A re-evaluation of the 'quantile approximation method' for random effects meta-analysis

Dan Jackson*,† and Jack Bowden

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

SUMMARY

The quantile approximation method has recently been proposed as a simple method for deriving confidence intervals for the treatment effect in a random effects meta-analysis. Although easily implemented, the quantiles used to construct intervals are derived from a single simulation study. Here it is shown that altering the study parameters, and in particular introducing changes to the distribution of the within-study variances, can have a dramatic impact on the resulting quantiles. This is further illustrated analytically by examining the scenario where all trials are assumed to be the same size. A more cautious approach is therefore suggested, where the conventional standard normal quantile is used in the primary analysis, but where the use of alternative quantiles is also considered in a sensitivity analysis. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS: meta-analysis; random effects model; quantile approximation method

1. INTRODUCTION

Meta-analysis, the statistical process of combining the results from separate trials concerned with the same treatment or issue, is frequently used in medical statistics. A standard model for performing meta-analyses is the random effects model and the most commonly used procedure for implementing this is that suggested by DerSimonian and Laird [1]. Despite this it should be noted that other approaches are possible when using this model [2–4].

The random effects model assumes that the outcome from the ith of k trials is distributed as $y_i|\mu_i \sim N(\mu_i, \sigma_i^2)$, where μ_i is the true underlying treatment effect for this trial and σ_i^2 is the corresponding within-study variance. This variance is estimated in practice, but assumed fixed and known when pooling the results, and can be obtained for a range of measures used in meta-analysis

*Correspondence to: Dan Jackson, MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR, U.K.
†E-mail: daniel.jackson@mrc-bsu.cam.ac.uk

as described in detail by Sutton *et al.* [5]. This conditional distribution is justified by the Central Limit and Slutsky's theorems but requires sufficiently large trials in order to provide a suitable approximation. The random effects model further assumes that $\mu_i \sim N(\mu, \tau^2)$, where μ and τ^2 denote the overall treatment effect and between-study variance, respectively, and that the trials are independent. This provides the marginal distributions $y_i \sim N(\mu, \sigma_i^2 + \tau^2)$.

DerSimonian and Laird estimate τ^2 using the Q statistic,

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

where $w_i = \sigma_i^{-2}$ and $\bar{y} = \sum_{i=1}^{k} w_i y_i / \sum_{i=1}^{k} w_i$. Evaluating $E[Q]$ as a function of τ^2 and matching this moment provides the DerSimonian and Laird estimate

$$\hat{\tau}^2 = \max \left(0, \frac{Q - (k-1)}{\sum_{i=1}^{k} w_i - \sum_{i=1}^{k} w_i^2 / \sum_{i=1}^{k} w_i} \right)$$

If $\hat{\tau}^2$ is truncated to zero then the resulting confidence intervals and results from hypothesis tests are the same as those from a fixed effects analysis, which assumes from the outset that $y_i \sim N(\mu, \sigma_i^2)$. Hence $\hat{\tau}^2 = 0$ effectively means that a fixed effects procedure is adopted, although a random effects perspective may be maintained when interpreting the results.

Since $\hat{\tau}^2$ is a consistent estimate, *assuming that the number of trials is sufficiently large* we can, as a further approximation, replace τ^2 by this estimate when making inferences about the overall treatment effect μ. Hence we use the model $y_i \sim N(\mu, \sigma_i^2 + \hat{\tau}^2)$ for this purpose, which results in

$$\hat{\mu} = \frac{\sum_{i=1}^{k} w_i(\hat{\tau}) y_i}{\sum_{i=1}^{k} w_i(\hat{\tau})}$$

where $w_i(\hat{\tau}) = 1/(\sigma_i^2 + \hat{\tau}^2)$. The variance of this estimate is approximately $\text{var}(\hat{\mu}) = 1/\sum_{i=1}^{k} w_i(\hat{\tau})$ so that a nominal 95 per cent confidence interval is given by $\hat{\mu} \pm 1.96\sqrt{\text{var}(\hat{\mu})}$. However this relies on the notion that, approximately,

