STATISTICS IN MEDICINE
Statist. Med. 2002; **21**:1539–1558 (DOI: 10.1002/sim.1186)

Quantifying heterogeneity in a meta-analysis

Julian P. T. Higgins[*,†] and Simon G. Thompson

MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR, U.K.

SUMMARY

The extent of heterogeneity in a meta-analysis partly determines the difficulty in drawing overall conclusions. This extent may be measured by estimating a between-study variance, but interpretation is then specific to a particular treatment effect metric. A test for the existence of heterogeneity exists, but depends on the number of studies in the meta-analysis. We develop measures of the impact of heterogeneity on a meta-analysis, from mathematical criteria, that are independent of the number of studies and the treatment effect metric. We derive and propose three suitable statistics: H is the square root of the χ^2 heterogeneity statistic divided by its degrees of freedom; R is the ratio of the standard error of the underlying mean from a random effects meta-analysis to the standard error of a fixed effect meta-analytic estimate, and I^2 is a transformation of H that describes the proportion of total variation in study estimates that is due to heterogeneity. We discuss interpretation, interval estimates and other properties of these measures and examine them in five example data sets showing different amounts of heterogeneity. We conclude that H and I^2, which can usually be calculated for published meta-analyses, are particularly useful summaries of the impact of heterogeneity. One or both should be presented in published meta-analyses in preference to the test for heterogeneity. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: meta-analysis; heterogeneity

1. INTRODUCTION

A systematic review of studies addressing a common question will inevitably bring together material with an element of diversity. Studies will differ in design and conduct as well as in participants, interventions, exposures or outcomes studied. Such diversity is commonly referred to as methodological or clinical heterogeneity, and may or may not be responsible for observed discrepancies in the results of the studies. Statistical heterogeneity exists when the true effects being evaluated differ between studies, and may be detectable if the variation between the results of the studies is above that expected by chance.

[*] Correspondence to: Julian Higgins, MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR, U.K.

[†] E-mail: julian.higgins@mrc-bsu.cam.ac.uk

Contract/grant sponsor: MRC; contract/grant number: G9815466

Addressing statistical heterogeneity (referred to simply as heterogeneity in this paper) is one of the most troublesome aspects of many systematic reviews. The interpretative problems depend on how substantial the heterogeneity is, since this determines the extent to which it might influence the conclusions of the meta-analysis. It is therefore important to be able to quantify the extent of heterogeneity among a collection of studies. An obvious means of achieving this is by estimating the between-study variance of the parameters of interest. This is done as part of a random effects meta-analysis [1]. The variance can be used to describe the extent of variability in effect across studies, for example, as a range of odds ratios or risk ratios. However, such a measure does not facilitate comparisons of heterogeneity across meta-analyses of different types of outcomes, such as dichotomous and continuous outcomes. Further, interpretation of this estimate in isolation can be difficult, since it is specific to the chosen measure of effect in the meta-analysis. For example, in clinical trials a common measure of treatment effect for dichotomous outcome data is the odds ratio. The extent of heterogeneity is quantified on the scale of the log-odds ratio, an unintuitive scale to most.

A more common way of indicating the extent of heterogeneity is a statistical test, often described as Cochran's χ^2 test or the Q-test [2, 3]. A p-value is frequently quoted as an indication of the extent of between-study variability. It is widely appreciated that the test has poor power in the common situation of few studies, and excessive power to detect clinically unimportant heterogeneity when there are many studies [4]. The test does not therefore provide a relevant summary of the extent to which heterogeneity impacts on the meta-analysis.

We here aim to develop measures of the extent of heterogeneity in a meta-analysis that overcome the shortcomings of existing measures. Our focus is on the impact of heterogeneity on the results of a meta-analysis and therefore, more loosely, on the degree to which conclusions might be generalized to situations outside those investigated in the studies at hand. We desire measures which are easily interpretable by non-statisticians and which do not intrinsically depend on the number of studies or type of outcome data. A measure with such properties could have many useful applications. Apart from our primary motivation of developing a simple, universal statistic that summarizes the impact of heterogeneity in a wide range of meta-analyses, it may enable quantification of how much heterogeneity can be accounted for by study-level covariates, or by particularly influential studies.

In the following section we introduce some motivating examples illustrating a variety of meta-analytic situations. In Section 3 we then present our desirable properties formally, and develop potential measures of the impact of heterogeneity for the mathematically tractable special case of studies which have equally precise estimates. We explore properties and interpretations of those that show promise in the general case in Section 4, and propose three measures for general use within systematic reviews and apply them to our example data sets. Finally, in Section 5 we discuss the strengths and limitations of our proposals.

2. MOTIVATING EXAMPLES

We introduce five data sets from systematic reviews of clinical trials, to which we apply our methods in Section 4. They have been chosen to provide a range of meta-analyses with regard to numbers of studies, measures of treatment effect and extent of heterogeneity.

2.1. Homogeneous set of trials: Human albumin (Figure 1(a))

A systematic review collated randomized controlled trials of human albumin solution for resuscitation and volume expansion in critically ill patients [5]. An overall detrimental effect

