{0.8}$	$I^2_{0.9}$
5	0.18	0.37	0.65	0.83	0.89
15	0.11	0.23	0.44	0.60	0.67
25	0.09	0.19	0.36	0.50	0.57
35	0.08	0.16	0.31	0.45	0.51
45	0.07	0.15	0.28	0.41	0.46
55	0.06	0.13	0.26	0.37	0.43
65	0.06	0.12	0.24	0.35	0.40

Table III. The power of the standard test for $\alpha = 0.1$, assuming all studies are the same size. The tabulated I_β^2 denote the values of I^2 required to provide a power of β.

Number of studies	$I_{0.2}^2$	$I_{0.5}^2$	$I_{0.8}^2$	$I_{0.9}^2$
5	0.23	0.57	0.79	0.86
15	0.14	0.37	0.55	0.63
25	0.11	0.30	0.46	0.53
35	0.09	0.26	0.40	0.47
45	0.08	0.23	0.36	0.42
55	0.08	0.21	0.33	0.39
65	0.07	0.20	0.31	0.37

than 50 per cent of the studies' variances, so that $I^2 = 0.6$ can be interpreted as being quite large. Hence really rather severe heterogeneity is needed to achieve high power when there are 15 studies present in the meta-analysis. Higgins and Thompson also suggest that mild heterogeneity be described as providing less than 30 per cent of the studies' variances. Table II shows that to achieve a power of 0.8 for such mild heterogeneity more than 65 studies are required.

Table III shows an inevitable increase in power when using $\alpha = 0.1$ but similar arguments as in the previous paragraph lead to the conclusion that the test can hardly be considered high powered even when using this value. The claim that the standard test has low power seems to be entirely justified by these findings.

6. CONCLUSIONS

The standard test for the presence of heterogeneity has frequently been said to have low power. This claim has previously been supported by the results from simulations. The main contribution of this paper has been to provide power formulae, which are an alternative to simulations for assessing this.

Although previous findings are of considerable interest, simulation studies can only ever consider a selection of special cases from an infinite range of possibilities. However, by examining the special case of there being an odd number of studies with the same within-study variance, a simple formula for the power has been obtained. This can be used to give an indication of the power of the test for a particular data set, assuming that the study sizes are not too disparate, although a more careful analysis for a particular example can also be performed as shown in Sections 3 and 4. It seems a reasonable suggestion that those conducting the test, irrespective of their fixed or random effects preferences, use these techniques to investigate the power of the test prior to choosing a significance level, so that they can adopt a value best suited for their purpose. This can be done initially using (2), and possibly Figure 5, and then using equation (1) (possibly with simulation to support this) if a more detailed investigation is required.

Perhaps more importantly, the results provide an indication of the power of the test more generally, and in terms of the proportion of the studies' variances that is provided by between-study variation. A consideration of this readily interpretable quantity provides further evidence of the low power of the test. This indicates that adopting a significance level of 0.1, as sometimes suggested, appears justified, although this also provides a test that appears to be fairly low powered. The results show that the test is only high powered when there are large numbers of big studies, which is not a common occurrence in meta-analysis. Note that all the results obtained are based on the assumption that the random effects model described in Section 2 is appropriate, which requires sufficiently large studies so that the σ_i^2 and hence the w_i can be reasonably regarded as known. Although the findings indicate low power unless the studies are large, the random effects model and therefore the results are less valid for meta-analyses with many small studies and this should not be forgotten when interpreting the findings.

Finally, it should be noted that this paper has only considered the standard test. Many others are possible (see Reference [4], for example, pp. 39–40) and no attempt has been made to assess these. It should therefore not be inferred that the standard test is inferior to others on the basis of this research. These other tests for the presence of heterogeneity also make modelling assumptions, which are invalid in more complicated settings where there are covariates or publication bias, where studies with more optimistic or interesting results are more likely to be present in the meta-analysis. The implications of such phenomena for the standard test are currently being investigated.

REFERENCES

1. Biggerstaff BJ, Tweedie RL. Incorporating variability of estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in Medicine* 1997; **16**:753–768.
2. Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. *Statistics in Medicine* 1996; **15**:619–629.
3. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**:177–188.
4. Sutton AJ, Abrams KR, Jones DR, Sheldon DR, Song F. *Methods for Meta-analysis in Medical Research*. Wiley: New York, 2002.
5. Hardy RJ, Thompson SG. Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine* 1998; **17**:841–856.
6. Higgins JPT, Thompson SG. Quantifying heterogeneity in meta-analysis. *Statistics in Medicine* 2002; **21**:1539–1558.
7. Jones MP, O'Gorman TW, Lemke JH, Woolson RF. A Monte Carlo investigation of homogeneity tests of the odds ratio under various sample size configurations. *Biometrics* 1989; **45**:171–181.
8. Paul SR, Donner A. Small sample performance of tests of homogeneity of odds ratios in K 2 × 2 tables. *Statistics in Medicine* 1992; **11**:159–165.
9. Paul SR, Donner A. A comparison of tests of homogeneity of odds ratios in K 2 × 2 tables. *Statistics in Medicine* 1989; **8**:1455–1468.
10. Casella G, Berger RL. *Statistical Inference* (2nd edn). Duxbury: North Scituate, MA, 2002.