STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:4531–4543
Published online 2 April 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2883

A simple method for inference on an overall effect in meta-analysis

Sarah E. Brockwell[1] and Ian R. Gordon[2,*,†,‡]

[1] *Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA15261, U.S.A.*
[2] *Statistical Consulting Centre, Department of Mathematics and Statistics, University of Melbourne, VIC 3010, Australia*

SUMMARY

The random effects approach in meta-analysis due to DerSimonian and Laird is well established and used pervasively. It has been established by Brockwell and Gordon that this method, when used for confidence intervals, leads to coverage probabilities lower than the nominal value. A number of alternatives have been proposed, but these either have the defect of iterative and complicated calculation, or deficient coverage. In this paper we propose a new approach, which is simple to use, and has coverage probabilities better than the alternatives, based on extensive simulation. We call this approach the 'quantile approximation' method. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: meta-analysis; random effects; coverage properties

1. INTRODUCTION

Meta-analysis is frequently used to estimate the overall effect of an intervention across a number of information sources or studies. In many cases, the methods used to estimate the overall effect are relatively straightforward, both in terms of model structure and computational requirements. A review of published meta-analyses indicates that practising meta-analysts favour such simple methods.

For different meta-analyses, the nature of the effect of interest, and of the information sources themselves, may be very diverse, corresponding to the wide variety of problems to which meta-analyses are now applied. For our purposes, we take an effect to be the difference in efficacy between an intervention and a control and sources of information regarding this effect to be studies with quantitative results, such as clinical trials in medical research. We consider, in particular, meta-analyses in which the intervention effect is measured as a log odds ratio. Such a scenario occurs frequently in practice, particularly in the medical sciences. It is assumed that a suitable estimate

*Correspondence to: Ian R. Gordon, Statistical Consulting Centre, Department of Mathematics and Statistics, The University of Melbourne, VIC 3010, Australia.
†E-mail: i.gordon@ms.unimelb.edu.au
‡Associate Professor.

Received 13 February 2006
Accepted 15 February 2007

of the log odds ratio and a standard error of this estimate are available for each study. Methods for obtaining such estimates are considered in Cox [1].

When the aim of a meta-analysis is to estimate an overall effect, commonly used methods include the fixed effects method of Woolf [2] and the random effects method of DerSimonian and Laird [3]. Selection between these two approaches is often carried out using a simple test of homogeneity [4] of the studies involved (for example, Touloumi *et al.* [5]; Danesh *et al.* [6]). However, it is widely agreed that a random effects model should be routinely adopted [7, 8]. It has been demonstrated [9] that, in the presence of even slight between-study heterogeneity, the fixed effects model results in inferences which substantially underestimate the variation in the data and in parameter estimates.

The random effects model is expressed as follows. Let k denote the number of studies in the meta-analysis, $Y_i, i=1,2,\ldots,k$, the estimator of the intervention effect for study i, and $\hat{\sigma}_i$ the standard error of Y_i, $i=1,2,\ldots,k$. If θ_i is the true intervention effect for study i, then it is generally assumed that

$$Y_i = \theta_i + e_i \quad \text{where } e_i \overset{\mathrm{d}}{=} \mathrm{N}(0, \hat{\sigma}_i^2)$$

$$\theta_i = \mu + \varepsilon_i \quad \text{where } \varepsilon_i \overset{\mathrm{d}}{=} \mathrm{N}(0, \tau^2)$$

for $i=1,2,\ldots,k$, and the e_i and ε_i are assumed to be independent. Note that it is standard practice to assume that a normal distribution is suitable for Y_i, despite the estimate $\hat{\sigma}_i^2$ in $\mathrm{var}(Y_i)$ (see [7, 10]). The parameter of interest is generally μ, the overall measure of intervention effect, and using the assumptions of normality and independence, we can write

$$Y_i = \mu + e_i + \varepsilon_i, \quad \text{so that } Y_i \overset{\mathrm{d}}{=} \mathrm{N}(\mu, \hat{\sigma}_i^2 + \tau^2)$$

The variance of the random effect, τ^2, is a measure of the heterogeneity in the data.

The random effects method of DerSimonian and Laird [3], hereafter abbreviated by DL, is based upon a method of moments estimate of τ^2:

$$\hat{\tau}^2 = \max\left\{0, \frac{1}{W}(Q_{\hat{w}} - (k-1))\right\} = \max\left\{0, \frac{1}{W}\left(\sum \hat{w}_i (Y_i - \hat{\mu})^2 - (k-1)\right)\right\} \tag{1}$$

where $Q_{\hat{w}}$ is the Cochran statistic [4], used to test for study homogeneity; $\hat{w}_i = 1/\hat{\sigma}_i^2$, $\hat{\mu}$ is the fixed effects estimate of μ: $\hat{\mu} = \sum \hat{w}_i Y_i / \sum \hat{w}_i$, and $W = \sum \hat{w}_i - \sum \hat{w}_i^2 / \sum \hat{w}_i$; W is assumed to be constant. The DL estimate of μ is then the weighted average

$$\hat{\mu}_{\hat{\tau}} = \frac{\sum \hat{w}_i(\hat{\tau}) Y_i}{\sum \hat{w}_i(\hat{\tau})} \quad \text{where } \hat{w}_i(\hat{\tau}) = \frac{1}{\hat{\sigma}_i^2 + \hat{\tau}^2}$$

An estimate of the variance of $\hat{\mu}_{\hat{\tau}}$ is given by

$$\widehat{\mathrm{var}}(\hat{\mu}_{\hat{\tau}}) = \frac{1}{\sum \hat{w}_i(\hat{\tau})}$$

Confidence intervals for μ are commonly obtained under the assumption of normality i.e. a 95 per cent confidence interval (CI) is $\hat{\mu}_{\hat{\tau}} \pm 1.96\sqrt{\widehat{\mathrm{var}}(\hat{\mu}_{\hat{\tau}})}$. This method is regularly used to obtain an estimate and confidence interval for the overall effect (see, for example, [11–13]) and requires only that Y_i is approximately normally distributed and that $\hat{\sigma}_i^2$ provides a reasonable estimate of $\mathrm{var}(Y_i)$. Implementation of this method is straightforward.