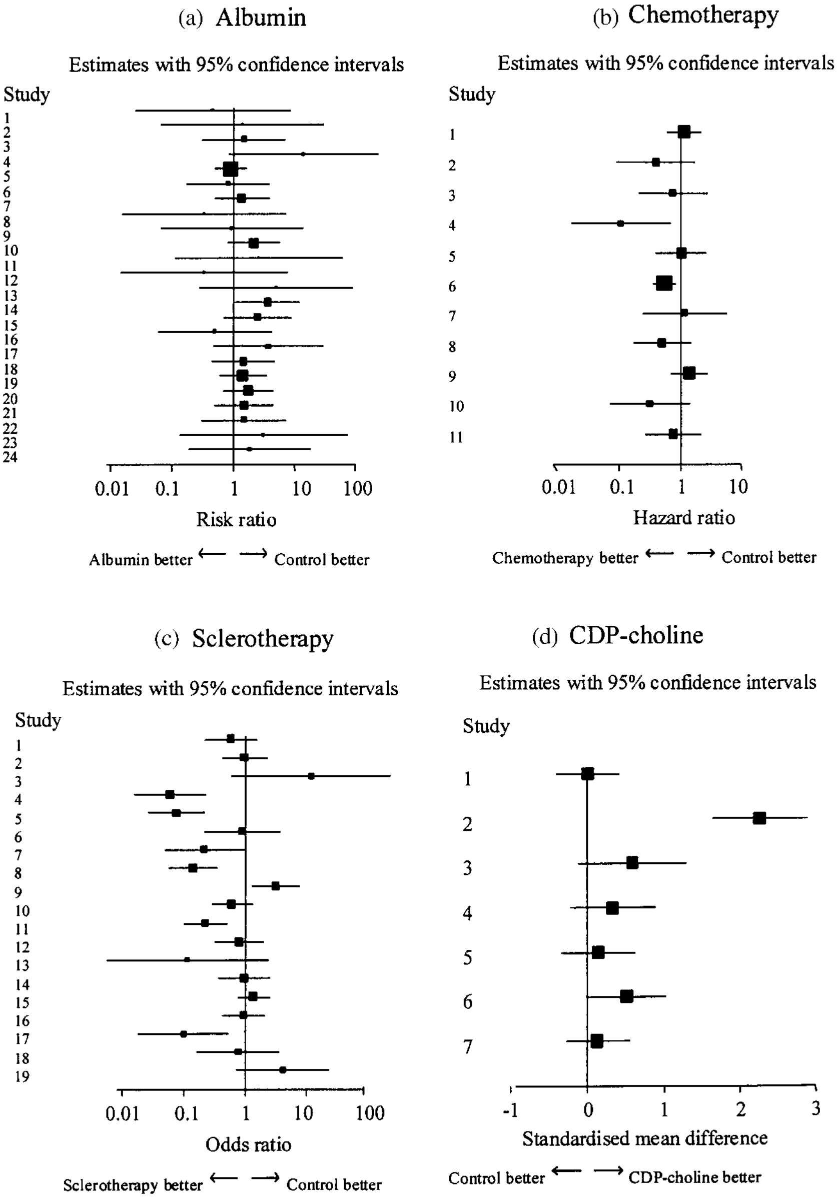

Figure 1. Confidence interval plots for four example data sets: (*a*) 24 trials of albumin versus placebo [5]; (*b*) 11 trials of adjuvant chemotherapy [7]; (*c*) 19 trials of sclerotherapy versus control [8]; (*d*) 7 trials of CDP-choline versus control [9]; (*e*) 3 trials of gamma nail versus sliding hip screws [10].

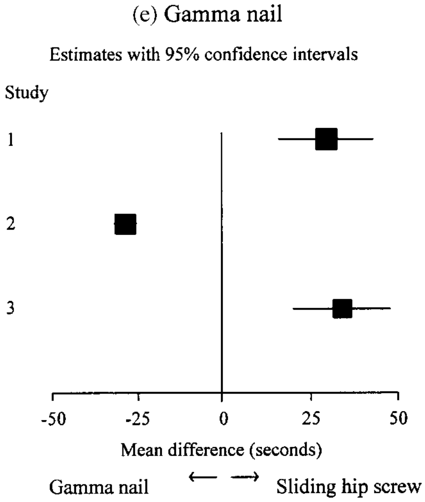

Figure 1. *Continued.*

of albumin on mortality was apparent, in part due to a striking lack of heterogeneity in relative risks between the trials, despite a degree of clinical diversity that led to controversy [6]. Figure 1(*a*) illustrates the relative risks of death on albumin relative to placebo for all 24 trials included in the meta-analysis.

2.2. Moderate heterogeneity: Adjuvant chemotherapy (Figure 1(b))

Our second example is a meta-analysis based on individual patient data undertaken by the Sarcoma Meta-analysis Collaboration [7]. Fourteen trials of adjuvant chemotherapy for localized resectable soft-tissue sarcoma of adults were identified. We address the outcome of local recurrence-free interval, for which 11 trials contributed data. This is a time-to-event outcome, and the treatment and control groups were compared using hazard ratios. There was no statistically significant heterogeneity, but the results were not as apparently homogeneous as the albumin trials.

2.3. Heterogeneous set of trials: Sclerotherapy (Figure 1(c))

A meta-analysis was undertaken of 19 trials of sclerotherapy versus control treatment for the prevention of first bleeding in cirrhosis [8]. Substantial heterogeneity in odds ratios was identified in terms of both size and direction of effect. In particular, six were statistically significant with odds ratios less than one (favouring sclerotherapy) and one was statistically significant with an odds ratio greater than one.

2.4. Outlying trial: CDP-choline (Figure 1(d))

Meta-analysis on a memory outcome (recall production) included seven trials of cytidinediphosphocholine (CDP-choline) for cognitive and behavioural disturbances associated with chronic cerebral disorders in the elderly [9]. The studies used a variety of instruments and so a standardized difference between mean measurements (denoted SMD) in the treatment and control groups was used. The SMD in one study differed substantially from the SMDs in the other studies. The authors noted that this study 'used a non-standard memory assessment and the results were disparate from the remaining studies. For this reason the analysis was repeated excluding this study'.

2.5. Extreme heterogeneity among few trials: gamma nail (Figure 1(e))

Our final example is an extreme set of three trials of gamma nails versus sliding hip screws for extracapsular hip fractures [10]. For the outcome of radiographic screening time, all three trials yielded a highly statistically significant mean difference, but two found longer screening times for gamma nails and one found longer screening times for sliding hip screws.

3. METHODS

3.1. Background

Different techniques for meta-analysis require differing types of information, ranging from the direction of average effect in each study to individual data from each participant in each study. We consider a widely applicable, and widely used, approach based on an observed estimate, and its corresponding precision, from each study. Methods described by DerSimonian and Laird [1] and Whitehead and Whitehead [3] follow this approach, and the so-called 'Peto method' [11] can be viewed within this framework. We denote an estimate of parameter θ_i from study i ($i=1,\ldots,k$) by y_i, and its precision (which we define as the reciprocal of the estimate's variance) by w_i. We make the conventional assumption that the precisions are known, although in reality these are estimated from the data in each study. In a traditional fixed effect meta-analysis the θ_i are assumed identical and a summary estimate, $\hat{\mu}_{\text{F}}$, is calculated as a weighted average of the study estimates, using the precisions as weights: $\hat{\mu}_{\text{F}}=\sum w_i y_i/\sum w_i$. The variance of $\hat{\mu}_{\text{F}}$ under the fixed effect assumption is $v_{\text{F}}=1/\sum w_i$. A basic random effects meta-analysis may be achieved by incorporating an estimate of the between-study heterogeneity, τ^2, into the weights [1] to produce a summary estimate $\hat{\mu}_{\text{R}}=\sum w_i^* y_i/\sum w_i^*$, where $w_i^*=(w_i^{-1}+\hat{\tau}^2)^{-1}$. An approximate variance of $\hat{\mu}_{\text{R}}$ under the random effects assumption is $v_{\text{R}}=1/\sum w_i^*$.

A test of homogeneity of the θ_i is provided by referring the statistic

$$Q=\sum w_i(y_i-\hat{\mu}_{\text{F}})^2$$

to a χ^2 distribution with $k-1$ degrees of freedom. A moment-based estimate of τ^2 may be obtained [1] by equating the observed value of Q with its expectation

$$E[Q]=\tau^2\left(\sum w_i-\frac{\sum w_i^2}{\sum w_i}\right)+k-1 \tag{1}$$

yielding

$$\hat{\tau}^2_{\text{DL}}=\frac{Q-(k-1)}{\sum w_i-\dfrac{\sum w_i^2}{\sum w_i}} \tag{2}$$

By convention this is replaced with zero if $Q < k-1$, with the consequence that, for a given set of studies, the precision of a random effects summary estimate will not exceed the precision of a fixed effect summary estimate.