$$M = \frac{\hat{\mu} - \mu}{\sqrt{\text{var}(\hat{\mu})}} \sim N(0, 1)$$

A recently proposed modification to the approach suggested by DerSimonian and Laird is the 'quantile approximation method' [6]. This ultimately follows a very similar procedure as DerSimonian and Laird but uses an alternative set of quantiles to the usual standard normal 0.975 quantile of 1.96 when constructing 95 per cent confidence intervals. These quantiles are derived from a simulation study using parameters that correspond to 'typical scenarios for estimating a log odds ratio' [6]. The resulting quantiles are all larger than 1.96 and the modification results in wider confidence intervals with greater coverage probabilities that are intended to correspond more closely to the nominal value than the more usual procedure achieves. A related argument that leads to the use of larger quantiles is the observation by Follmann and Proschan [7] that if the within-study variances are all equal, and one uses the untruncated version of the DerSimonian and Laird estimate, the resulting test statistic has a t rather than a standard normal null distribution. Sutton and Higgins [8, p. 628] refer to other recent studies that suggest using a t, rather than standard normal, null distribution in the context of meta-analysis. In particular, Sidik and Jonkman [9] suggest using a t distribution with a variety of variance estimators and emphasize that both

the within and between-study variances are estimates. Böckenhoff and Hartung [10] address the specific issue of estimating the within-study variances, but here it will be assumed that the studies are large enough for these to be regarded as known and that the random effects model is correct. The problems are therefore caused by having to estimate the between-study variance, and using this as if it was the true value, in the conventional way.

Although the quantile approximation method is intuitively appealing, a single simulation study provides little assurance that the properties of M will be similar under alternative circumstances. The intention of this paper is to re-assess this method. The rest of the paper is set out as follows. In Section 2 the procedure from which the quantile approximation method was derived is briefly described, and in Section 3 some further simulation studies are performed, and analytical results are derived, that show that the original simulation study need not provide a good indication of the properties of M for alternative scenarios. A more cautious approach is therefore suggested, where the usual standard normal quantile is used in the primary analysis, but the use of alternative quantiles is considered in a secondary or sensitivity analysis. This procedure is illustrated using an example data set in Section 4. Section 5, the discussion, concludes the paper.

2. THE QUANTILE APPROXIMATION METHOD

The quantile approximation method has been recently proposed by Brockwell and Gordon; for a full description of this and its justification, see their original paper [6]. In the original simulation study, values of σ_i^2 were simulated from a scaled and truncated χ^2 distribution (with one degree of freedom, using a scale factor of 0.25 and then truncated to lie in the interval [0.009, 0.6]). For each of $k = 2, 3, \ldots, 30$, 19 values $\tau^2 = 0, 0.01, 0.02, \ldots, 0.1, 0.15, 0.2, \ldots, 0.5$ were used, for each of which 25 000 simulated data sets were produced from the random effects model $y_i \sim \mathrm{N}(\mu, \sigma_i^2 + \tau^2)$; $\mu = 0.5$ was used but this choice is immaterial. This results in $19 \times 25\,000 = 475\,000$ simulated meta-analyses for each k. The random variable $M = (\hat{\mu} - \mu)/\sqrt{\mathrm{var}(\hat{\mu})}$ was evaluated for each of these sets of 475 000 meta-analyses and the 0.025 and 0.975 quantiles, denoted by $\hat{C}_{0.025}$ and $\hat{C}_{0.975}$, were obtained. Assuming symmetry, $\hat{f}(k) = 1/2(-\hat{C}_{0.025} + \hat{C}_{0.975})$ provides an estimate of the 0.975 quantile for any given k. In order to smooth the resulting quantiles, the $\hat{f}(k)$ were regressed on $1/k$, $1/\sqrt{k}$ and $1/\log(k)$ and this regression curve was used to provide the quantiles of M. Brockwell and Gordon propose that these quantiles be used when constructing 95 per cent confidence intervals, rather than the more conventional standard normal quantile of 1.96, and tabulate these values in their Table II. Other than this change in the quantile used, the quantile approximation method adopts the approach of DerSimonian and Laird in its entirety.