DOI: 10.1002/sim

However, the DL method, as generally implemented, does not adequately account for estimation error, particularly when the number of studies is relatively small. This has been demonstrated using coverage probabilities estimated from simulated data [9]. Simulation is required because analytic approaches to finding the coverage probability do not provide tractable solutions. The problem of inadequate coverage is due primarily to the use of a normal distribution along with an estimated variance in confidence intervals for μ. The normal approximation results in CIs for μ which are too narrow on average. The situation is analogous to (but not the same as) using a quantile from the normal distribution rather than the t distribution, when carrying out inferences on small samples of normal data.

In this paper, we consider alternatives to the DL random effects method in order to produce better confidence intervals for μ, which can be readily obtained. We extend the method comparison in Brockwell and Gordon [9] by considering an alternative from the literature—an extension to the DL method, developed by Biggerstaff and Tweedie [14]—and briefly discuss several other possible methods. We then propose an alternative method. This method, called the quantile approximation method, is based largely on numeric results, and is aimed, very directly, at obtaining better 95 per cent CIs for μ; that is, intervals with coverage closer to 0.95. The method is developed primarily for meta-analyses in which the effect of intervention is measured as a log odds ratio, and is very simply implemented.

In Section 2 we review the method proposed by Biggerstaff and Tweedie [14] and compare this to the DL method and the likelihood ratio method (hereafter, LR) outlined by Hardy and Thompson [15]. In Section 3, motivation for the quantile approximation method is presented, the method is developed, and then compared to those considered in Section 2. In Section 4, all four methods are applied to a collection of studies on the effect of intravenous magnesium on mortality in suspected acute myocardial infarction. This well-known example appears throughout the meta-analysis literature [16–18]. The practical usefulness and general applicability of the quantile approximation method are discussed in Section 5.

2. THE BIGGERSTAFF AND TWEEDIE METHOD

The DL method does not incorporate the uncertainty associated with the estimation of τ^2 into either the estimation of μ, the standard error of this estimator or into confidence intervals for μ. In an attempt to rectify this, Biggerstaff and Tweedie ([14], hereafter, BT) proposed an alternative to the DL method. This approach uses $\hat{\omega}_i(\hat{\tau}) = E(1/(\hat{\sigma}_i^2 + \hat{\tau}^2))$ in place of $\hat{w}_i(\hat{\tau}) = 1/(\hat{\sigma}_i^2 + \hat{\tau}^2)$ in estimating μ, where only the randomness of $\hat{\tau}^2$ is considered (the $\hat{\sigma}_i^2$ are assumed constant). We briefly review their method below and then compare this to the DL and LR method [15] both of which are reviewed in more detail in Brockwell and Gordon [9].

BT propose that the distribution of $Q_{\hat{w}} = \sum \hat{w}_i(Y_i - \hat{\mu})^2$ be approximated by a gamma distribution, with parameters r and λ specified by

$$r = \frac{E(Q_{\hat{w}})^2}{\text{var}(Q_{\hat{w}})} \quad \text{and} \quad \lambda = \frac{E(Q_{\hat{w}})}{\text{var}(Q_{\hat{w}})}$$

The mean and variance of $Q_{\hat{w}}$, given in BT, are known exactly (ignoring the variation in the $\hat{\sigma}_i^2$) and both are functions of τ^2. Hence, both r and λ must be estimated in practice. BT propose that this be done by substituting $\hat{\tau}^2$ for τ^2 in $E(Q_{\hat{w}})$ and $\text{var}(Q_{\hat{w}})$. Estimates obtained in this fashion are denoted $\hat{r}$ and $\hat{\lambda}$.

DOI: 10.1002/sim

The distribution of $\hat{\tau}^2_m = (Q_{\hat{w}} - (k-1))/W$, the value of $\hat{\tau}^2$ prior to truncation (as given in (1)), is then approximately a scaled, shifted gamma distribution, with density function

$$f(t;\hat{\tau}^2) = \frac{W\hat{\lambda}^{\hat{r}}}{\Gamma(\hat{r})}(Wt+k-1)^{\hat{r}-1}\,\mathrm{e}^{-\hat{\lambda}(Wt+k-1)}I_{[-\frac{(k-1)}{W},\infty)}(t), \quad t\in\mathbb{R}$$

BT propose new weights (used to estimate μ) which are proportional to

$$\hat{\omega}_i(\hat{\tau}) = \widehat{E[\hat{w}_i(\hat{\tau})]} = F(0;\hat{\tau}^2)\hat{w}_i + \int_0^\infty \hat{w}_i(t)\,f(t;\hat{\tau}^2)\,\mathrm{d}t \tag{2}$$

where F denotes the cumulative distribution function corresponding to f. An estimator for μ is then the weighted average

$$\hat{\mu}_{\hat{\omega}} = \frac{\sum\hat{\omega}_i(\hat{\tau})Y_i}{\sum\hat{\omega}_i(\hat{\tau})} \tag{3}$$

with estimated variance

$$\widehat{\mathrm{var}}(\hat{\mu}_{\hat{\omega}}) = \frac{\sum\hat{\omega}_i(\hat{\tau})^2(\hat{\sigma}^2_i+\hat{\tau}^2)}{[\sum\hat{\omega}_i(\hat{\tau})]^2} \tag{4}$$