3.2. Derivation of candidate measures

We derive candidate measures of heterogeneity by considering the special case in which the sampling variances of estimates from each study are known and equal, say to $1/w_i = \sigma^2$ for all i. Measures which do not fulfil appropriate properties in this situation will not be useful for consideration in the general case. In developing mathematical criteria to match the desired properties of a measure of heterogeneity, we set out the scenario as follows. We have k studies with true underlying treatment effects θ_i such that $E[\theta_i] = \mu$ and $\text{var}(\theta_i) = \tau^2$. From each study an estimate y_i of θ_i is available such that $E[y_i \mid \theta_i] = \theta_i$ and $\text{var}(y_i \mid \theta_i) = \sigma^2$. No particular distributions are assumed. The parameters underlying the scenario are μ, τ^2, σ^2 and k, where μ and τ^2 are unknown, k is known and σ^2 is assumed known.

Under these simplifying assumptions we find

$$\hat{\tau}^2_{\text{DL}} = \sigma^2 \left(\frac{Q}{k-1} - 1 \right) \tag{3}$$

$$\hat{\mu}_{\text{R}} = \hat{\mu}_{\text{F}} = \bar{y}$$

$$v_{\text{F}} = \frac{\sigma^2}{k} \tag{4}$$

$$v_{\text{R}} \approx \frac{(\sigma^2 + \tau^2)}{k}$$

and note that the unconditional variance of an individual y_i is given by

$$\text{var}(y_i) = \sigma^2 + \tau^2$$

Let us denote our measure of heterogeneity by $f(\mu, \tau^2, \sigma^2, k)$. We formulate our criteria as follows:

(i) *Dependence on the extent of heterogeneity.* Our first criterion requires that

$$f(\mu, \tau'^2, \sigma^2, k) > f(\mu, \tau^2, \sigma^2, k) \quad \text{whenever } \tau'^2 > \tau^2$$

This criterion is self-evident.

(ii) *Scale invariance.* A linear transformation of the parameter space from $\mathbb{R}$ to $a + b\mathbb{R}$ suggests the requirement that

$$f(a + b\mu, b^2\tau^2, b^2\sigma^2, k) = f(\mu, \tau^2, \sigma^2, k) \quad \text{for any } a, b$$

We impose this criterion in order that comparisons may be made across meta-analyses using different scales of measurement and using different types of outcome data. It also reflects the idea that the axes of confidence interval plots (such as Figure 1) are unimportant in describing the impact of heterogeneity.

(iii) *Size invariance.* This criterion states that the measure is not dependent on the number of studies:

$$f(\mu,\tau^2,\sigma^2,k') = f(\mu,\tau^2,\sigma^2,k) \quad \text{for all } k,k'$$

This criterion arises because the number of studies, although related to the evidence for heterogeneity, should not intrinsically affect its extent.

Note that we do not propose that our measure should be independent of the precisions of estimates observed in the studies. Thus sets of studies with identical heterogeneity τ^2, but with different degrees of sampling error σ^2, will produce different measures. Our aim is to describe the impact of the heterogeneity on the meta-analysis, its conclusions and its interpretation, rather than describing the underlying between-study variability. The latter can best be achieved simply by estimating the between-study variance, τ^2.

Criterion (iii) implies that the measurement must not involve k and (ii) implies it must not involve μ. Further, criterion (i) implies that τ^2 must be involved (as does common sense) and that $f(.)$ should increase monotonically with τ^2. Criterion (ii) thence implies that σ^2 must be involved, and also that $f(.)$ must be a function of the ratio of τ^2 to σ^2. Our goal is therefore to find a monotonic increasing function of $\rho = \tau^2/\sigma^2$ which is easily interpretable, and where the dependence on σ^2 is implicit rather than explicit so that the measure may be applied in the more general case of unequal study precisions.

First, we consider

$$\rho + 1 = \frac{\tau^2 + \sigma^2}{\sigma^2} \tag{5}$$

Using an estimate in place of τ^2 will lead to an estimate of this. The moment estimate yields, from (3)

$$H^2 = \frac{Q}{k-1} \tag{6}$$

as a possible measure of the amount of heterogeneity. Alternatively, noticing that the numerator in (5) may be estimated by kv_{R} from (4) and the denominator is kv_{F}, we might consider the statistic

$$R^2 = \frac{v_{\mathrm{R}}}{v_{\mathrm{F}}} \tag{7}$$

as an estimate of $\rho + 1$ and hence an alternative measure of the amount of heterogeneity. Both H^2 and R^2 have appealing interpretations, which we discuss in Section 4.

Now we consider a different function of ρ

$$\frac{\rho}{1+\rho} = \frac{\tau^2}{\tau^2 + \sigma^2} \tag{8}$$

Although explicitly involving the within-study variance, which will in practice vary between studies, we consider a statistic of the form

$$I^2 = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}^2}$$

We discuss approaches to deriving $\hat{\sigma}^2$, and demonstrate why this construction has potential application in the general case, in the following section.

3.3. *Generalizability*

The formal criteria we used to derive the measures do not generalize conveniently to the situation in which precisions differ between studies, where σ^{-2} is replaced by a set of values $\{w_i,\ i=1,\ldots,k\}$. Now the data are the set of k pairs $\{y_i, w_i\}$. Whereas generalizations of criteria (i) and (ii) are immediate, that of (iii) is prevented by the inseparable association between actual w_is and the number of studies, k. However, the measures H^2 and R^2 can both easily be calculated for studies with different precisions, since they do not explicitly involve σ^2 in their calculation, although they are no longer identical.

The statistic I^2 requires a number, $\hat{\sigma}^2$, that describes the 'typical' within-study variance. We consider two possibilities. The first has been suggested by Takkouche *et al.* [12], who present a statistic in the form of I^2 as a quantification of heterogeneity in a meta-analysis. They take the reciprocal of the arithmetic mean weight, to give $\hat{\sigma}^2 = kv_{\mathrm{F}}$. For the second, we note that using (1) we can write

$$E[H^2] = \frac{\tau^2 + s^2}{s^2}$$

for the general case, where

$$s^2 = \frac{\sum w_i(k-1)}{\left(\sum w_i\right)^2 - \sum w_i^2} \tag{9}$$

Hence we use $\hat{\sigma}^2 = s^2$. We prefer this second method, for there is a convenient relationship between I^2 and H^2:

$$I^2 = \frac{H^2 - 1}{H^2} \tag{10}$$

4. INTERPRETATION AND PROPERTIES OF PROPOSED MEASURES OF HETEROGENEITY

Here we develop the statistics H^2, R^2 and I^2. While we have introduced them through defining their squares, we shall henceforward address the square roots of the first two (H and R) because we believe that clinicians are, in general, more familiar with standard deviations and confidence intervals than with variances. However, we present I^2 as it stands since the concept of 'proportion of variance (un)explained' is widely familiar.