3. ADDITIONAL SIMULATION STUDIES AND ANALYTICAL RESULTS

In order to explore the applicability of the quantiles provided, two similar but slightly different simulation studies will be performed. These use exactly the same parameter values as those utilized by Brockwell and Gordon, except that the within-study variances will be truncated to lie in the intervals [0.06, 0.6] and [0.0009, 0.6], respectively. The first scenario is similar to the one explored previously, but with the important difference that large (small within-study variance) trials are not observed. Meta-analyses based on several relatively small trials are commonplace, and is the
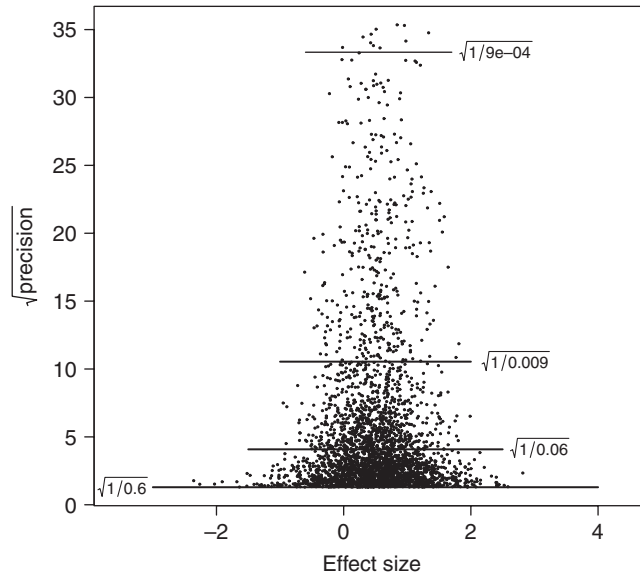
Figure 1. Funnel plot showing the trial sizes permitted by three alternative simulation studies.

scenario, which this simulation study attempts to imitate. Since the σ_i^2 are simulated for every single trial in the simulation study, $475\,000k$ within-study variances are needed for each of $k=2,3,\ldots,30$, and a very large number of within-study variances is needed in total. The variances were simulated directly from a scaled χ_1^2 distribution and values that lay out of the allowable range of $[0.06, 0.6]$ were discarded. If the range were restricted further then it would be considerably more efficient to simulate directly from the distribution of the within-study variances.

The second scenario is also similar to that assumed by Brockwell and Gordon but permits very large 'mega-trials', which may occur in situations where such trials have been performed where these are deemed necessary in order to try to obtain a definitive estimate of treatment effect.

The funnel plot shown in Figure 1 illustrates the types of trials permitted by these simulation studies. Here 3000 within-study variances have been simulated from the scaled χ^2 distribution described above, but truncated to lie within $[0.0008, 0.6]$, from which indicative y_i have been simulated from $y_i \sim N(0.5, \sigma_i^2 + 0.2)$. Here the precision is defined as $1/\sigma_i^2 = w_i(0)$ and the square-root of the precision is plotted against y_i. The first simulation study performed here only permits small trials that lie beneath the horizontal line labelled $\sqrt{1/0.06}$, whereas the simulation study performed by Brockwell and Gordon permit those that lie beneath $\sqrt{1/0.009}$. Finally the second simulation performed here permits trials that lie under $\sqrt{1/0.0009}$ and allows very large mega trials as shown in Figure 1. The within-study variances used in Figure 1 were truncated to lie within $[0.0008, 0.6]$ to emphasize that even the second simulation study described above does not permit trials of inordinate size, and the axes chosen were adopted to adequately display the variation in the trial sizes.