It is suggested that CIs for μ be constructed assuming normality, giving 95 per cent CIs of the form

$$\hat{\mu}_{\hat{\omega}} \pm 1.96\sqrt{\widehat{\mathrm{var}}(\hat{\mu}_{\hat{\omega}})}$$

Since the distribution of $Q_{\hat{w}}$ depends on τ^2 (and we are ignoring the variation in the $\hat{\sigma}^2_i$), each $\hat{\omega}_i(\hat{\tau})$ is, in practice, only an estimate of the expected value of $(\hat{\sigma}^2_i+\hat{\tau}^2)^{-1}$. Furthermore, the variance of $\hat{\mu}_{\hat{\omega}}$ is unknown and so is estimated by the expression in (4). Hence, like the DL method, the BT approach involves using a normal distribution along with an estimated variance in confidence intervals for μ. The effect of this is considered below. Finally, we note that implementation of the BT method is somewhat more complex, requiring numeric integration to evaluate each $\hat{\omega}_i(\hat{\tau})$, given in (2).

The BT method is compared to both the DL and LR methods by estimating the coverage probabilities of confidence intervals for μ produced by each method. The LR method, derived from standard results for the asymptotic distribution of a likelihood ratio statistic, gives confidence intervals specified by

$$\{\mu_0 : \ln L(\mu_0, \hat{\tau}^2(\mu_0)) > \ln L(\hat{\mu}_{\mathrm{ml}}, \hat{\tau}^2_{\mathrm{ml}}) - \tfrac{1}{2}C_{1-\alpha}(\chi^2_1)\}$$

where $C_{1-\alpha}(\chi^2_1)$ is the $(1-\alpha)$ quantile of the χ^2_1 distribution, and $\hat{\mu}_{\mathrm{ml}}$ and $\hat{\tau}^2_{\mathrm{ml}}$ are the maximum likelihood estimators of μ and τ^2. Details can be found in Hardy and Thompson [15]. Note that $L(\mu_0, \hat{\tau}^2_{\mathrm{ml}}(\mu_0))$, the profile likelihood function for μ_0, is the likelihood function evaluated at μ_0 and $\hat{\tau}^2_{\mathrm{ml}}(\mu_0)$, where $\hat{\tau}^2_{\mathrm{ml}}(\mu_0)$ is the maximum likelihood estimate of τ^2 given $\mu=\mu_0$. Unlike the single parameter likelihood function $L(\mu_0, \hat{\tau}^2_{\mathrm{ml}})$, the profile likelihood function permits some variation in the estimate of τ^2 with varying values of μ_0.

2.1. Data simulation and method comparison

The exact distributions of the CIs considered here are unknown. Hence, the coverage probabilities for a random interval (A, B) for μ, defined as

$$\text{coverage probability} = \Pr(\mu\in(A,B))$$

Table I. Coverage probability summary statistics for the DL, BT, and likelihood ratio (LR) methods.

Method	Mean	Min	MAD	$q_{0.95}$AD	$p(0.01)$	$p(0.02)$	$p(0.03)$	$p(<0.95)$
DL	0.9191	0.8599	0.0300	0.0668	0.0435	0.1897	0.5020	0.9091
BT	0.9204	0.8610	0.0296	0.0661	0.0870	0.2213	0.5217	0.9091
LR	0.9417	0.9146	0.0145	0.0267	0.2372	0.7826	0.9723	0.7826

Each descriptive statistic is obtained from 253 estimated coverage probability values. MAD = median absolute deviation (from 0.95); $q_{0.95}\text{AD} = 0.95$ quantile of the absolute deviations; $p(0.01)$, $p(0.02)$, and $p(0.03)$ denote the proportion of estimated coverages within 0.01, 0.02, 0.03, respectively, of the nominal level; $p(<0.95)$ = proportion of estimated coverages below the nominal level (of 0.95).

must be estimated from simulated data. The coverage probability is estimated by the proportion of simulated confidence intervals containing the true value of the parameter of interest. We consider nominal 95 per cent CIs for μ for the DL, LR, and BT methods, for a range of values of k and τ^2. In all three cases, the coverage probability is not dependent on the value of μ.

Data are simulated to mimic meta-analyses for which the effect of intervention is measured as a log odds ratio. For each simulated meta-analysis we generate realizations $y_i,\ i=1,2,\dots,k$ from a $\mathrm{N}(\mu, \hat{\sigma}_i^2+\tau^2)$ distribution. In all cases we take $\mu = 0.5$. The k values of $\hat{\sigma}_i^2$ are simulated (for each meta-analysis) from a scaled χ_1^2 distribution, then truncated to lie within [0.009, 0.6].§ Using the simulated y_i and $\hat{\sigma}_i^2$, 95 per cent CIs for μ are obtained for each of the three methods considered. This process is repeated 25 000 times and the coverage probability for a given method (for given values of k and τ^2) is estimated as the proportion of intervals (out of 25 000) containing $\mu = 0.5$.

Coverage probabilities for each method are estimated for all integer values of k between 3 and 25 and $\tau^2 = 0, 0.01, \dots, 0.1$. This gives 253 coverage probability estimates for each method. Summary statistics for these estimates are presented in Table I.

A selection of estimated coverage probabilities for each method is shown in Figure 1. Two features are readily apparent. First, there is very little difference between estimated coverages for the BT method and those of the DL method. This is also evident from the very similar summary statistics (given in Table I) observed for these two methods.

Second, coverage probabilities for the LR method are, in general, closer to the nominal level than those of either the DL or BT methods. For example, the mean estimated LR method coverage is 0.942 and 95 per cent of estimated coverages lie within 0.027 of the nominal level. For the BT method, corresponding figures are 0.920 and 0.066, respectively.