4.1. *The H statistic*

The statistic H in (6) describes the relative excess in Q over its degrees of freedom. The ratio of Q to its degrees of freedom has been suggested previously as a measure of the extent of heterogeneity [13]. Since $E[Q] = k - 1$ in the absence of heterogeneity, $H = 1$ indicates homogeneity of treatment effects. An appealing interpretation for H may be achieved through consideration of the radial plot of Galbraith [14, 15] of the standardized treatment effect

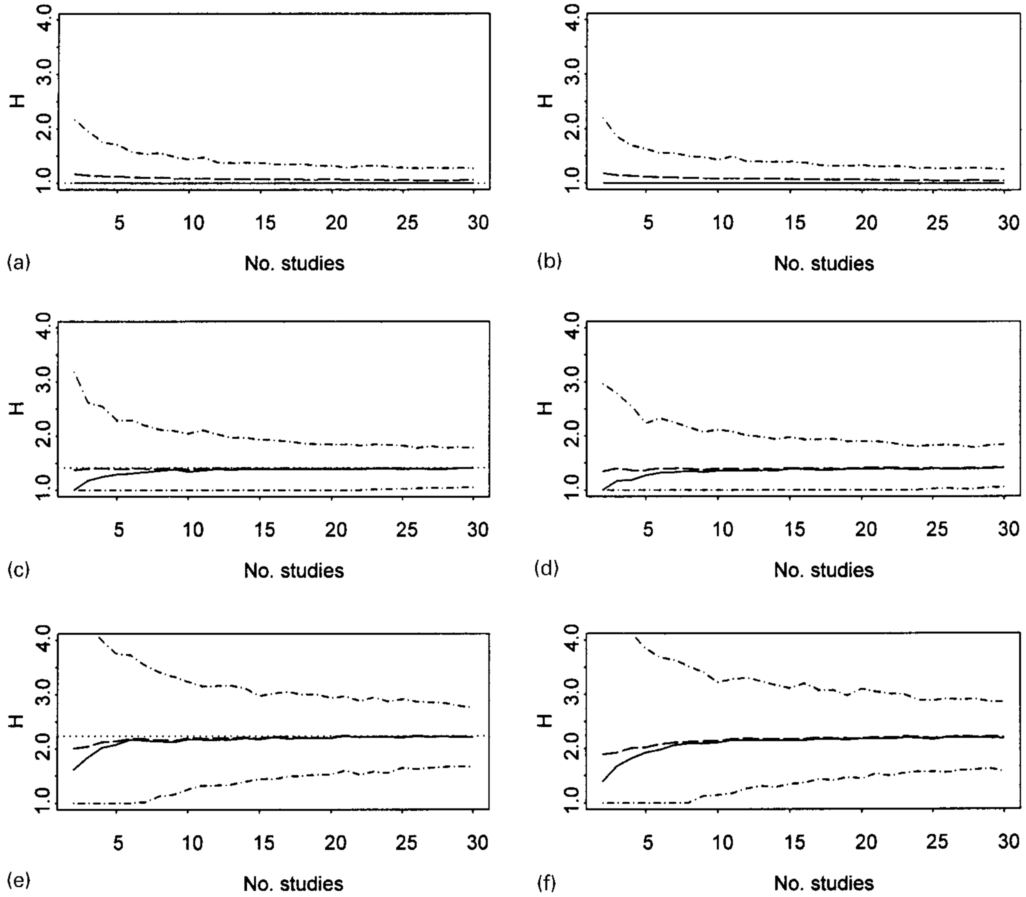

Figure 2. Values of H (out of 1000) from simulated meta-analysis data sets for increasing numbers of studies from 2 to 30, and with different values of within-study precision (w) and between-study variance (τ^2). (a) $w=1$, $\tau^2=0$; (b) variable w (between 0.1 and 1.9, average 1), $\tau^2 = 0$; (c) $w = 1$, $\tau^2=0.25$; (d) variable w, $\tau^2=0.5$; (e) $w=1$, $\tau^2=1$; (f) variable w, $\tau^2=1$. ——, median of H; ------, mean of H; -·-·-, 95 per cent reference range for H; ...., true value of H (known for cases (a), (c) and (e)).

estimates $y_i\sqrt{w_i}$ against $\sqrt{w_i}$. The slope of the unweighted least squares regression line though the origin on such a plot is given by $\hat{\mu}_F = \sum w_i y_i / \sum w_i$, that is, the traditional fixed effect meta-analytic pooled estimate. The estimated residual standard deviation from this regression is given by

$$\sqrt{\left\{\frac{\sum (y_i\sqrt{w_i} - \hat{\mu}_F\sqrt{w_i})^2}{k-1}\right\}} = \sqrt{\left(\frac{Q}{k-1}\right)}$$

which is our definition of H.

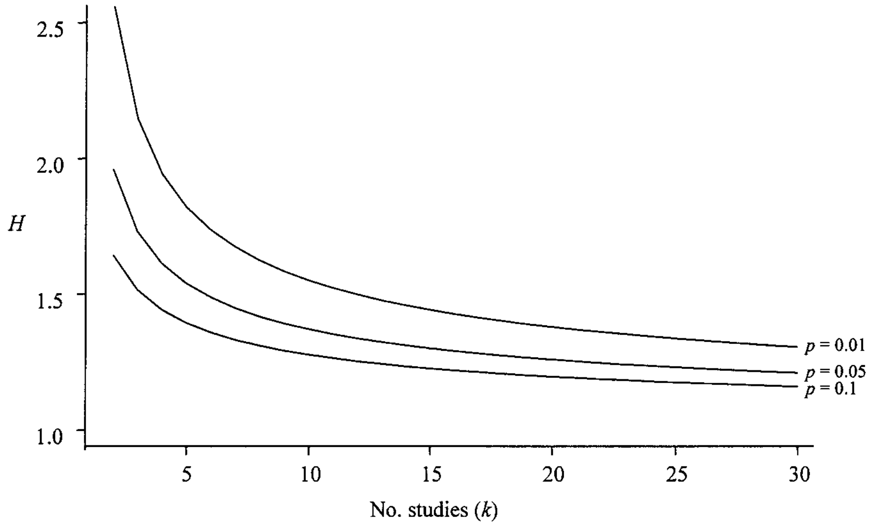

Figure 3. The mathematical relationship between H and the number of studies in a meta-analysis for three fixed p-values from the heterogeneity test ($p = 0.1$, $p = 0.05$ and $p = 0.01$).

The consistency of the statistic H across different sizes of meta-analysis is illustrated from simulations in Figure 2. The value of H does not intrinsically depend on the number of studies (unlike Q), and increases appropriately as τ^2 increases. There is a slight average bias for small numbers of studies (say less than eight). Variability in H is large for small numbers of studies, so it will be difficult in practice for moderate heterogeneity to be distinguished from chance. The variability is slow to reduce as the number of studies increases. The behaviour of H is similar when precisions vary to when they do not.

Figure 3 illustrates the mathematical relationship between the statistical test for heterogeneity (based on Q) and the value of H over varying numbers of studies. With a small number of studies, statistically significant heterogeneity would be evident only when the impact of heterogeneity, as measured by the H statistic, is high. This explicitly highlights the poor properties of the test when there are few studies. The graph may also be used as an aid to the interpretation of H in relation to the more familiar test result.

4.1.1. Calculation and uncertainty. Calculation of H is straightforward, and possible from the majority of published metaanalyses, where either the Q-statistic or its p-value is presented (since the number of studies is generally known). We choose to present the maximum out of H and 1, though we recognize that this will prevent identification of excessive *homogeneity* – that is less variability than would be expected by chance – perhaps due to studies being replicated within a meta-analytic data set.

