STATISTICS IN MEDICINE
Statist. Med. 2001; **20**:825–840

A comparison of statistical methods for meta-analysis

Sarah E. Brockwell and Ian R. Gordon*,†

Department of Mathematics and Statistics, Richard Berry Building, The University of Melbourne, Victoria 3010, Australia

SUMMARY

Meta-analysis may be used to estimate an overall effect across a number of similar studies. A number of statistical techniques are currently used to combine individual study results. The simplest of these is based on a fixed effects model, which assumes the true effect is the same for all studies. A random effects model, however, allows the true effect to vary across studies, with the mean true effect the parameter of interest. We consider three methods currently used for estimation within the framework of a random effects model, and illustrate them by applying each method to a collection of six studies on the effect of aspirin after myocardial infarction. These methods are compared using estimated coverage probabilities of confidence intervals for the overall effect. The techniques considered all generally have coverages below the nominal level, and in particular it is shown that the commonly used DerSimonian and Laird method does not adequately reflect the error associated with parameter estimation, especially when the number of studies is small. Copyright © 2001 John Wiley & Sons, Ltd.

1. INTRODUCTION

Meta-analysis refers to the process of locating, selecting, assessing and combining information relevant to a particular research question. Over the past 20 years the number of published meta-analyses and discussions on meta-analysis methodology has dramatically increased. This has occurred particularly in the areas of medical and epidemiological research [1], but also in the sociological and behavioural sciences [2]. A Medline search found 269 meta-analyses published in 1990. This figure has risen steadily, to 575 in 1997. Owing to this rapid rise in the popularity of meta-analysis, it is becoming increasingly important that the methodology and statistics used are sound.

In general it is instructive to identify the sources of heterogeneity between studies, possibly by modelling the outcome of interest in terms of features of the studies ('meta-regression'). However, in many meta-analyses the number of studies is small and such an approach is not feasible. This paper focuses on the non-Bayesian statistical methods commonly used for meta-analysis, when the goal of the analysis is to estimate an effect from a relatively small number of similar studies. Our results suggest that, as usually applied, these methods have important deficiencies. In particular,

*Correspondence to: Ian R. Gordon, Department of Mathematics and Statistics, Richard Berry Building, The University of Melbourne, Victoria 3010, Australia
†E-mail: i.gordon@ms.unimelb.edu.au

Received September 1998
Accepted February 2000

confidence intervals obtained from the combined information do not adequately account for the variation introduced both from the data, and the estimation procedure itself.

The statistical methods are generally based on standard fixed or random effects models. These are outlined briefly below, and the random effects model is discussed in more detail in the following two sections.

Consider a collection of k studies, the ith of which has estimated effect size Y_i and true effect size θ_i. A general model is then specified by

$$Y_i = \theta_i + e_i \quad \text{where } e_i \stackrel{\mathrm{d}}{=} \mathrm{N}(0, \sigma_i^2), \quad i = 1, 2, \ldots, k$$

The e_i indicate random deviations from the true effect size and are assumed independent with mean zero and variance σ_i^2. This implies that the estimated effect size Y_i is normally distributed with mean θ_i and variance σ_i^2. Y_i can be any measure of effect, provided the assumption of normality is (at least approximately) appropriate. Common examples are a log-odds ratio or difference in means.

In general the parameter of interest is the overall effect, denoted by μ. The fixed effects model assumes $\theta_i = \mu$ for $i = 1, 2, \ldots, k$, implying that each study in the meta-analysis has the same underlying effect. Note that even if the θ_i are assumed to be the same, the Y_i are not identically distributed due to the possibility of differing σ_i^2. The estimator of μ is generally a simple weighted average of the Y_i, with the optimal weights proportional to $w_i = 1/\mathrm{var}(Y_i)$. In practice the variances are not known so estimated variances $\hat{\sigma}_i^2$ are used to estimate both μ and $\mathrm{var}(\hat{\mu})$. Any effect of this is generally ignored in practice, but to indicate this estimation we use the notation $\hat{\sigma}_i^2$ throughout. Hence we define $\hat{w}_i = 1/\hat{\sigma}_i^2$ giving

$$\hat{\mu} = \frac{\sum \hat{w}_i Y_i}{\sum \hat{w}_i} = \sum \frac{Y_i}{\hat{\sigma}_i^2} \Big/ \sum \frac{1}{\hat{\sigma}_i^2} \quad \text{and} \quad \widehat{\mathrm{var}}(\hat{\mu}) = 1 \Big/ \sum \frac{1}{\hat{\sigma}_i^2}$$

In contrast to the fixed effects model, the random effects model does not assume that the θ_i are equal, but that they are normally distributed. This gives the two-stage model

$$\left.\begin{aligned} Y_i &= \theta_i + e_i \quad \text{where } e_i \stackrel{\mathrm{d}}{=} \mathrm{N}(0, \hat{\sigma}_i^2) \\ \theta_i &= \mu + \varepsilon_i \quad \text{where } \varepsilon_i \stackrel{\mathrm{d}}{=} \mathrm{N}(0, \tau^2) \end{aligned}\right\} \tag{1}$$

The error terms e_i and ε_i are assumed to be independent. In this case, the true effect for study i is centred around the overall effect, allowing individual studies to vary both in estimated effect and true effect. The random effects variance parameter τ^2 is a measure of the heterogeneity between studies. Note that the fixed effects model is a special case of the random effects model, with $\tau^2 = 0$.