It can be shown, assuming that τ^2 is known, that intervals from the DL and LR methods are identical. The difference in estimated coverage probabilities for these two methods can therefore be attributed to the manner in which τ^2 is estimated and the mechanism by which the uncertainty associated with this estimation is incorporated into CIs for μ. Estimated coverages suggest that the LR method does this in a more satisfactory fashion. However, although almost all estimated LR method coverages are within 0.03 of the nominal level, less than a quarter are within 0.01 and, for large τ^2, many are considerably below the nominal level. Furthermore, unlike the DL method, the LR method is not straightforward in implementation, requiring numerical minimization and search routines to obtain maximum likelihood estimates and confidence intervals.

§The scale factor used here is 0.25. This method results in simulated standard errors, $\hat{\sigma}_i$, with median of approximately 0.346 and upper and lower quartiles of 0.518 and 0.212, respectively.

DOI: 10.1002/sim

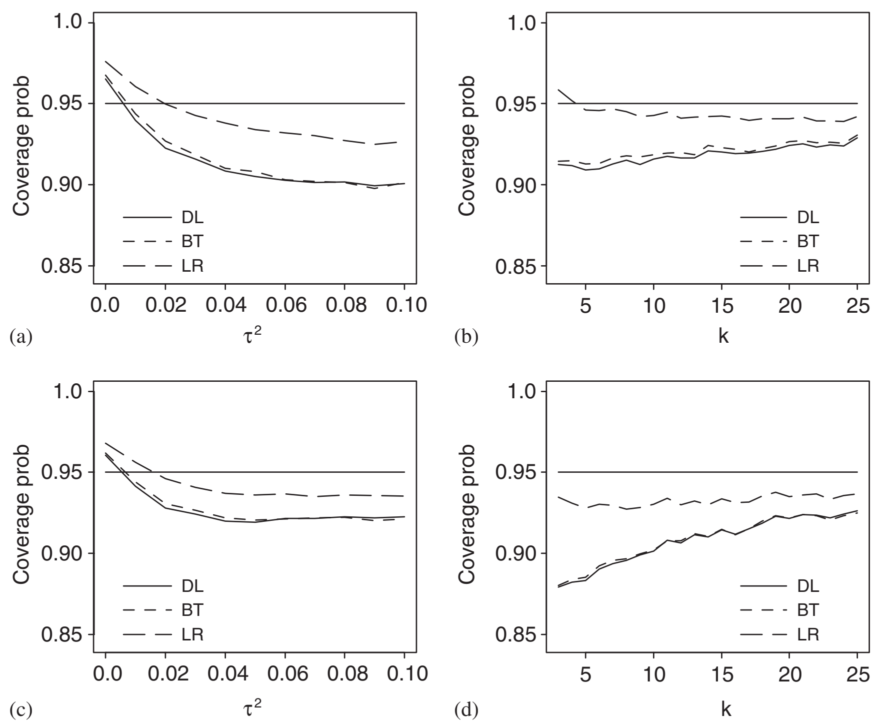

Figure 1. Estimated coverage probabilities for the DerSimonian and Laird (DL), Biggerstaff and Tweedie (BT) and likelihood ratio (LR) methods, for varying k and τ^2. The horizontal line is the nominal level of 0.95: (a) $k=10$; (b) $\tau^2=0.03$; (c) $k=20$; and (d) $\tau^2=0.07$.

The BT method also requires numeric routines to obtain confidence intervals for μ. For this method, estimated coverage probabilities are frequently well below the nominal level of 0.95. This is particularly so when k is small. The low observed coverage probabilities for this method can be attributed to the use of a normal distribution along with an estimated variance in confidence intervals for μ. Hence, the normal approximations used in both the DL and BT methods do not, in general, produce confidence intervals of a sufficiently large average width.

3. THE QUANTILE APPROXIMATION METHOD

Consider CIs obtained using the DL method. In order to improve these intervals, in terms of their coverage, we might attempt to either eliminate the necessity of estimating $\mathrm{var}(\hat{\mu}_{\hat{\tau}})$ or obtain a better approximation to the distribution of the statistic $M=(\hat{\mu}_{\hat{\tau}}-\mu)/\sqrt{\widehat{\mathrm{var}}(\hat{\mu}_{\hat{\tau}})}$. Exact expressions for both $\mathrm{var}(\hat{\mu}_{\hat{\tau}})$ and the cumulative distribution function of $\hat{\mu}_{\hat{\tau}}$ are considered in Sheehan [19]. Both of these expressions are intractable functions of τ^2 and so must be estimated in practice. As a result, exact approaches do not provide a substantial improvement of DL interval coverage probabilities. Furthermore, since these functions are complicated, there is no apparent method of removing their dependence on τ^2 (and hence the necessity for estimation).

DOI: 10.1002/sim

In many standard applications involving normally distributed data, uncertainty associated with the variance estimate is incorporated into CIs *via* a t distribution. This approach does not apply in the usual way here, due to the manner in which both the estimator of μ and its variance depend on τ^2. However, if we assume that $M \overset{\mathrm{d}}{\approx} t_{k-1}$, the resulting CIs for μ have estimated coverages which are an improvement over those of the DL method. Summary statistics for estimated coverages (obtained using data simulated as described in the next section) are presented in Table III. Of the 9 per cent of estimated t-distribution coverages outside the range [0.92, 0.98], all of those observed are greater than 0.98. These values occur for small k, indicating that in many cases, confidence intervals are considerably wider than required. Estimated t distribution method coverages are below the nominal level for larger values of k.

In general, CIs derived without foundation in proven distribution theory are not optimal. However, in some cases, such theory does not provide suitable intervals and alternative methods must be considered. The t distribution method is one such alternative. We propose a second such alternative in an attempt to produce confidence intervals whose coverage is better than those of the t-distribution, DL, or LR methods.