One possible complication in practice is that the Q-statistic is not always calculated using the inverse-variance weighted average $\hat{\mu}_{\mathrm{F}}$. For example, the software MetaView used in the Cochrane Database of Systematic Reviews [16] makes use of the Mantel–Haenszel fixed effect summary estimate in place of $\hat{\mu}_{\mathrm{F}}$ for dichotomous outcome data. The difference between the results using these two methods is however usually small.

In the Appendix we discuss eight methods of calculating confidence intervals for H and report brief results of a Monte Carlo investigation of their performance. We prefer to interpret these as reference intervals for H since we do not consider H to be estimating an intuitive parameter. The intervals describe the variability associated with the value of H for studies with precisions identical to those observed in the current meta-analysis. They reflect uncertainty in the extent of heterogeneity. For practical application, we recommend a simple construction for an interval (method III in the Appendix) involving only Q and k, derived from a test-based standard error for $\ln(Q)$. Intervals are of the form

$$\exp(\ln H \pm Z_\alpha \times \mathrm{SE}[\ln(H)])$$

where Z_α is the $(1-\alpha/2)$ quantile of the standard normal distribution and

$$\mathrm{SE}[\ln(H)] = \frac{1}{2}\,\frac{\ln(Q)-\ln(k-1)}{\sqrt{(2Q)}-\sqrt{(2k-3)}} \quad \text{if } Q > k$$

$$\sqrt{\left\{\frac{1}{2(k-2)}\left(1-\frac{1}{3(k-2)^2}\right)\right\}} \quad \text{if } Q \leqslant k$$

In the examples that follow we also consider, for comparison, some alternative methods: a bootstrap interval (method VIII); a maximum likelihood method (method IV), and a Bayesian approach (method VII). The last two arise from viewing H as an estimate of

$$\eta = \sqrt{\left\{\frac{(\sum w_i - \sum w_i^2/\sum w_i)\tau^2}{k-1}+1\right\}} \tag{11}$$

as discussed at greater length in the Appendix.

4.1.2. Application. Table I lists values for H for our five examples. The homogenous albumin trials give a value of $Q/(k-1)$ below 1, which we replace by $H=1$. In such situations we calculate a confidence interval from the standard error of $\ln(Q/(k-1))$ under homogeneity, giving an upper limit of 1.34. Too few of the bootstrap samples yielded values for H greater than 1, resulting in both the lower and upper limits of a 95 per cent interval being 1. The upper end of the Bayesian credible interval was 1.26, supporting a conclusion that there is minimal between-study variability in this data set.

A value of $H=1.19$ for the adjuvant chemotherapy trials indicates the presence of some heterogeneity. All 95 per cent intervals include 1, and upper bounds range from 1.46 for the ML interval to 2.31 for the Bayesian interval. In this and subsequent examples, the Bayesian interval has a noticeably larger upper limit than the other intervals. This arises since the method places no restrictions on the distribution of $\hat{\tau}^2$.

The heterogeneous sclerotherapy trials give a value of H of 2.13. This indicates that the residual standard deviation on the Galbraith plot is just over twice the value than if the studies had been homogenous. All intervals of uncertainty exclude 1, providing strong evidence that the heterogeneity is real. The test-based interval is the narrowest interval. The bootstrap interval is slightly wider, and the ML interval wider still. ML intervals will tend to be wider since they are based on symmetric confidence intervals for τ^2, which has a highly skewed distribution. The Bayesian version of H again has the widest uncertainty interval.

Table I. Summary statistics from the five example data sets and values of H, R and I^2 with approximate confidence intervals, (a) and (b) are based on H as defined in (6); (c) and (d) are based on H viewed as an estimate of (11). (e) is based on R as defined in (7); (f) and (g) are based on R viewed as (13). (h) to (k) that follow are calculated from transformation (10) applied to (a) to (d).

Data set	Albumin	Adjuvant chemotherapy	Sclerotherapy	CDP-choline	Gamma nail
Number of trials	24	11	19	7	3
Treatment effect	log RR	log HR	log OR	SMD	MD
Heterogeneity variance, $\hat{\tau}^2$	0	0.09	0.98	0.39	1745
'Typical within-study variance', s^2	0.52	0.22	0.28	0.067	27
$\hat{\tau}^2/s^2$ (as a percentage)	0%	41%	353%	591%	6415%
Test for heterogeneity	$Q = 14.4$ $(p = 0.91)$	$Q = 14.1$ $(p = 0.17)$	$Q = 81.5$ $(p = 4.7 \times 10^{-10})$	$Q = 41.5$ $(p = 2.3 \times 10^{-7})$	$Q = 130.3$ $(p < 10^{-16})$
Method	H (95% CI)	H (95% CI)	H (95% CI)	H (95% CI)	H (95% CI)
(a) Point value with test-based confidence interval	1 (1, 1.34)	1.19 (1, 1.69)	2.13 (1.71, 2.64)	2.63 (1.90, 3.65)	8.07 (6.08, 10.72)
(b) Point value with bootstrap confidence interval (1000 samples)	1 (1, 1)	1.19 (1, 1.48)	2.13 (1.48, 2.55)	2.63 (1, 3.83)	8.07 (1, 8.26)
(c) Based on ML estimate of τ^2	1 (1, 1.17)	1.11 (1, 1.46)	2.18 (1.23, 2.82)	2.68 (1, 3.84)	5.55 (1, 8.97)
(d) Based on Bayesian estimate of τ^2 using BUGS*	1.02 (1.00, 1.26)	1.25 (1.00, 2.31)	2.36 (1.68, 3.62)	3.35 (1.89, 7.51)	9.75 (3.88, 51.5)
Method	R (95% CI)	R (95% CI)	R (95% CI)	R (95% CI)	R (95% CI)
(e) Point value with bootstrap confidence interval (1000 samples)	1 (1, 1)	1.32 (1, 1.68)	2.27 (1.58, 2.71)	2.68 (1, 4.04)	15.3[†] (1, 15.8)
(f) Based on ML estimate of τ^2	1 (1, 1.29)	1.21 (1, 1.66)	2.32 (1.30, 2.97)	2.72 (1, 3.90)	10.5 (1.00, 17.0)
(g) Based on Bayesian estimate of τ^2 using BUGS*	1.04 (1.00, 1.41)	1.40 (1.01, 2.58)	2.50 (1.80, 3.77)	3.40 (1.93, 7.60)	18.4 (7.38, 97.2)
Method	I^2 (95% CI)	I^2 (95% CI)	I^2 (95% CI)	I^2 (95% CI)	I^2 (95% CI)
(h) Point value with test-based confidence interval	0% (0%, 45%)	20% (0%, 65%)	78% (66%, 86%)	86% (72%, 92%)	98% (97%, 99%)
(i) Point value with bootstrap confidence interval (1000 samples)	0% (0%, 0%)	29% (0%, 54%)	78% (54%, 85%)	86% (0%, 94%)	98% (0%, 99%)
(j) Based on ML estimate of τ^2	0% (0%, 27%)	19% (0%, 53%)	79% (34%, 87%)	86% (72%, 98%)	97% (0%, 99%)
(k) Based on Bayesian estimate of τ^2 using BUGS*	4% (0%, 47%)	36% (0%, 81%)	82% (65%, 92%)	91% (72%, 98%)	99% (93%, 100%)

*Median and 95 per cent interval from posterior distribution.