It is generally agreed [3, 4] that in the presence of heterogeneity, the random effects model should be used. Heterogeneity is commonly tested using a statistic defined by Cochran [5]: $Q_w = \sum w_i (Y_i - \hat{\mu})^2$. The null hypothesis of homogeneity is $H_0 : \tau^2 = 0$ against a one-sided alternative. If we assume the variances σ_i^2 are known, then under H_0, $Q_w \stackrel{\mathrm{d}}{=} \chi_{k-1}^2$. In practice, estimates $\hat{w}_i = 1/\hat{\sigma}_i^2$ are used giving the statistic

$$Q_{\hat{w}} = \sum \hat{w}_i (Y_i - \hat{\mu})^2$$

which is approximately χ_{k-1}^2 under H_0. A large value of $Q_{\hat{w}}$ indicates large study-to-study variation, and the hypothesis of homogeneity is rejected. In such cases the random effects model is generally

Table I. Columns 2–5 show the sample sizes and observed proportions dying for the six studies on the effect of aspirin after myocardial infarction. Columns 6 and 7 give the observed log-odds ratio and estimated variances for each of the studies. For a discussion of the fixed effects and DerSimonian and Laird random effects weights see Section 5.

Study	Treatment		Control		$\log(\mathrm{OR}_i)$	$\hat{\sigma}_i^2$	Weights	
	n_{ti}	$\hat{p}_{ti}$	n_{ci}	$\hat{p}_{ci}$			FE*	RE*
1	615	0.0797	624	0.1074	−0.3289	0.0389	0.11	0.15
2	758	0.0580	771	0.0830	−0.3845	0.0412	0.10	0.14
3	317	0.0852	309	0.1054	−0.2158	0.0753	0.05	0.09
4	832	0.1226	850	0.1482	−0.2196	0.0205	0.20	0.20
5	810	0.1049	406	0.1281	−0.2257	0.0352	0.12	0.15
6	2267	0.1085	2257	0.0970	0.1246	0.0096	0.43	0.26

* Fixed effects weights $\dfrac{\hat{\sigma}_i^{-2}}{\sum \hat{\sigma}_i^{-2}}$; random effects weights $\dfrac{(\hat{\sigma}_i^2 + \hat{\tau}^2)^{-1}}{\sum (\hat{\sigma}_i^2 + \hat{\tau}^2)^{-1}}$.

adopted. Hardy and Thompson [6] suggest, however, that the power of this test can be low. They therefore recommend that this should not be the only means by which the fixed effects model is rejected.

The random effects model is outlined and discussed in detail in Sections 2 and 3. In particular we consider methods for estimating μ and τ^2, and for incorporating $\hat{\tau}^2$ into $\hat{\mu}_{\hat{\tau}}$ – the random effects estimator of μ – and the variance of $\hat{\mu}_{\hat{\tau}}$. In Section 2 we outline the commonly-used DerSimonian and Laird [7] random effects method, and in Section 3, likelihood techniques.

In Section 4, confidence intervals for μ, calculated using several methods, are compared using coverage probabilities. We first compare the fixed effects method and DerSimonian and Laird random effects method, where these are used irrespective of the observed value of $Q_{\hat{w}}$. These are also compared to a Q-based method which reflects the common practice of using the χ^2-test to determine whether a fixed or random effects approach should be adopted. We also compare the DerSimonian and Laird intervals to two likelihood based intervals.

In Section 5 four methods are applied to a collection of studies on the effect of aspirin after myocardial infarction. This set of studies appears repeatedly in the literature [4, 7, 8], owing to its interesting range of observed measures of effect and sample sizes.

The collection consists of six studies, each examining the effect of aspirin after myocardial infarction. In each study the number of patients who died after having been given either aspirin or a control drug is recorded. Sample sizes for all the studies are quite large – the smallest involving 626 patients. Of the six studies however, one is particularly large, involving a total of 4524 patients. Table I gives the sample sizes, n_{ti} and n_{ci} and the proportions dying $\hat{p}_{ti}$ and $\hat{p}_{ci}$ for each of the treatment (t) and control (c) groups, where i denotes the study number. As shown, the large sixth study is the only one for which $\hat{p}_t > \hat{p}_c$.