Both exact approaches and commonly used approximating methods have failed to provide a satisfactory expression for (or approximation to) the distribution of $M=(\hat{\mu}_{\hat{\tau}}-\mu)/\sqrt{\widehat{\mathrm{var}}(\hat{\mu}_{\hat{\tau}})}$. However, knowing the distribution of M, and in particular the appropriate quantiles of this distribution, would enable a meta-analyst to obtain confidence intervals for μ with coverages close to the nominal level. We propose a method of approximating the quantiles of M required to obtain a 95 per cent CI for μ. The resulting approximate quantile values can then be used (along with $\hat{\mu}_{\hat{\tau}}$ and $\widehat{\mathrm{var}}(\hat{\mu}_{\hat{\tau}})$) to obtain intervals for μ. This method is simple to implement and, as is demonstrated below, produces intervals with coverage probabilities closer to the nominal value than those of DL, t-distribution, or LR methods. The method used to obtain quantile approximations is outlined below.

3.1. Derivation of the quantile approximation method

In the absence of a good algebraic approximation to the distribution of M, we obtain estimates of the quantiles of this distribution using numerical techniques. We consider in particular the 0.025 and 0.975 quantiles, required for a 95 per cent CI for μ. Estimates of these quantiles are obtained *via* a function b_k (which approximates the 0.975 quantile of M), providing an alternative multiplier for symmetric CIs of the form: $\hat{\mu}_{\hat{\tau}} \pm b_k\sqrt{\widehat{\mathrm{var}}(\hat{\mu}_{\hat{\tau}})}$. The quantile approximation function is a monotonically decreasing non-linear function of k. The form of b_k is derived by simulating samples of M and modelling the resulting sample quantiles using standard regression techniques.

Samples of M are generated for all values of k between 2 and 30. This range of k is likely to include the scope of almost all meta-analyses. Here we use a wider range of τ^2 than that used previously, to better reflect values likely to be found in practice, and to explore the validity of the QA method over a wider range of τ^2. The values used were $\tau^2=0, 0.01, \ldots, 0.10, 0.15, \ldots, 0.5$. For each (k, τ^2) combination we simulate 25 000 meta-analyses. As mentioned previously, we simulate data under the random effects model with $\mu=0.5$ and $\hat{\sigma}_i^2 \in [0.009, 0.6]$. Recall that these parameter values, and the values of τ^2, are selected to correspond to typical scenarios for estimating a log odds ratio. The value of M is evaluated for each simulated meta-analysis, giving a sample of 475 000 realizations for each k.

From each sample we obtain both the 0.025 and 0.975 sample quantiles, denoted by $\hat{C}_{0.025}$ and $\hat{C}_{0.975}$, respectively. We assume that the distribution of M is symmetric about zero. This is

DOI: 10.1002/sim

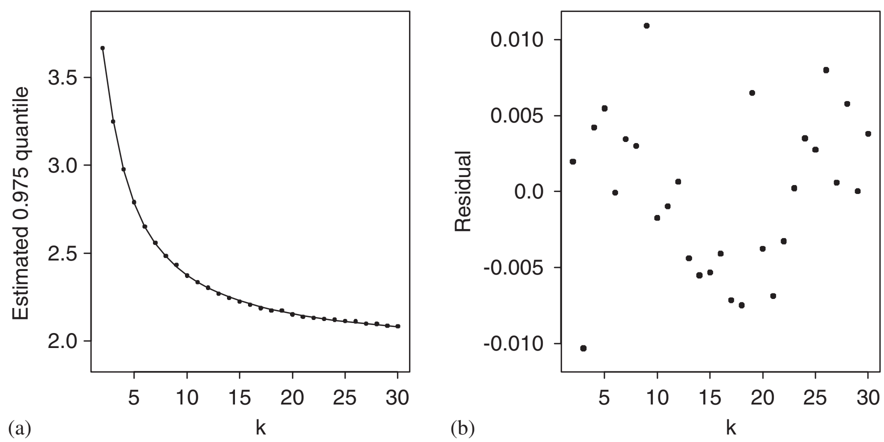

Figure 2. Observed quantile values and the approximation function b_k. Plot (a): observed quantile values and the fitted quantile approximation function $b_k \approx C_{0.975}$. Plot (b): residuals for the regression used to obtain b_k. These are plotted against k.

supported by the simulation results, which show no significant difference between $-\hat{C}_{0.025}$ and $\hat{C}_{0.975}$. So, one function, b_k, is sufficient; $C_{0.975}(M) \approx b_k$ and $C_{0.025}(M) \approx -b_k$.

In order to obtain the function b_k, we use 'observed values' $\hat{f}_k = \frac{1}{2}(-\hat{C}_{0.025} + \hat{C}_{0.975})$. The function b_k is obtained by regressing the 29 observed values $\hat{f}(k)$ onto three variables, each a simple function of k. The variables included in the regression were selected after consideration of the pattern of observed quantiles (shown in Figure 2) and the desired form of the function—namely that it decreases in a non-linear fashion as k increases. The resulting regression equation is

$$b_k = 2.061 + \frac{4.902}{k} + \frac{0.756}{\sqrt{k}} - \frac{0.958}{\ln(k)}, \quad k = 2, 3, \ldots, 30 \tag{5}$$

with a regression R^2 of 0.9998 and a residual standard deviation of 0.0055. Figure 2 contains a plot of both the observed and predicted quantile values as well as a plot of the regression residuals. As expected, for all k in the range considered, b_k is greater than the corresponding standard normal distribution quantile, 1.96. Values of b_k for k between 2 and 30, obtained by evaluation of (5), are given in Table II.

Using b_k and $-b_k$ to approximate the 0.975 and 0.025 quantiles of M, we propose an interval for μ of the form

$$\hat{\mu}_{\hat{\tau}} \pm b_k \sqrt{\widehat{\text{var}}(\hat{\mu}_{\hat{\tau}})} \tag{6}$$

The estimates of μ and $\text{var}(\hat{\mu}_{\hat{\tau}})$ in (6) are those proposed by DL. We denote this method of obtaining intervals for μ as the 'quantile approximation' method. Like the DL method, quantile approximation intervals are very simple to obtain, and hence have a considerable practical advantage over other alternatives, such as the LR or BT methods.