The results from the simulation studies are shown in Figure 2. For the first simulation study the plotted points, shown as triangles, are the simulated $\hat{f}(k)$ and the corresponding curve shows the fitted regression of $\hat{f}(k)$ on k; the fitted model is $1.906 + 0.304/k + 0.774/\sqrt{k} - 0.095/\log(k)$.
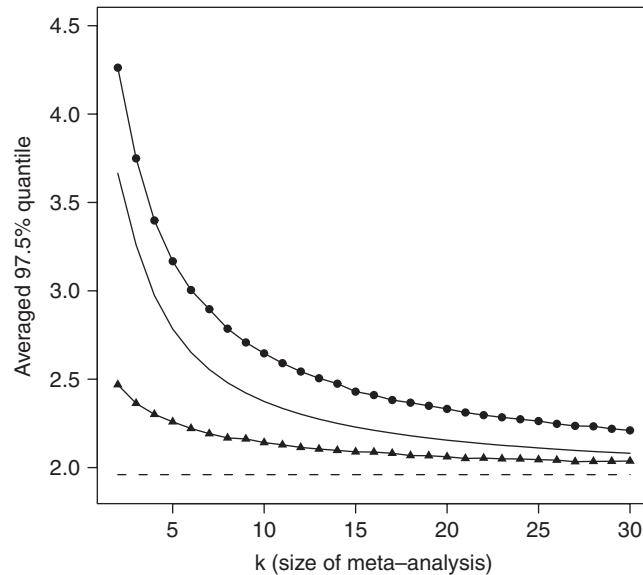
Figure 2. 97.5 per cent quantiles of M from the simulation studies. Triangles show the results for the first scenario, where within-study variances lie in the interval $[0.06, 0.6]$; circles show the results for the second scenario, where the within-study variances lie in the interval $[0.0009, 0.6]$. The corresponding solid curves show the fitted regressions and the other solid curve follows the quantiles from the quantile approximation method. The dashed line shows the standard normal quantile, for comparison.

The results from the second simulation study are similarly shown in Figure 2 using circles as plotting points. Here the fitted model is $1.936+4.089/k+2.069/\sqrt{k}-0.819/\log(k)$ and much larger quantiles result. The other solid line follows the quantiles given by Brockwell and Gordon in their Table II, and the dashed line shows the standard normal quantile of 1.96 for comparison.

It is clear from Figure 2 that the distribution used for the within-study variances can have considerable implications for the resulting quantiles of M. The asymptotic results that justify the standard procedure ensure that, as $k \rightarrow \infty$, the quantiles tend towards 1.96. Regression equations that have this limit might be considered, as the fitted regression equations do not provide this and should not be extrapolated for much larger values of k. We have not investigated this further, as this exercise merely serves to highlight that alternative quantiles result from different scenarios, and hence no such regression equation can give reliable quantiles for all circumstances.

3.1. Why do the simulation studies provide such different results?

The scenarios examined above were motivated by a consideration of Higgins and Thompson's [11] typical within-study variance, σ_t^2. This is given by

$$\sigma_t^2 = \frac{(k-1)\sum_{i=1}^{k} w_i}{(\sum_{i=1}^{k} w_i)^2 - \sum_{i=1}^{k} w_i^2}$$

This formula applies to a finite sample of within-study variances, but in our simulation studies we assume a population distribution for these. Taking the limit as $k \to \infty$ however provides this population and $\sigma_t^2 \to 1/E[W_i]$, where $E[W_i]$ denotes the expectation of the reciprocal of the within-study variances. If the within-study variances are drawn from $0.25\chi_1^2$, and then truncated to lie within $[a, b]$, then the density of their reciprocal is

$$f(w) = \frac{4\chi_1^2(4/w)}{w^2 f(a, b)}$$

for $w \in [1/b, 1/a]$ where $\chi_1^2(\cdot)$ denotes the probability density function of a χ^2 distribution with one degree of freedom, and $f(a, b)$ denotes the probability that this random variable lies in the interval $[4a, 4b]$. Hence

$$E[W_i] = \int_{1/b}^{1/a} w f(w) \, \mathrm{d}w$$

and the reciprocal of this integral gives the typical within-study variance for the population. For the simulation study conducted by Brockwell and Gordon, this typical within-study variance is 0.056.

The first simulation study however does not permit within-study variances less than 0.06, and the typical within-study variance is 0.160. For the second simulation study the typical variance is a mere 0.017. These very different typical study sizes help to explain the impact of the changes in the procedure and the difficulty in generalising findings from a particular simulation study to meta-analysis more generally, as the results depend on the value of τ^2 in relation to the sizes of the studies, as demonstrated analytically in the next section.

3.2. The analytical form of M for the case where all trials are the same size

In order to further examine distribution of M, we will consider the special case where $\sigma_i^2 = \sigma^2$ for all i, i.e. all trials are of the same 'size'.