Table I also gives the observed log-odds ratios ($\log(\mathrm{OR})$) and corresponding estimated variances where

$$\mathrm{OR}_i = \frac{\hat{p}_{ti}(1-\hat{p}_{ci})}{(1-\hat{p}_{ti})\hat{p}_{ci}} \quad \text{and} \quad \hat{\sigma}_i^2 = \widehat{\mathrm{var}}[\log(\mathrm{OR}_i)] = \frac{1}{x_{ti}} + \frac{1}{n_{ti}-x_{ti}} + \frac{1}{x_{ci}} + \frac{1}{n_{ci}-x_{ci}}$$

and x_{ti} and x_{ci} denote the observed number of deaths for the treatment and control groups respectively for study i. From this table we can see that for the sixth study the log-odds ratio is

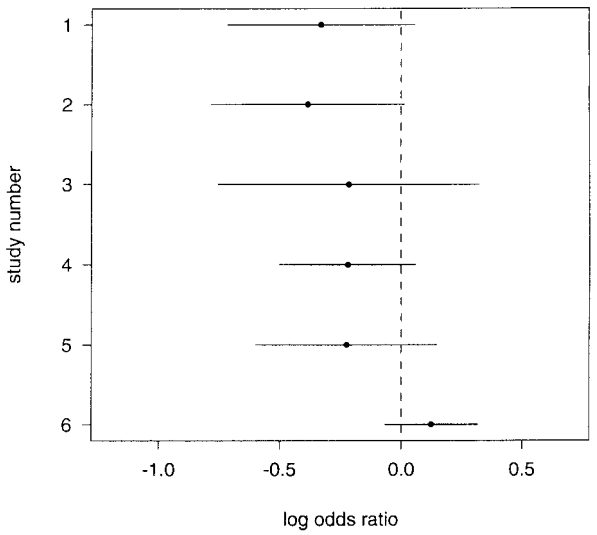

Figure 1. Estimated log-odds ratios and confidence intervals for the six studies used in the aspirin meta-analysis.

considerably different from that of the remaining five studies. This table also shows the effect of the large differences in sample sizes with the estimated variance for the sixth study being considerably smaller than for the other five studies. Figure 1 shows the estimated log-odds ratios and corresponding confidence intervals (using a normal approximation) for each of the six studies.

The results of combining these studies, using four different methods, are presented and discussed in Section 5. In Section 6 some conclusions are drawn and alternative methods are briefly discussed.

2. THE RANDOM EFFECTS MODEL AND ESTIMATION OF τ^2

The random effects model given in (1) can also be written

$$Y_i = \mu + \varepsilon_i + e_i \quad \text{where } e_i \overset{\text{d}}{=} \text{N}(0, \hat{\sigma}_i^2) \quad \text{and} \quad \varepsilon_i \overset{\text{d}}{=} \text{N}(0, \tau^2)$$

relating the Y_i directly to the overall measure of effect μ. By the independence of ε_i and e_i we then have $Y_i \overset{\text{d}}{=} \text{N}(\mu, \hat{\sigma}_i^2 + \tau^2)$. A weighted average is again used to estimate μ, giving

$$\hat{\mu}_\tau = \frac{\sum \hat{w}_i(\tau) Y_i}{\sum \hat{w}_i(\tau)}$$

with variance

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

where $\hat{w}_i(\tau) = [\tau^2 + \hat{w}_i^{-1}]^{-1}$ and $\hat{w}_i$ are as defined above. Assuming τ^2 is known, we then have

$$\hat{\mu}_\tau \stackrel{\mathrm{d}}{=} \mathrm{N}\left(\mu, \frac{1}{\sum \hat{w}_i(\tau)}\right)$$

Note that $\hat{w}_i(\tau) \leqslant \hat{w}_i$. This implies $\mathrm{var}(\hat{\mu}_\tau) \geqslant \mathrm{var}(\hat{\mu})$, and hence random effects model confidence intervals for μ are generally wider than those constructed from the fixed effects model.

In practice, τ^2 is unknown. The most commonly used estimator of τ^2 is a method of moments based estimator proposed by DerSimonian and Laird [7], derived by equating an estimate of the expected value of $Q_{\hat{w}}$ with its observed value. Note that

$$E(Q_{\hat{w}}) = k - 1 + \tau^2\left(\sum \hat{w}_i - \frac{\sum \hat{w}_i^2}{\sum \hat{w}_i}\right)$$

Suppose that t is obtained by solving

$$q_{\hat{w}} = k - 1 + t\left(\sum \hat{w}_i - \frac{\sum \hat{w}_i^2}{\sum \hat{w}_i}\right) \quad \text{giving } t = \frac{q_{\hat{w}} - (k-1)}{\sum \hat{w}_i - \sum \hat{w}_i^2 / \sum \hat{w}_i}$$

It is possible that this value is negative, which is unacceptable as a value for τ^2, so we define

$$\hat{\tau}^2 = \begin{cases} t & \text{if } t > 0 \\ 0 & \text{if } t \leqslant 0 \end{cases}$$

Note that due to the truncation, $\hat{\tau}^2$ is a biased estimator for τ^2.