3.2. Method comparison

We again compare CIs using coverage probabilities estimated from data simulated under the random effects model. Coverage probabilities are estimated for $k = 2, 3, \ldots, 30$ and $\tau^2 = 0, 0.01, \ldots, 0.10$,

DOI: 10.1002/sim

Table II. Values of the quantile approximation function for all values of k between 2 and 30.

k	b_k	k	b_k	k	b_k
		11	2.335	21	2.145
2	3.665	12	2.302	22	2.135
3	3.260	13	2.274	23	2.126
4	2.974	14	2.250	24	2.119
5	2.784	15	2.229	25	2.111
6	2.652	16	2.211	26	2.104
7	2.555	17	2.195	27	2.097
8	2.480	18	2.180	28	2.091
9	2.422	19	2.167	29	2.086
10	2.374	20	2.155	30	2.081

$0.15, \ldots, 0.5$. For each (k, τ^2) combination, the coverage probability is estimated as the proportion of intervals, out of 25 000, containing $\mu = 0.5$. Data for each meta-analysis are simulated as described in Section 2.1, reflecting what might typically occur in practice when estimating a log odds ratio.

A selection of estimated coverage probabilities for the DL, LR, and QA methods are presented in Figure 3. Summary statistics for the 551 coverage probabilities estimated for each method are given in Table III.

These results indicate that coverage probabilities for the QA method are a substantial improvement on those of the DL method. For example, the mean estimated QA coverage is 0.95, whilst that for the DL method is 0.91. Furthermore, an estimated 79 per cent of QA coverages are within 0.01 of the nominal level and 93 per cent within 0.03. Corresponding figures for the DL method are 2 and 55 per cent . The DL coverages are always less than the QA ones and the QA coverages are almost always closer to the nominal value than the DL ones. This implies that QA method intervals have better average width.

The differences between QA and LR method coverage probabilities are not quite so substantial, but a considerably higher proportion of estimated QA coverages are very close to the nominal level: 79 per cent compared to 27 per cent. These results indicate that, in the great majority of cases, QA method intervals are better than LR intervals, and they are considerably simpler to calculate.

The statistics in Table III indicate that both the t distribution and QA methods have mean coverage of approximately 0.95. However, the spread of estimated t distribution method coverage probabilities is markedly greater than that of QA method coverages. For example, an estimated 42 per cent of t distribution coverage probabilities are within 0.01 of the nominal level, as opposed to 79 per cent of QA coverages. These observations therefore suggest that b_k better approximates the relevant quantiles of M than corresponding t_{k-1} distribution values.

A comparison of values of b_k, given in Table II, and 0.975 quantiles of the t_{k-1} distribution indicates that $b_k < C_{0.975}(t_{k-1})$ for $k \leqslant 4$ and $b_k > C_{0.975}(t_{k-1})$ otherwise. Estimated t distribution method coverage probabilities are, in general, well above the nominal level for small k; note, in particular, that $C_{0.975}(t_1) = 12.706$. In such cases, the QA method quantile is smaller and estimated coverages closer to the nominal level, indicating that b_k better approximates the relevant quantile of M. Similarly, for large k, estimated t distribution coverages are below the nominal level.

DOI: 10.1002/sim

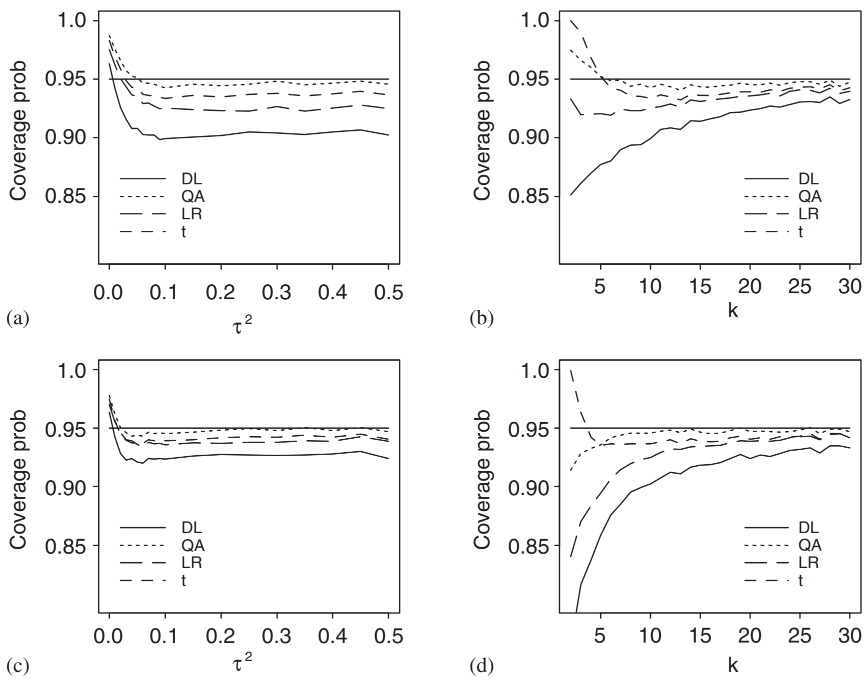

Figure 3. Estimated coverage probabilities for the DerSimonian and Laird, quantile approximation (QA) and likelihood ratio (LR) methods, for varying k and τ^2. The horizontal line is the nominal level of 0.95: (a) $k = 10$; (b) $\tau^2 = 0.10$; (c) $k = 20$; and (d) $\tau^2 = 0.50$.