[†] Mean value of R over bootstrap samples was 7.03.

The H statistic for the CDP-choline data set is 2.63, with a 95 per cent test-based confidence interval from 1.90 to 3.65, indicating strong evidence of genuine heterogeneity. The ML and bootstrap confidence intervals have a lower limit of 1. It is clear that the bootstrap approach will produce misleading results for data sets such as this one, which has a single outlying trial. Sampling with replacement from the 7 trials, 34 per cent of samples will not include the outlying trial, and will therefore yield small values for H (the value of H for the 6 trials excluding the outlier is 1). The Bayesian interval is much the widest, although it excludes 1, reflecting the considerable uncertainty around the value of the between-trial variance in this peculiar data set.

The three conflicting trials of gamma nails versus sliding hip screws give a huge value of $H=8.07$ (95 per cent test based interval from 6.08 to 10.72). The ML and bootstrap intervals again both begin at 1, and are not realistic summaries of uncertainty in the impact of heterogeneity in this meta-analysis. The Bayesian interval stretches from 3.88 to 51.5, reflecting the high uncertainty surrounding estimation of τ^2 from only three trials.

4.2. *The R statistic*

Interpretation of the square of R, as defined in (7), may be made in a similar manner to that of a design effect in cluster sampling [17]. It describes the inflation in the confidence interval for a single summary estimate under a random effects model compared with a fixed effect model. A value of 1 indicates identical inferences under the two models, that is when treatment effects are homogenous and the fixed effect model is sufficient. Consistency of R over various scenarios is very similar to that for H illustrated in Figure 2.

4.2.1. Calculation and uncertainty. A value for R may be obtained using any method of obtaining the appropriate variances. We concentrate on the computationally straightforward calculation [1, 3]

$$R=\sqrt{\left\{\frac{\sum w_i}{\sum w_i^*}\right\}}=\sqrt{\left\{\frac{\sum w_i}{\sum (w_i^{-1}+\hat{\tau}^2)^{-1}}\right\}} \tag{12}$$

We may thus view R as an estimate of a function of τ^2 alone (since the w_i are assumed known). This leads to the selection of estimates and approximate confidence intervals based on methods for estimating τ^2, as described for H in the Appendix. A Bayesian interval and a bootstrap confidence interval are also available.

4.2.2. Application. Table I lists values for R for our five examples. A test-based interval such as that based on the significance test for Q is not available for R. Both values of R and confidence intervals are very similar to those for H. This is to be expected in many circumstances since H and R coincide when all estimates have equal precision. The gamma nail trials yield values substantially larger than for H, again with wide 95 per cent intervals reflecting the uncertainty associated with the small number of trials.

4.3. *The I^2 statistic*

The denominator of the right-hand side of (8) is the unconditional variance of the y_i, which comprises additive components due to within-study variation (usually between-patient variation) and between-study variation (heterogeneity). The quantity therefore has an appealing interpretation as the proportion of total variation in the estimates of treatment effect that is due to heterogeneity between studies. It is similar in concept to the intraclass correlation coefficient in cluster sampling. This interpretation may be made approximately for the statistic I^2 in the general case.

4.3.1. Calculation and uncertainty. We define I^2 in terms of H using equation (10). This allows us to express inferences made on H in terms of I^2. Confidence limits for I^2 are therefore readily available from confidence limits for H as discussed above and in the Appendix.

4.3.2. Application. Values for I^2 expressed as a percentage are included in Table I for the five examples. There is a direct correspondence between these and results for H. The change of scale, however, reveals new interpretations for the findings. We focus here on the point values alone. An I^2 of 0 per cent for the albumin trials indicates that all variability in effect estimates is due to sampling error within trials, and that none is due to heterogeneity. On the other hand, despite a non-significant heterogeneity test result ($p = 0.17$), some 20 per cent of variability in the adjuvant chemotherapy trials may be attributable to between-study variation. For the sclerotherapy, CDP-choline and gamma nail examples, values for I^2 are 78, 86 and 98 per cent, respectively.

5. DISCUSSION

We have developed three measures, H, R and I^2, for quantifying the impact of heterogeneity in a meta-analysis. The measures H and R are identical if all studies have identical precisions, and will be similar in most other situations. The measure I^2 is a transformation of H that has a different, yet intuitive, interpretation. Our measures are more relevant than the result of the test for heterogeneity that is commonly presented in a meta-analysis. The test has poor power with few studies and inappropriately high power with many studies, and it can therefore be difficult to decide either whether heterogeneity is present or whether it is clinically important. H, R and I^2 do not depend on the number of studies. We recommend that one of these be presented in place of the test.

Our measures quantify the impact rather than the extent of heterogeneity in a meta-analysis. The impact depends on the precisions of the study-specific estimates, which a measure of extent should not. The between-study variance measures extent of heterogeneity and is an important parameter in its own right, but it is specific to a particular treatment effect metric. A measure that compares the extent of heterogeneity across different scales would be useful. One approach might be to measure treatment effects in a consistent way across studies, irrespective of the data type measured on individuals [18]. However, this would involve a different approach to the meta-analysis, and would not lead to a measure that is simple to calculate from published reviews. In common with Q or its p-value [19], our measures can be compared across different treatment effect metrics. For example, values of H for the sclerotherapy data (Figure 1(c)) are 2.13 if the log-odds ratio is used, compared to 1.92 for the log relative risk and 2.60 for the risk difference scales. An analysis focusing on relative risks would therefore suffer from the least impact of heterogeneity. The principal advantage of H, R and I^2 over Q is that comparisons can also be made across meta-analyses of different sizes.

How should particular values for H, R and I^2 and their confidence intervals be interpreted in practice? Some indications for H are given by Figure 3. For ten studies, statistically significant heterogeneity at $p=0.1$, $p=0.05$ and $p=0.01$ would be identified from the test when $H=1.28$, $H=1.37$ and $H=1.55$, respectively. For 30 studies, heterogeneity would be identified when $H=1.16$, $H=1.21$ and $H=1.31$, respectively. No universal rule could cover definitions for 'mild', 'moderate' or 'severe' heterogeneity, but it would seem that values exceeding 1.5 might induce considerable caution and values below 1.2 might cause little concern. These correspond to values of I^2 of 56 per cent and 31 per cent. Thus, mild heterogeneity might account for less than 30 per cent of the variability in point estimates, and notable heterogeneity substantially more than 50 per cent. However, these suggestions are tentative, not least because the practical impact of heterogeneity in a meta-analysis also depends on the size and direction of treatment effects. The interpretation of heterogeneity in a systematic review will depend critically on these, as well as on considerations of clinical and methodological diversity in the studies.

Our measures may be used to give an indication of the contribution of individual studies [20] and covariates [21] to the impact of heterogeneity. Derivation of a Bayesian interval for the difference between the measures with and without particular studies is possible in BUGS [22]. By replacing $\hat{\mu}_\text{F}$ in the definition of Q with the fitted value from a regression equation, a version of H (and therefore I^2) is defined for the impact of heterogeneity in a meta-regression. Similarly, by taking the ratio of the standard errors of slopes from fixed effect and random effects meta-regressions, a version of R is defined for univariate meta-regression.