DerSimonian and Laird proposed that this estimate can then be incorporated into the random effects weights giving

$$\hat{w}_i(\hat{\tau}) = (\hat{\tau}^2 + \hat{\sigma}_i^2)^{-1}$$

An estimator of μ is then given by

$$\hat{\mu}_{\hat{\tau}} = \frac{\sum \hat{w}_i(\hat{\tau}) Y_i}{\sum \hat{w}_i(\hat{\tau})}$$

with variance estimated by

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

Note that this is simply a straight substitution of $\hat{\tau}^2$ into the variance of $\hat{\mu}_\tau$, derived assuming τ^2 is known. In obtaining confidence intervals for μ using $\hat{\mu}_{\hat{\tau}}$ and $\widehat{\mathrm{var}}(\hat{\mu}_{\hat{\tau}})$, it is common practice to maintain the assumption of normality for $\hat{\mu}_{\hat{\tau}}$, despite the use of $\hat{\tau}^2$ and $\hat{\sigma}_i^2$ in place of τ^2 and σ_i^2, respectively.

Two main issues arise from the general application of the random effects model as described above:

(i) The assumption of normality poses problems, first in its validity, and secondly in our ability to check that validity for meta-analyses based on a small number of studies. In particular, the assumption of normally distributed random effects, or between study errors ε_i is not easily verified or justified. The issue of validating the assumption of normality is addressed by Hardy and Thompson [6], however they consider only relatively large values of k.

(ii) The variation between true study effect sizes is taken into account via the inclusion of errors with variance τ^2. It is however only an estimate of this variance which is added into the weights, and the model takes no account of the uncertainty associated with this estimate. In particular the distribution used for $\hat{\mu}_{\hat{\tau}}$ is not altered. As shown in Section 4.2, this results in confidence intervals for μ which are narrower on average than they should be. It is common practice to use a t-distribution to account for the error associated with a variance estimate [9]. This approach is not valid in the random effects meta-analysis context.

3. LIKELIHOOD METHODS

Maximum likelihood theory is widely used for estimation and inference. In this section we review two methods which use established likelihood theory to obtain confidence intervals for μ. These are considered as alternatives to the DerSimonian and Laird method.

3.1. Estimating μ and τ^2 using maximum likelihood

Recall that the standard random effects model has $Y_i \overset{\mathrm{d}}{=} \mathrm{N}(\mu, \hat{\sigma}_i^2 + \tau^2)$, $i = 1, 2, \ldots, k$ and that the $\hat{\sigma}_i^2$ are treated as known constants. The log-likelihood function is

$$\log L(\mu, \tau^2) = -\frac{1}{2}\sum \log(2\pi(\hat{\sigma}_i^2 + \tau^2)) - \frac{1}{2}\sum \frac{(y_i - \mu)^2}{\hat{\sigma}_i^2 + \tau^2}, \quad \mu \in \mathbb{R},\ \tau^2 \geqslant 0 \tag{2}$$

Maximum likelihood estimates $\hat{\mu}_{\mathrm{ml}}$ and $\hat{\tau}^2_{\mathrm{ml}}$ can be found in standard ways (see Appendix A for details).

One major advantage of maximum likelihood estimation lies in the large body of asymptotic theory existing for such estimators. In regular cases a maximum likelihood estimator from a sample of k independent and identically distributed random variables has a normal distribution as $k \to \infty$. It should be noted that the k variables Y_i, $i = 1, 2, \ldots, k$ from a meta-analysis are independent but not identically distributed, since $\mathrm{var}(Y_i) = \hat{\sigma}_i^2 + \tau^2$. In any realistic meta-analysis with large k, the standard assumptions will still apply however.

Using this asymptotic distribution, it is possible to construct a confidence interval for μ. However, this is only an approximate interval, since the asymptotic variance of $\hat{\mu}_{\mathrm{ml}}$ depends on the unknown τ^2. This variance is derived from the covariance matrix of $(\hat{\mu}_{\mathrm{ml}}, \hat{\tau}^2_{\mathrm{ml}})$ and is given by

$$\mathrm{var}(\hat{\mu}_{\mathrm{ml}}) = \frac{1}{\sum (\hat{\sigma}_i^2 + \tau^2)^{-1}}$$

Under the assumption of asymptotic normality we therefore have

$$\hat{\mu}_{\mathrm{ml}} \overset{\mathrm{d}}{\sim} \mathrm{N}\left(\mu, \frac{1}{\sum (\hat{\sigma}_i^2 + \tau^2)^{-1}}\right) \quad \text{as } k \to \infty$$

This distribution is used for $\hat{\mu}_{\mathrm{ml}}$ even though the likelihood estimate of τ^2 may lie on the boundary of the parameter space, namely, $\tau^2 = 0$.

Note that the variance of $\hat{\mu}_{\mathrm{ml}}$ is of the same form as that for the DerSimonian and Laird random effects method. In this case $\mathrm{var}(\hat{\mu}_{\mathrm{ml}})$ is estimated using $\hat{\tau}^2_{\mathrm{ml}}$ without any modification to

the distribution of $\hat{\mu}_{\mathrm{ml}}$; 95 per cent confidence intervals are therefore

$$\hat{\mu}_{\mathrm{ml}} \pm \frac{1.96}{\sqrt{\{\sum(\hat{\sigma}_i^2 + \hat{\tau}_{\mathrm{ml}}^2)^{-1}\}}}$$

This method is referred to as the simple likelihood method.