Table III. Coverage probability summary statistics for the DL, likelihood ratio (LR), and quantile approximation (QA) and t distribution (t) methods.

Method	Mean	Min	MAD	$q_{0.95}$AD	$p(0.01)$	$p(0.02)$	$p(0.03)$	$p(<0.95)$
DL	0.9144	0.7639	0.0281	0.0897	0.0236	0.2396	0.5481	0.9474
LR	0.9360	0.8401	0.0142	0.0460	0.2650	0.7132	0.8838	0.8657
QA	0.9514	0.9142	0.0050	0.0337	0.7895	0.8693	0.9347	0.7151
t	0.9481	0.9306	0.0110	0.0448	0.4156	0.8639	0.9147	0.7713

MAD = median absolute deviation (from 0.95); $q_{0.95}$AD = 0.95 quantile of the absolute deviations; $p(0.01)$, $p(0.02)$, and $p(0.03)$ denote the proportion of estimated coverages within 0.01, 0.02, 0.03, respectively, of the nominal level; $p(< 0.95)$ = proportion of estimated coverages below the nominal level (of 0.95).

By using a slightly larger quantile value in this range, QA intervals have, in general, better coverage probabilities.

4. AN EXAMPLE: MAGNESIUM AND ACUTE MYOCARDIAL INFARCTION

A famous example in the debate about the merits of meta-analysis relates to the treatment of suspected acute myocardial infarction with intravenous magnesium. There were seven small

DOI: 10.1002/sim

Table IV. Meta-analysis estimates and 95 per cent confidence intervals for the overall effect of intravenous magnesium on mortality in suspected acute myocardial infarction.

Method	Estimated odds ratio	95 per cent CI	Estimate of τ^2
FE	0.471	(0.280, 0.791)	—
DL	0.448	(0.233, 0.861)	0.171
QA	0.448	(0.204, 0.985)	0.171
LR	0.449	(0.192, 0.903)	0.162

randomized trials carried out, which were used by Teo *et al*. [16] in an overview. The estimated odds ratio for short-term mortality was 0.45, with a 95 per cent confidence interval of (0.28, 0.71), showing 'strong evidence' of the benefits of the treatment.

Subsequently, the very large ISIS 4 trial of the same treatment showed an odds ratio of 1.06 (95 per cent confidence interval: 1.00, 1.12), and a controversy about the earlier meta-analysis ensued [17, 18]. A number of possible reasons for the apparent anomaly between the two results were canvassed. However, statistical issues were largely dismissed. Egger and Smith commented that 'a few [meta-analyses] will produce misleading results by chance alone, though this is unlikely in the present case.' Presumably, this comment was based on the upper limit of the 95 per cent CI for the pooled odds ratio being considerably below 1; the p-value for the usual null hypothesis was given as $p<0.001$.

The original analysis by Teo *et al*. used Peto's fixed effect method. The result differed somewhat from the usual fixed effect method, which gives an odds ratio of 0.47 and a 95 per cent CI of (0.28, 0.79) for these data, but both have upper confidence limits which is well below one. The DL interval is wider, because there is modest evidence of between study heterogeneity ($Q=7.57$, cf. χ^2_6); its upper bound is 0.86. The use of this interval would probably still lead to the same qualitative conclusion as that of Teo *et al*.

The LR method also gives an upper limit that is markedly below one—see Table IV for details.

On the other hand, the QA method gives a 95 per cent CI of (0.20, 0.98). While this interval also has an upper bound below one, we suggest that this interval, had it been obtained in the original meta-analysis, would have led to far more circumspect conclusions at the time, and less subsequent controversy about the difference between the meta-analysis and the mega-trial.

5. COMMENTS

Several features of the QA method require some discussion. First, plots (a) and (c) of Figure 3 indicate that, particularly for small k, the coverage probabilities of all three methods depend on τ^2. In the case of the QA method this could, in theory, be remedied by allowing the quantile approximation function f to depend on both k and τ^2. This approach has been considered, but found unsuitable for several reasons. We discuss these below.

It is possible to derive a quantile approximation function $g(k, \tau^2)$ by regressing sample quantiles of M onto variables which are non-linear functions of k and τ^2. If $g(k, \tau^2)$ is then used to obtain approximate quantile values, the resulting CIs have coverage probabilities very close to the nominal level. In practice, however, τ^2 is unknown and hence $g(k, \tau^2)$ must be estimated, possibly by using $\hat{\tau}^2$ in place of τ^2. This estimation introduces error into the quantile approximation function g, thereby

DOI: 10.1002/sim

affecting the resulting coverage probability values. Not surprisingly, the magnitude of this error varies with k—since the sampling error in the estimation of τ^2 decreases as k increases. The error introduced into g is also dependent on τ^2, with resulting coverage probabilities decreasing as τ^2 increases, similar to the dependence on τ^2 evident in Figure 3. For intervals using either b_k or $g(k, \hat{\tau}^2)$, the variation across values of τ^2 is greatest for small k.

These observations suggest that little real improvement is made, in practice, by incorporating τ^2 into the quantile approximation function, and by avoiding this, the method is far simpler. Similar problems arise if the approximation function is derived directly as a function of $\hat{\tau}^2$. In light of this we have proposed a quantile approximation function which depends only on k; k, after all, is known. This implies that coverage probabilities for the resulting confidence intervals vary across τ^2, particularly for small k. In this they are not alone, however, as is demonstrated in Figure 3.

The quantile approximation function is obtained from data simulated to mimic quite specific types of meta-analyses. Several issues therefore arise concerning the validity and general applicability of this method. As for all results obtained using numerical methods, the QA method is, to some degree, subject to particularities of the data simulation techniques. In this case, these simulation techniques influence both the derivation and assessment of the method. However, the range of parameter values considered is broad, and satisfactorily covers what might be observed in practice for meta-analyses where the overall intervention effect is measured as a log odds ratio. Like all of the methods considered here, the validity of the QA method relies on one or more assumptions. For the QA method, we assert that the range of parameter values considered represents what might occur in practice.