Approximate confidence intervals are available for all three measures, but a test-based interval for H (or I^2) is particularly easy to calculate, being based only on the values of Q and k, and has reasonable coverage. We have not found a good, simple interval for R. Bayesian intervals for all statistics are available and might be considered the gold standard. However, they require computer-intensive methods and software expertise to calculate them, and can be sensitive to prior distributions, especially when there is little information concerning the heterogeneity variance as in the gamma nail example. A bootstrap-generated interval provides a reasonable interval when there are many studies and heterogeneity permeates the whole data set (rather than being due to a small number of outlying studies), but is poor otherwise.

In conclusion, we propose H and I^2 as our favoured measures for quantifying heterogeneity in a meta-analysis. H may be interpreted approximately as the ratio of confidence interval widths for single summary estimates from random effects and fixed effect meta-analyses (that is, is approximately equal to R). I^2 describes the percentage of variability in point estimates that is due to heterogeneity rather than sampling error. Both may be readily calculated from most published meta-analyses, and a closed form uncertainty interval is available.

APPENDIX

We outline four approaches to calculating uncertainty intervals for H, leading to eight distinct methods. The approaches are (i) based on the distribution of Q, (ii) based on the statistical significance of Q, (iii) based on the estimation of τ^2, and (iv) using a non-parametric bootstrap procedure.

A1. (i) Intervals based on the distribution of Q

Biggerstaff and Tweedie [23] discuss the distribution of Q. If the fixed effect model is true then Q has a χ^2 distribution with $k-1$ degrees of freedom. Otherwise it has a non-central χ^2 distribution. Biggerstaff and Tweedie approximate the general distribution of Q by a gamma distribution with mean and variance equal to the mean and variance of Q, which they derive.

Method I. A symmetric Wald-type uncertainty interval for H may be based on the variance of Q given in equation (7) of Biggerstaff and Tweedie [23]. A 95 per cent uncertainty interval for H may be calculated as

$$\sqrt{\left\{\frac{1}{k-1}\left(Q \pm 1.96\sqrt{\operatorname{var}(Q)}\right)\right\}}$$

Method II. The gamma approximation to the distribution of Q, as described by Biggerstaff and Tweedie, provides a more appropriate interval. This requires evaluation of quantiles from the cumulative distribution function of the gamma distribution, requiring more advanced software than method I.

A2. (ii) Intervals based on the statistical significance of Q

Method III. Test-based methods [24] provide a second approach to calculating confidence intervals for Q, and hence for H. We base these on $\ln(Q)$ in order to remove some of the skew inherent in the distribution of Q. One approach, which can be calculated without the use of statistical tables or computer software, is based on a normal approximation to the χ^2 distribution for large degrees of freedom: $Z=\sqrt{(2Q)}-\sqrt{(2k-3)}$ has approximately a standard normal distribution (formula 26.4.13 in Abranowitz and Stegun [25]) so, equating Z with $(\ln(Q)-\ln(k-1))/\mathrm{SE}[\ln(Q)]$, we can estimate a standard error for $\ln(Q)$ using

$$\mathrm{SE}[\ln(Q)]=\frac{\ln(Q)-\ln(k-1)}{\sqrt{(2Q)}-\sqrt{(2k-3)}}$$

Now, since $Q=(k-1)H^2$, and k is a constant, we have $\operatorname{var}[\ln(Q)]=4\operatorname{var}[\ln(H)]$ and hence a test-based standard error for $\ln(H)$ is

$$\mathrm{SE}_1[\ln(H)]=\frac{1}{2}\,\frac{\ln(Q)-\ln(k-1)}{\sqrt{(2Q)}-\sqrt{(2k-3)}}$$

A drawback to the test-based standard error is that it approaches zero as H approaches 1, whereas we define H to be 1 whenever $Q\leqslant k-1$. To overcome this, for small H, we take a standard error based on the approximate variance of $\ln(Q/(k-1))=2\ln(H)$ when Q truly has a χ^2 distribution with $k-1$ degrees of freedom (formula 26.4.36 in Abranowitz and Stegun [25]):

$$\mathrm{SE}_0[\ln(H)]=\sqrt{\left\{\frac{1}{2(k-2)}\left(1-\frac{1}{3(k-2)^2}\right)\right\}}$$

Using one or other of these standard errors, a 95 per cent uncertainty interval for H follows as

$$\exp(\ln H \pm 1.96 \times \mathrm{SE}[\ln(H)])$$

Note that since $\mathrm{SE}_1[\ln(H)]$ approaches 0 as H approaches 1 (with equality at $H=1$, that is when $Q \leqslant k-1$), we do not advocate switching to $\mathrm{SE}_0[\ln(H)]$ exactly at this point. Empirical examination of the behaviour of $\mathrm{SE}_1[\ln(H)]$ in the range $k=2,\ldots,100$ suggests than $\mathrm{SE}_0[\ln(H)]$ should be used whenever $Q \leqslant k$, and this is the policy we follow.

A3. (iii) Intervals based on the estimation of τ^2

Using (2) we may view H as an estimate of η in (11). Thus a third approach to the calculation of H and uncertainty intervals for H is to base them on estimates and uncertainty intervals for τ^2, since all quantities in η other than τ^2 are assumed known. Various approaches to estimating τ^2 with associated uncertainty are available. Biggerstaff describes several approaches (unpublished data). We consider three of them here:

Method IV. A maximum likelihood (ML) method based on an iterative estimate of τ^2 and a closed-form confidence interval.

Method V. A restricted maximum likelihood (REML) method based on an iterative estimate of τ^2 and a closed-form confidence interval.

Method VI. A closer approximation than the gamma distribution of Biggerstaff and Tweedie to the distribution of Q is a Pearson type III distribution based on equating the mean, variance and skewness. Appropriate quantiles, giving rise to an approximate 95 per cent confidence interval for Q, may be determined from the cumulative distribution function of the Pearson distribution and converted to a 95 per cent confidence interval for τ^2.

Further approaches to estimating τ^2 with confidence intervals are available, for example the likelihood approach described by Hardy and Thompson [26]. We do not include this approach here. However, we do consider a Bayesian approach:

Method VII. The expression (11) also allows us to compute a Bayesian estimate of η (as a version of H), for example by adding this calculation into BUGS [22] code for a simple random effects meta-analysis with non-informative priors [27]. In our analyses we use locally non-informative N(0,1000) prior distributions for μ and Uniform(0,1000) prior distributions for τ.

A4. (iv) Bootstrap intervals

Method VIII. Finally, a bootstrap interval for H may be obtained by taking samples of size k with replacement from the pairs $\{y_i, w_i\}$ and calculating quantiles for the H statistic over repeated samples.

A5. Simulation study

We compared the eight methods of obtaining confidence intervals for H described above by simulation. We assessed the coverage of a 95 per cent confidence interval for a variety of simulated data sets in which (i) the number of studies, (ii) the relative study weights, (iii)

Table A1. Percentage coverage of eight methods for calculating 95 per cent intervals for H. Results from 1000 simulations. For equal precisions the precisions were set to 1. Unbalanced precisions took values of (1,1,20) for three studies, eight 1s and two 20s for 10 studies and sixteen 1s and four 20s for 30 studies. 'Typically variable' precisions took values from the sclerotherapy data set, in the order they appear in Figure 1(c). The heterogeneity variance, τ^2, was calculated as the stated proportion of s^2 in (9).