3.2. Profile likelihood intervals

An alternative method, which uses a profile likelihood function, is proposed by Hardy and Thompson [10]. Unlike the simple likelihood method this method allows for asymmetric intervals and some imprecision in the estimate of τ^2.

The profile likelihood function for μ is defined as

$$L_\mu(\mu_0) = L(\mu_0, \hat{\tau}^2(\mu_0))$$

where $\hat{\tau}^2(\mu_0)$ satisfies

$$\hat{\tau}^2(\mu_0) = \sum\left[\frac{(y_i - \mu_0)^2 - \hat{\sigma}_i^2}{(\hat{\sigma}_i^2 + \hat{\tau}^2(\mu_0))^2}\right] \Big/ \sum\frac{1}{(\hat{\sigma}_i^2 + \hat{\tau}^2(\mu_0))^2} \tag{3}$$

Clearly $\hat{\tau}^2(\mu_0)$ is not assumed fixed for all μ_0.

An approximate 95 per cent confidence interval for μ is then given by values of μ_0 which satisfy

$$\log(L_\mu(\mu_0)) > \log(L(\hat{\mu}_{\mathrm{ml}}, \hat{\tau}_{\mathrm{ml}}^2)) - \tfrac{1}{2}C_{0.95}(\chi_1^2) \tag{4}$$

where $C_\gamma(\chi_1^2)$ is the γ-quantile of the χ_1^2 distribution. Details of the derivation and implementation of this method are given in Appendix B.

Biggerstaff and Tweedie [11] use a similar method to find a confidence interval for τ^2, but as the focus of this paper is finding confidence intervals for μ, intervals for τ^2 are not considered in detail here.

4. COMPARISON OF METHODS

4.1. Coverage probabilities and simulation methods

The coverage probability of a random interval (A, B) for μ is defined as $\Pr(\mu \in (A, B))$ which – for a nominal 95 per cent confidence interval – should be close to 0.95. The exact coverage can only be found if the distribution of the interval is known. If, however, as is more common, the distribution is unknown, the coverage probability must be estimated using simulation. This is done by simulating a large number of meta-analyses and for each meta-analysis calculating the appropriate confidence interval. The estimated coverage probability is then the proportion of these intervals which contain μ.

The coverage probability is usually dependent on the parameters of the model and so the coverages presented below are estimated for a range of values of τ^2 and k. For all the intervals considered here, the coverage probability is not dependent on the value of μ since the procedure is invariant with respect to a location shift. The data for each meta-analysis are simulated using the random effects model described in Section 1, assuming normal errors e_i and ε_i, with zero mean and variances $\hat{\sigma}_i^2$ and τ^2, respectively. For all simulations we use $\mu = 0.5$.

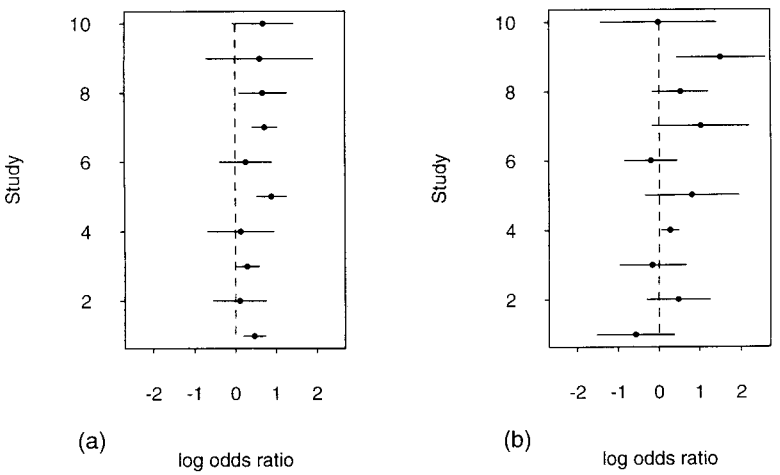

Figure 2. Simulated meta-analyses with $\mu = 0.5$ and $k = 10$. In graph (a) $\tau^2 = 0.03$, a small amount of between study variation. In graph (b) $\tau^2 = 0.07$; the larger amount of between-study variation manifests itself in a larger spread of estimates. The width of the confidence intervals for μ indicate the range of values of $\hat{\sigma}_i^2$ used.

The coverage probability is then estimated by simulating 25000 meta-analyses of k studies, with $\mu = 0.5$ and τ^2 as specified. A 95 per cent confidence interval is then calculated for each of these meta-analyses and the coverage is estimated as the proportion, out of 25000, which contain the parameter value, $\mu = 0.5$. Since the true coverages are generally greater than 0.9, the standard errors of the estimated coverage probabilities are essentially $\leqslant 0.002$.