Under this simplification we have that the y_i are i.i.d. from a normal distribution and therefore all the usual standard results apply. In particular the distribution of the sample mean of the y_i, $\bar{y}$, and the corresponding sample variance, $s^2 = \sum_{i=1}^k (y_i - \bar{y})^2/(k-1)$, are related to those of well-known distributions and are independently distributed. Furthermore, the untruncated DerSimonian and Laird estimate of τ^2 has the very simple form $\hat{\tau}_u^2 = s^2 - \sigma^2$. As the sample variance s^2 is a continuous random variable, and hence the probability that it takes the value σ^2 is precisely zero, there are therefore two distinct possibilities for the more usual estimate $\hat{\tau}^2$: either this is truncated to zero (if $s^2 < \sigma^2$, so that a fixed effects model is adopted) or positive and equal to $\hat{\tau}_u^2$ (if $s^2 \geqslant \sigma^2$). Let E denote an indicator random variable for the event that a random effects procedure is adopted, i.e. $E = 0$ if $\hat{\tau}^2 = 0$ and $E = 1$ otherwise. We therefore have that the distribution of M is given by

$$f(m) = P(E=0) f(m|E=0) + \{1 - P(E=0)\} f(m|E=1) \tag{1}$$

In order to obtain the distribution of M we evaluate the three expressions on the right hand of (1). These terms may be derived in a very similar manner as shown by Jackson [12], where the

distribution of M, with $\mu = 0$, is derived under the hypothesis $H_0 : \mu = 0$. Defining the incomplete gamma function as

$$\Gamma_1(a, x) = \frac{1}{\Gamma(a)} \int_0^x t^{(a-1)} \exp(-t) \, dt$$

where $\Gamma(\cdot)$ denotes the usual Gamma function, we obtain

$$P(E = 0) = P(s^2 < \sigma^2) = \Gamma_1 \left(\frac{k-1}{2}, \frac{k-1}{2(1 + w\tau^2)} \right) \qquad (2)$$

where $w = \sigma^{-2}$. Conditioning on $E = 0$, and noting that $\sigma_i^2 = \sigma^2$ for all i gives $M = (\bar{y} - \mu)/\sqrt{\sigma^2/k}$. Noting that $E = 0$ is equivalent to the event that $s^2 < \sigma^2$, and that $\bar{y}$ and s^2 are independent, we obtain

$$f(m | E = 0) = \frac{1}{\sqrt{1 + w\tau^2}} \phi \left(\frac{m}{\sqrt{1 + w\tau^2}} \right) \qquad (3)$$

where $\phi(\cdot)$ denotes the standard normal density function. We show in the appendix that

$$f(m | E = 1) = f_{\{t, (k-1)\}}(m) \left[\frac{1 - \Gamma_1 \left\{ \frac{k}{2}, \frac{k-1+m^2}{2(1 + w\tau^2)} \right\}}{1 - \Gamma_1 \left\{ \frac{k-1}{2}, \frac{k-1}{2(1 + w\tau^2)} \right\}} \right] \qquad (4)$$

where $f_{(t, v)}(\cdot)$ denotes the density of Student's t distribution with v degrees of freedom. Substituting (2), (3) and (4) into (1) provides the required density $f(m)$. The density of M depends only on k and $w\tau^2$, and can also be expressed in terms of k and $I^2 = \tau^2/(\sigma^2 + \tau^2) = w\tau^2/(1 + w\tau^2)$, the proportion of the variation provided by between-study variation. Higgins and Thompson [11] define a corresponding statistic, denoted here as $\hat{I}^2$, $\hat{I}^2 = \hat{\tau}^2/(\hat{\sigma}_t^2 + \hat{\tau}^2)$. If $\sigma_i^2 = \sigma^2$ for all i, $\sigma_t^2 = \sigma^2$ and I^2 correspond to the true underlying value that Higgins and Thompson's statistic describes.