Given the frequency with which binary data arise and are summarized by a log odds ratio, the QA method is of considerable use in practice and, for meta-analyses similar to those simulated here, it provides CIs with better width than either the DL or LR methods. While it is not firmly established that the QA method is suitable for meta-analyses on other scales, the development here relies on a structure which is not specific to log odds ratios, since the essence is estimators that are approximately normally distributed. It seems likely that the superiority of the QA method over the DL method, clearly demonstrated for log odds ratios, translates to inference on other parameters for which inferences are sometimes made using meta-analysis.

Sidik and Jonkman [20] proposed a method based on the use of t distribution quantiles, which is different from the t based method considered in this paper. A direct comparison is possible with our approach, since they used our settings for simulation. Their method is an improvement on DL, but the quantile approximation method outperforms their approach, as is clearly evident from a comparison of their Figure 1 with our Figure 3. As they note, this is attributable to the use of $\hat{\tau}^2$, without adequate allowance for the uncertainty in its estimation. Since it is by no means clear how to make such allowance analytically, we have adopted the approach of estimating an appropriate adjustment. One further note to make about any t based intervals is that they will clearly be too wide when k is very small, and especially when $k=2$. It might be said that the use of meta-analysis to summarize information from only two or three studies is stretching its applicability, but the widespread availability of the meta-analysis package RevMan [21], in particular, has meant that such meta-analyses are frequently conducted, sometimes as a sub-analysis in a more substantive study.

The QA method is proposed in order to address the need for a satisfactory approximation to the distribution of $M = (\hat{\mu}_{\hat{\tau}} - \mu)/\sqrt{\widehat{\mathrm{var}}(\hat{\mu}_{\hat{\tau}})}$. This method provides an approximation to particular quantiles of this distribution, assuming that the distribution is symmetric about zero. Until and unless an algebraic expression for the distribution of M is identified, the QA method enables the

DOI: 10.1002/sim

meta-analyst to obtain confidence intervals for μ which better reflect the true distribution of M and account for the estimation of $\hat{\tau}^2$. The method is very easy to apply, and provides better CI coverage than alternatives.

ACKNOWLEDGEMENTS

The authors are very grateful to Associate Professor Ray Watson, Department of Mathematics and Statistics, University of Melbourne, for much helpful advice and encouragement.

REFERENCES

1. Cox DR. *The Analysis of Binary Data*. Chapman & Hall: London, 1970.
2. Woolf B. On estimating the relation between blood group and disease. *Annals of Human Genetics* 1955; **19**:251–253.
3. DerSimonian R, Laird NM. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**:177–188.
4. Cochran WG. Problems arising in the analysis of a series of similar experiments. *Journal of the Royal Statistical Society* 1937; **4**(Suppl.):102–118.
5. Touloumi G, Katsouyanni K, Zmirou D, Schwartz J, Spix C *et al*. Short term effects of ambient oxidant exposure on mortality: a combined analysis within the APHEA project. *American Journal of Epidemiology* 1997; **146**:177–185.
6. Danesh J, Collins R, Appleby P, Peto R. Association of fibrinogen, C-reactive protein, albumin, or leukocyte count with coronary heart disease. *JAMA* 1998; **279**:1477–1482.
7. Draper D, Graver DP, Goel PK, Greenhouse JB, Hedges LV, Morris CN, Tucker JR, Waternaux CM. *Combining Information*: *Statistical Issues and Opportunities for Research.* National Academy Press: Washington, 1992.
8. Mengersen KL, Tweedie RL, Biggerstaff BJ. The impact of method choice in meta-analysis. *Australian Journal of Statistics* 1995; **37**:19–44.
9. Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. *Statistics in Medicine* 2001; **20**:825–840.
10. Fleiss JL. The statistical basis of meta-analysis. *Statistical Methods in Medical Research* 1993; **2**:121–145.
11. Choma D, Daures JP, Quantin X, Pujol JL. Aneuploidy and prognosis of non-small-cell lung cancer: a meta-analysis of published data. *British Journal of Cancer* 2001; **85**:14–22.
12. Camma C, Giunta M, Fiorica F, Pagliaro L, Craxi A, Cottone M. Preoperative radiotherapy for resectable rectal cancer: a meta-analysis. *JAMA* 2000; **284**:1008–1015.
13. SanGiovanni JP, Parra-Cabrera S, Colditz GA, Berkey CS, Dwyer JT. Meta-analysis of dietary essential fatty acids and long-chain polyunsaturated fatty acids as they relate to the visual resolution acuity in healthy preterm infants. *Pediatrics* 2000; **105**:1292–1298.
14. 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.
15. Hardy RJ, Thompson SG. Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine* 1998; **17**:841–856.
16. Teo KK, Yusuf S, Collins R, Held PH, Peto R. Effects of intravenous magnesium in suspected acute myocardial infarction: overview of randomised trials. *British Medical Journal* 1991; **303**:1499–1503.
17. Yusuf S, Flather M. Magnesium in acute myocardial infarction. *British Medical Journal* 1995; **310**:751–752.
18. Egger M, Smith GD. Misleading meta-analysis. *British Medical Journal* 1995; **310**:752–754.
19. Sheehan SE. An evaluation of practical methods for meta-analysis. *Ph.D. Thesis*, The University of Melbourne, Victoria, Australia, 2001.
20. Sidik K, Jonkman JN. A simple confidence interval for meta-analysis. *Statistics in Medicine* 2002; **21**:3153–3159.
21. Review Manager (RevMan) [Computer program]. Version 4.2 for Windows. The Cochrane Collaboration: Oxford, England, 2003.