To give the simulations authenticity, parameter values are chosen to correspond to a typical scenario for estimating a common log-odds ratio (that is, $\exp(\mu)$ is the parameter of interest). Accordingly, the Y_i may be considered sample log-odds ratios, and the $\hat{\sigma}_i^2$ their estimated variances. These variances are realizations from a χ_1^2 distribution, multiplied by 0.25 and then restricted to lie within the interval $(0.009, 0.6)$. Values produced in this way are consistent with the typical distribution of $\hat{\sigma}_i^2$ for log-odds ratios in practice. The $\hat{\sigma}_i^2$ are varied for each of the 25000 simulations in order to allow for the sampling error in these values, which although present in practice, is not accounted for in the methods considered.

The simulation procedure described above is implemented for a given k and τ^2. The procedure is repeated for each of 11 values of τ^2 between 0.0 and 0.1, and values of k between 3 and 35. Figure 2 shows two simulated meta-analyses with $\tau^2 = 0.03$ and $\tau^2 = 0.07$; in each case $\mu = 0.5$ and $k = 10$. Such graphs were in part used to determine suitable values for τ^2 and the $\hat{\sigma}_i^2$.

The simulations are implemented using a program written in C++, with each simulation generating k observations from normal distributions with mean μ and variance $\hat{\sigma}_i^2 + \tau^2$. The data are then used to calculate the fixed effects estimate of μ, the observed value of $Q_{\hat{w}}$, the DerSimonian and Laird estimate of τ^2 and the corresponding confidence intervals for μ. Since in practice the outcome of the χ^2-test using $Q_{\hat{w}}$ is often used to determine which method shall be used, a combined fixed effects/random effects method is considered here. This method reflects the χ^2-test procedure, selecting the fixed or random effects methods as determined by comparing the observed value of $Q_{\hat{w}}$ with the 0.95 quantile of the χ_{k-1}^2 distribution. This is referred to as the Q-based method.

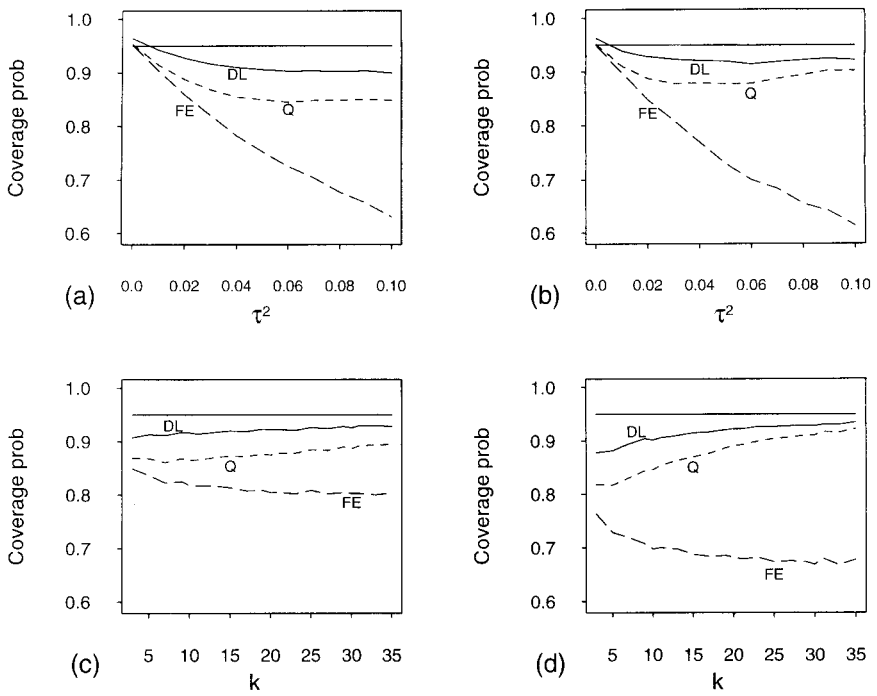

Figure 3. Estimated coverage probabilities for the fixed effects (FE), DerSimonian and Laird random effects (DL) and Q-based (Q) methods, for varying k and τ^2. In graph (a) $k = 10$ and in graph (b) $k = 20$. In graph (c) $\tau^2 = 0.03$ and in graph (d) $\tau^2 = 0.07$.

Estimated coverage probabilities for the fixed effects method, the Q-based method, the DerSimonian and Laird random effects method, the simple likelihood method and the profile likelihood method are presented in the following section.

4.2. Simulation results

The results of the simulations are presented in two figures showing the estimated coverage probabilities for different values of k and τ^2. Figure 3 shows the estimated coverage probabilities for the fixed effects method (FE). Here the fixed effects method is used, despite the data being simulated with $\tau^2 > 0$ in most cases. For $\tau^2 = 0$ the estimated coverage probability is close to 0.95 as expected, however increasing τ^2 substantially reduces the coverage. This reflects the theoretical coverage probability result for the fixed effects model specified by

$$\text{coverage probability} = \Phi(z) - \Phi(-z) \quad \text{where } z = \frac{1.96}{\sqrt{\left(1 + \tau^2 \dfrac{\sum \hat{w}_i^2}{\sum \hat{w}_i}\right)}} \tag{5}$$

and Φ denotes the standard normal cumulative distribution function. A brief derivation of this result is given in Appendix C. Clearly this coverage probability decreases as τ^2 increases.