		Study precisions											
		Equal				Unbalanced				Typically variable			
$\tau^2/s^2 \times 100\%$		0%	25%	100%	400%	0%	25%	100%	400%	0%	25%	100%	400%
(I) Var(Q)	$k = 3$	100	100	100	100	100	100	100	100	100	100	100	100
	10	100	100	99	93	100	100	100	100	100	100	100	100
	30	98	98	97	94	100	100	100	96	99	100	100	100
(II) Gamma for Q	$k = 3$	100	100	100	75	100	100	100	79	100	100	100	73
	10	100	100	100	91	100	100	100	86	100	100	100	87
	30	99	100	93	93	100	100	89	90	100	100	93	92
(III) Test-based	$k = 3$	97	96	96	91	97	97	95	90	97	96	95	87
	10	97	97	95	80	98	97	91	60	98	96	93	74
	30	97	98	88	77	98	96	77	53	97	96	87	73
(IV) ML for τ^2	$k = 3$	100	100	100	55	100	100	23	54	100	100	100	47
	10	100	100	79	81	100	100	54	74	100	100	75	79
	30	100	100	90	89	100	68	76	88	100	100	88	89
(V) REML for τ^2	$k = 3$	100	100	100	67	100	100	94	69	100	100	100	62
	10	100	100	86	86	100	100	69	82	100	100	82	83
	30	100	100	92	91	100	78	81	90	100	100	91	90
(VI) Type III for τ^2	$k = 3$	98	97	98	98	98	98	98	97	98	97	98	97
	10	98	98	97	95	98	98	98	86	98	98	97	94
	30	97	98	95	94	98	98	94	94	97	98	96	95
(VII) Bayesian for τ^2	$k = 3$	0*	95	95	94	0	95	95	92	0	95	95	92
	10	0	98	97	94	0	97	97	95	0	98	96	94
	30	0	99	93	94	0	97	94	94	0	95	94	95
(VIII) Bootstrap	$k = 3$	100	44	43	44	100	53	52	56	100	50	45	46
	10	100	79	76	79	100	77	71	63	100	79	77	76
	30	99	89	89	88	99	88	83	74	99	89	88	88

*The Bayesian model restricts the between-study variance to be non-negative, so when τ^2 is truly zero a credible interval will have zero coverage.

the amount of heterogeneity and (iv) the symmetry of the distribution of true study effects were varied. Some results are given in Table A1. The Bayesian approach was found to have good coverage, and might be considered the gold standard as it assumes no particular distribution for the between-study variance. The intervals based on symmetric intervals for Q or τ^2 had excessive coverage. The Pearson type III approximation worked well in all situations, though it is complicated to calculate. The bootstrap confidence interval had poor coverage. The test-based confidence interval also performed well in all but the most extreme situations. We recommend this one for its ease of calculation.

ACKNOWLEDGEMENTS

We thank Doug Altman and Jon Deeks for early discussions, Brad Biggerstaff for providing S-plus code for calculating confidence intervals for τ^2 and for further help, Ian White for helpful discussions at a later stage, and an editor and two referees for useful comments. This research was funded by MRC project grant G9815466.

REFERENCES

1. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**:177–188.
2. Cochran WG. The combination of estimates from different experiments. *Biometrics* 1954; **10**:101–129.
3. Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomised clinical trials. *Statistics in Medicine* 1991; **10**:1665–1677.
4. Hardy RJ, Thompson SG. Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine* 1998; **17**:841–856.
5. Cochrane Injuries Group Albumin Reviewers. Human albumin administration in critically ill patients: systematic review of randomised controlled trials. *British Medical Journal* 1998; **317**:235–240.
6. Beale RJ, Wyncoll DLA, McLuckie A. Human albumin administration in critically ill patients: Analysis is superficial and conclusions exaggerated. *British Medical Journal* 1998; **317**:884.
7. Sarcoma Meta-analysis Collaboration. Adjuvant chemotherapy for localised resectable soft-tissue sarcoma of adults: meta-analysis of individual data. *Lancet* 1997; **350**:1647–1654.
8. Pagliaro L, D'Amico G, Sorensen TIA, Lebrec D, Burroughs AK, Morabito A, Tine F, Politi F, Traina M. Prevention of first bleeding in cirrhosis: a meta-analysis of randomized trials of nonsurgical treatment. *Annals of Internal Medicine* 1992; **117**:59–70.
9. Fioravanti M, Yanagi M. Cytidinediphosphocholine (CDP choline) for cognitive and behavioural disturbances associated with chronic cerebral disorders in the elderly (Cochrane Review). *Cochrane Database of Systematic Reviews*. Update Software: Oxford, 2000, Issue 3.
10. Parker MJ, Handoll HHG. Gamma and other cephalocondylic intramedullary nails versus extramedullary implants for extracapsular hip fractures (Cochrane review). *Cochrane Database of Systematic Reviews*. Update Software: Oxford, 2000, Issue 3.
11. Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: an overview of the randomised trials. *Progress in Cardiovascular Diseases* 1985; **27**:335–371.
12. 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.
13. National Health and Medical Research Council (Australia). *How to Review the Evidence*: *Systematic Identification and Review of the Scientific Literature*. National Health and Medical Research Council: Canberra, 2000.
14. Galbraith RF. Some applications of radial plots. *Journal of the American Statistical Association* 1994; **89**: 1232–1242.
15. Galbraith RF. A note on graphical presentation of estimated odds ratios from several clinical trials. *Statistics in Medicine* 1988; **7**:889–894.
16. The Cochrane Collaboration. Cochrane Database of Systematic Reviews. Update Software: Oxford, 2001, Issue 2.
17. Kish L. *Survey Sampling*. Wiley: New York, 1965.
18. Colditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of therapy. 1: Medical. *Statistics in Medicine* 1989; **8**:441–454.
19. Engels EA, Schmid CH, Terrin N, Olkin I, Lau J. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. *Statistics in Medicine* 2000; **19**:1707–1728.

20. Thompson SG. Controversies in meta-analysis: the case of the trials of serum cholesterol reduction. *Statistical Methods in Medical Research* 1993; **2**:173–192.
21. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in Medicine* 1999; **18**:2693–2708.
22. Spiegelhalter DJ, Thomas A, Best NG, Gilks WR. *BUGS*: *Bayesian Inference Using Gibbs Sampling. Version 0.50*. MRC Biostatistics Unit: Cambridge, 1995.
23. Biggerstaff BJ, Tweedie RL. Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in Medicine* 1997; **16**:753–768.
24. Miettinen O. Estimability and estimation in case-referent studies. *American Journal of Epidemiology* 1976; **103**:226–235.
25. Abramowitz M, Stegun IA. *Handbook of Mathematical Functions with Formulas*, *Graphs*, *and Mathematical Tables*. Dover Publications: New York, 1965.
26. Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. *Statistics in Medicine* 1996; **15**:619–629.
27. Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: a comparative study. *Statistics in Medicine* 1995; **14**:2685–2699.