As $f(m)$ is an even function, for a fixed k and I^2 the 0.975 quantile of M can be obtained numerically as the root of the equation

$$f(x) = \int_{-x}^{x} f(m) \, dm - 0.95$$

and a contour plot of these quantiles is shown in Figure 3. Note that the usual procedure of truncating $\hat{\tau}^2$, so that this is nonnegative, has been adopted. For the special case considered, if the untruncated version of $\hat{\tau}^2$ were used, M would follow a t distribution and the quantiles tend towards t quantiles as $I^2 \to 1$ in Figure 3. The quantile required is generally greater than the conventional 1.96 and for moderate k is highly sensitive to the value of I^2. If the degree of heterogeneity is mild the appropriate quantile is slightly less than the usual standard normal quantile.

The special case considered in this section illustrates that the appropriate quantile depends on the degree of heterogeneity in relation to the size of the trials. It is therefore at the very least extremely difficult to ascertain an appropriate 0.975 quantile of M for meta-analysis in any degree of generality, which is precisely what the quantile approximation method professes to do.
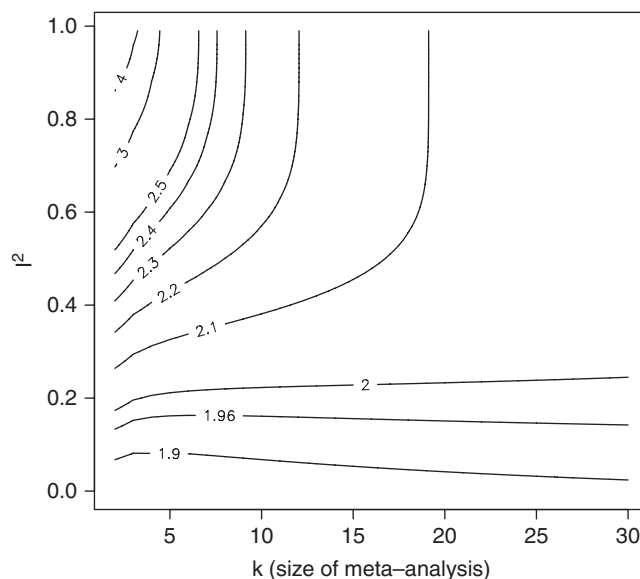
Figure 3. 97.5 per cent quantiles of M assuming that all trials are of the same size.

3.3. A suggested procedure

With the above findings and discussion in mind, we make the following suggestion:

A primary analysis, using the conventional standard normal quantile of 1.96, is appropriate as this value is usual and justified asymptotically. For $k<30$ in particular however, the simulation studies and Figure 3 show that this may be too small in application and hence that nominal 95 per cent confidence intervals can provide somewhat smaller coverage probabilities. In particular, $\hat{I}^2$ can be evaluated as $\hat{\tau}^2/(\sigma_t^2 + \hat{\tau}^2)$ and Figure 3 can then be used as a rough guide to assess if the conventional quantile is likely to be appropriate. We further suggest computing a confidence interval for τ^2 [2, 4, 13, 14], which can easily be transformed to provide a corresponding interval for I^2. A range of plausible values of I^2 may then be considered when inspecting Figure 3 and determining whether or not a secondary or sensitivity analysis is likely to be necessary.

The secondary analysis should involve a simulation study, using the observed within-study variances for the particular meta-analysis in question, and a range of τ^2 that lie in the confidence interval, in order to investigate the appropriate quantile for M. We propose that the largest quantile for M uncovered by this simulation study should be used in the secondary analysis. For example, if a statistically significant result is overturned by this process, then extreme caution needs to be exercised when interpreting the primary analysis.

4. EXAMPLE: GLYCEROL FOR ACUTE STROKE

This example has previously been used by Jackson [15] and Baker and Jackson [16]. It concerns the use of glycerol for preventing death in patients who suffer an acute stoke, and involves nine trials with the results summarized as two by two tables. The data are shown in Table I, where y_i denotes

Table I. The glycerol data.

Trial (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

the log odds ratio of the ith trial and σ_i^2 is the corresponding within-study variance, obtained in the usual way [5]; a negative log odds ratio indicates that the use of glycerol is beneficial.

The DerSimonian and Laird estimate of τ^2 is 0.08. Using this estimate in a conventional random effects meta-analysis as described in Section 1 provides a 95 per cent confidence interval for μ of $(−0.55, 0.22)$. The analysis is therefore inconclusive, but if this is valid it does allow us to infer that the use of glycerol is neither particularly beneficial nor harmful. A 95 per cent confidence interval for τ^2, using Viechtbauer's Q-profile method [14], is $(0, 0.95)$.