In practice it is more common for the fixed effects method to be used only after the χ^2-test, based on $Q_{\hat{w}}$, suggests that the hypothesis $H_0 : \tau^2 = 0$ should not be rejected. However if this test is

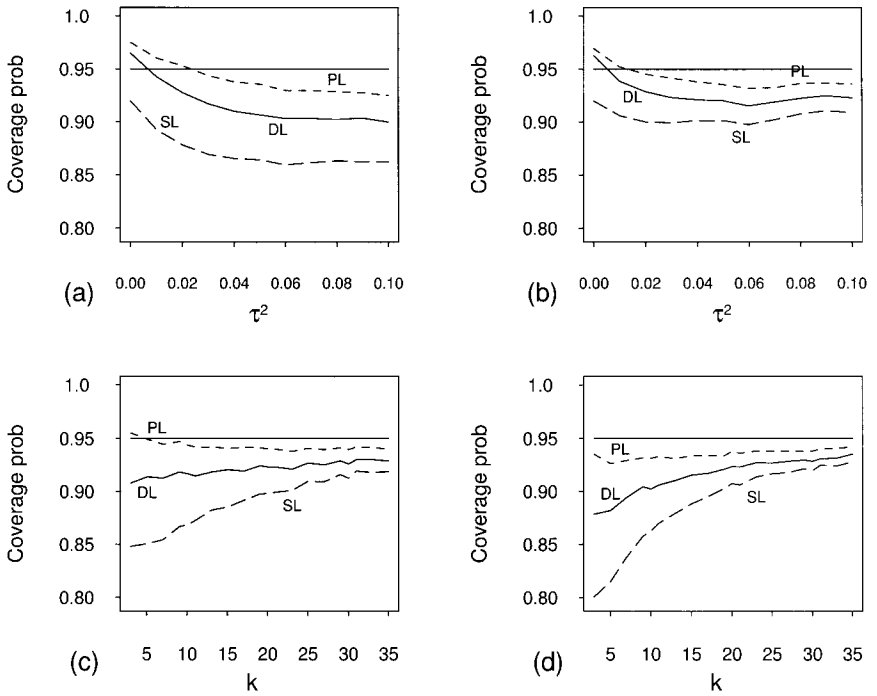

Figure 4. Coverages for the DerSimonian and Laird random effects (DL), profile likelihood (PL) and simple likelihood (SL) methods, for varying k and τ^2. Note that the scale differs from that in Figure 3. In graph (a) $k = 10$ and in graph (b) $k = 20$. In graph (c) $\tau^2 = 0.03$ and in graph (d) $\tau^2 = 0.07$.

used to select an appropriate method, the fixed effects method is selected a considerable proportion of the time, even when the true value of τ^2 is large. For example, with $k = 10$ and $\tau^2 = 0.06$, the fixed effects method is selected an estimated 58 per cent of the time. This percentage decreases as k increases, with the fixed method selected in 26 per cent of the cases with $k = 25$ and $\tau^2 = 0.06$.

Coverage probabilities for the Q-based method are also given in Figure 3. Not surprisingly, these coverage probabilities lie between the fixed effects and DerSimonian and Laird random effects coverages (denoted DL in Figure 3). These latter coverage probabilities are for confidence intervals for μ constructed using the DerSimonian and Laird method in all cases, irrespective of the value of $Q_{\hat{w}}$. As indicated in Figure 3, any use of the fixed effects method substantially reduces the coverage probability.

Figure 4 shows estimated coverage probabilities for the two random effects, likelihood based methods. Coverage probabilities for the DerSimonian and Laird method are also included for comparison. Note that the scale differs from that of Figure 3. These plots indicate that, except when τ^2 is close to zero, the coverage probabilities for all three methods are below 0.95. For all values of k and τ^2 the coverage probabilities for the simple likelihood method are below those of the other two methods. Similarly the profile likelihood method consistently has the highest estimated coverage probability. For the two cases where data are simulated with $\tau^2 = 0$, all three methods have coverage probabilities above 0.95, some significantly so. For these simulations we obtain an estimated value $\hat{\tau}^2 \geqslant 0$, resulting in wider confidence intervals and hence coverage probabilities higher than the nominal value.

Table II. Estimated overall log-odds ratios and corresponding confidence intervals for four different meta-analyses of the effect of aspirin after myocardial infarction.