The typical within-study variance is 0.25 and a 95 per cent confidence interval for I^2 is $(0, 0.79)$. Since $k=9$, Figure 3 suggests that much larger quantiles than the standard 1.96 are plausible and might be considered in a secondary analysis. Six simulation studies were therefore performed using the empirical distribution of within-study variances, with $\tau^2=0$, 0.2, 0.4, 0.6 and 0.8 and 0.95. These values lie within the confidence interval for τ^2 and cover a wide range of plausible possibilities. For each of these τ^2, 25 000 meta-analyses involving nine trials with the σ_i^2 shown in Table I were simulated from the random effects model $y_i \sim N(\mu, \sigma_i^2 + \tau^2)$; $\mu=0.5$ was again adopted although this choice is immaterial. Following the same procedure as before, M was computed for each simulated meta-analysis and the 0.025 and 0.975 quantiles were used to estimate the 0.975 quantile, resulting in quantiles of 1.853, 2.252, 2.389, 2.410, 2.389 and 2.385, for $\tau^2=0$, 0.2, 0.4, 0.6, 0.8 and 0.95, respectively. A worst case scenario is that a quantile of 2.410, rather than the standard 1.96, should be used when constructing the 95 per cent confidence interval for μ, which means that the usual interval should be widened by around $2.410/1.96 \approx 23$ per cent. This indicates that there is somewhat more uncertainty in the true treatment effect than the usual procedure suggests. For comparison, the quantile approximation method [6] gives the even larger quantile of 2.422, a value that does not seem particularly appropriate for the primary analysis for this particular example.

5. DISCUSSION

The usual random effects model for meta-analysis makes a series of assumptions and approximations. The first of these concerns the conditional normal distribution for $y_i|\mu_i$ with known variance. This is a justifiable approximation provided that the trials are sufficiently large and this has been assumed throughout.

The normal model for the random effect is hard to verify empirically with typically small numbers of trials. In part due to the mathematical tractability of this assumption however, this is a

fairly natural assumption to make. Assuming that the trials are independent is reasonable as they are conducted separately.

The final approximation, and one that is less likely to be appropriate in practice however, is the use of $\tau^2 = \hat{\tau}^2$ in the conventional way. This is only justified as the number of trials becomes large and the distribution of M may not closely follow a standard normal for the small sample sizes frequently encountered in practice. Despite this, under the random effects model, the density of M is not a function of μ: noting that Q is location invariant, $\hat{\tau}^2$ and therefore M can be written as a function of the k variables $\Delta_i = (y_i - \mu) \sim N(0, \sigma_i^2 + \tau^2)$, whose distributions do not depend on μ. Since the number and size of trials are clearly known, the only unknown variable that the distribution of M depends on is τ^2. When assuming particular values of this in the context of a sensitivity analysis, M becomes a pivotal quantity for the remaining parameter μ. Hence M can be used to construct confidence intervals for the overall treatment effect using this pivotal quantity in the way proposed, for a range of values of τ^2, after obtaining the appropriate quantiles approximately from simulation. The dependency of the quantile on τ^2 is mentioned by Brockwell and Gordon, who discuss a quantile approximation function $g(k, \tau^2)$. Our simulation studies highlight that this is also a function of the σ_i^2, and Figure 3 shows that $g(k, \tau^2)$ is sensitive to the value of τ^2. The uncertainty in this value should be taken into account when deciding which quantiles to use.

Figure 3 illustrates the scenarios where the standard normal quantile provides a poor approximation, but is no substitute for obtaining quantiles by simulation, because this figure only takes into account the size of studies and considers a simple special case. An improvement might be to produce a series of figures akin to Figure 3 where both an average and a measure of variability of the study sizes is reflected in the quantiles plotted. These might be obtained from further calculations, if a suitable distribution of study sizes could be found, or more likely from additional simulation studies. Deriving the exact distribution of M, in full generality, remains one of the most important unsolved problems in meta-analysis.

Although the quantile approximation method attempts to take into account the exact distribution of M for finite sample sizes, its 'one size fits all' approach can be misleading. We propose that the type of sensitivity analysis illustrated in Section 4 using the glycerol data be used in its stead. This requires more work and attention to detail, but is relatively straightforward to implement nonetheless.

APPENDIX A: PROOF OF EQUATION (4)

Conditioning on $E = 1$, and noting that $\sigma^2 = \sigma_i^2$ for all i, provides $\hat{\tau}^2 = \hat{\tau}_u^2 = s^2 - \sigma^2$ and $M = (\bar{y} - \mu)/\sqrt{s^2/k}$. A little rearrangement gives

$$M = \frac{\dfrac{\sqrt{k}(\bar{y} - \mu)}{\sqrt{\sigma^2 + \tau^2}}}{\sqrt{\dfrac{s^2(k-1)}{(\sigma^2 + \tau^2)(k-1)}}}$$

$$X_1 = \frac{\sqrt{k}(\bar{y} - \mu)}{\sqrt{\sigma^2 + \tau^2}} \sim N(0,1) \quad \text{and} \quad X_2 = \frac{s^2(k-1)}{\sigma^2 + \tau^2} \sim \chi_{k-1}^2$$

where X_1 and X_2 are independent. Furthermore $M = X_1/\sqrt{X_2/(k-1)}$, conditional on $E = 1$, which is equivalent to the event that $X_2 \geqslant (k-1)/(1+w\tau^2) = c$. The density $f(x_1, x_2 | E = 1)$ is given by $f(x_1)f(x_2)I_{(x_2 \geqslant c)}/P(X_2 \geqslant c)$, where $I_{(x_2 \geqslant c)} = 1$ if $x_2 \geqslant c$ and $I_{(x_2 \geqslant c)} = 0$ otherwise. Furthermore $P(X_2 \geqslant c)$ is given by the complement of equation (2). Changing variables in the usual way with $M = X_1/\sqrt{X_2/(k-1)}$ and $M_2 = X_2$, and integrating out M_2 provides (4).

REFERENCES

1. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**:177–188.
2. 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.
3. Emerson JD, Hoaglin DC, Mosteller F. Simple robust procedures for combining risk differences in sets of 2×2 tables. *Statistics in Medicine* 1996; **15**:1465–1488.
4. Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. *Statistics in Medicine* 1996; **15**:619–629.
5. Sutton AJ, Abrams KR, Jones DR, Sheldon DR, Song F. *Methods for Meta-Analysis in Medical Research*. Wiley: Chichester, 2002.
6. Brockwell SE, Gordon IR. A simple method for inference on an overall effect in meta-analysis. *Statistics in Medicine* 2007; **26**:4531–4543.
7. Follmann DA, Proschan MA. Valid inference in random effects meta-analysis. *Biometrics* 1999; **55**:732–737.
8. Sutton AJ, Higgins JPT. Recent developments in meta-analysis. *Statistics in Medicine* 2008; **27**:625–650.
9. Sidik K, Jonkman JN. Robust variance estimation for random effects meta-analysis. *Computational Statistics and Data Analysis* 2006; **50**:3681–3701.
10. Böckenhoff A, Hartung J. Some corrections of the significance level in meta-analysis. *Biometrical Journal* 1998; **40**:937–947.
11. Higgins JPT, Thompson SG. Quantifying heterogeneity in meta-analysis. *Statistics in Medicine* 2002; **21**: 1539–1558.
12. Jackson D. The significance level of the standard test for a treatment effect in meta-analysis. *Statistics in Biopharmaceutical Research* 2008. Available at: http://www.amstat.org/publications/sbr/index.cfm?fuseaction= jackson.
13. Knapp G, Biggerstaff BJ, Hartung J. Assessing the amount of heterogeneity in random-effects meta-analysis. *Biometrical Journal* 2006; **48**:271–285.
14. Viechtbauer W. Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine* 2007; **26**:37–52.
15. Jackson D. The power of the standard test for the presence of heterogeneity in meta-analysis. *Statistics in Medicine* 2006; **25**:2688–2699.
16. Baker R, Jackson D. Using journal impact factors to correct for the publication bias of medical studies. *Biometrics* 2006; **62**:785–792.