Method	$\hat{\mu}$	95 per cent CI	$\hat{\tau}^2$
Fixed effects	-0.1015	$(-0.2269, 0.0238)$	$\tau^2 = 0$ (assumed)
DerSimonian and Laird	-0.1689	$(-0.3609, 0.0231)$	$\hat{\tau}^2 = 0.0269$
Profile likelihood	-0.1175	$(-0.3696, 0.0352)$	$\hat{\tau}^2_{\mathrm{ml}} = 0.0390$
Simple likelihood	-0.1175	$(-0.3902, 0.0352)$	$\hat{\tau}^2_{\mathrm{ml}} = 0.0390$

The DerSimonian and Laird method for establishing confidence intervals for μ is the most commonly used technique, yet estimated values indicate that even for a large number of studies the confidence intervals obtained have a coverage probability below 0.95. This suggests that the use of $\hat{\tau}^2$ in the estimation of μ and the standard error of $\hat{\mu}_{\hat{\tau}}$, combined with the use of a normal approximation for $\hat{\mu}_{\hat{\tau}}$, produce intervals which are on average too narrow. This problem also arises when using the simple likelihood method. The standard error of $\hat{\mu}_{\mathrm{ml}}$ is estimated using $\hat{\tau}^2_{\mathrm{ml}}$, without modification to the assumed distribution for $\hat{\mu}_{\mathrm{ml}}$. In both cases, coverage probabilities estimated using the true τ^2 are close to the nominal value.

5. AN EXAMPLE: ASPIRIN AFTER MYOCARDIAL INFARCTION

In this section, the four methods compared in Section 4 are applied to six studies on the effect of aspirin after myocardial infarction reviewed in Section 1. Recall that one of the six studies is considerably larger than the others and is the only study with an odds ratio greater than one.

The homogeneity statistic for these data is $q_{\hat{w}} = 9.88$ on 5 degrees of freedom, giving $P = 0.08$ for the hypotheses $H_0 : \tau^2 = 0$ versus $H_1 : \tau^2 > 0$. Whilst the null hypothesis is not rejected at the 0.05 level, the test does bring into question the validity of the fixed effects model. This is brought about solely by the large sixth study, since the homogeneity statistic for the first five studies is $q_{\hat{w}} = 0.63$ on 4 degrees of freedom, giving $P > 0.9$. Adopting the random effects model for all six studies, we obtain $\hat{\tau}^2 = 0.0269$, using the DerSimonian and Laird estimator.

The fixed effects method and three random effects methods have all been used to combine these data. Table II gives estimated values of μ, the overall effect of aspirin versus the control, and τ^2; 95 per cent confidence intervals for μ are also presented. The three random effects estimates of μ are all considerably smaller than that obtained for the fixed effects model. This is due to the reduction in the weight ascribed to the large sixth study, which has a positive log-odds ratio.

All of the confidence intervals obtained include zero, and the upper bounds for each are relatively similar. The fixed effects confidence interval is considerably narrower than the three random effects based methods. A rough estimate of the fixed effects coverage probability for these data can be obtained from the theoretical coverage given in Section 4.2. Using $\hat{\tau}^2 = 0.0269$ and the estimated values of σ_i^2 given in Table I we obtain

$$z = \frac{1.96}{\sqrt{(1 + 0.0269 \dfrac{\sum (\hat{\sigma}_i^{-2})^2}{\sum \hat{\sigma}_i^{-2}})}} = 1.193$$

and estimated coverage probability

$$\Phi(z) - \Phi(-z) = 0.77$$

well below the nominal level of 0.95.

The markedly lower bound for the three random effects based methods are due to the allowance of a non-zero value for τ^2 in the model and the corresponding change in weights. Table I shows the weights allocated to each study, using both the fixed and random effects models. As can be seen in Table I, the fixed effects model assigns a large weight to studies with small variance, and small weights to those with large variance – generally corresponding to smaller sample sizes. In the random effects model, however, the weights tend to be relatively similar, and the sixth study now counts for only 26 per cent of the total weight, as opposed to 43 per cent in the fixed effects model. In general, in the random effects approach, large studies are downweighted relative to the fixed effects model, and smaller studies given a greater weighting in estimation.

These tendencies produce interesting results for the aspirin meta-analysis. It is generally expected that making adjustments to the weights does not substantially affect the estimate of μ, giving $\hat{\mu} \approx \hat{\mu}_{\hat{\tau}}$. Similarly, since $\hat{\sigma}_i^{-2} = \hat{w}_i > \hat{w}_i(\hat{\tau}) = (\hat{\sigma}_i^2 + \hat{\tau}^2)^{-1}$ we have $\mathrm{SE}(\hat{\mu}) < \mathrm{SE}(\hat{\mu}_{\hat{\tau}})$, where $\mathrm{SE}(\hat{\mu})$ is the standard error of $\hat{\mu}$. The expected result is therefore that

$$|z_{\mathrm{FE}}| = \left| \frac{\hat{\mu}}{\mathrm{SE}(\hat{\mu})} \right| > \left| \frac{\hat{\mu}_{\hat{\tau}}}{\mathrm{SE}(\hat{\mu}_{\hat{\tau}})} \right| = |z_{\mathrm{RE}}|$$

For the aspirin meta-analysis, however, we observe the reverse effect. Although $\mathrm{SE}(\hat{\mu}) = 0.0640 < 0.0980 = \mathrm{SE}(\hat{\mu}_{\hat{\tau}})$, this difference is negated by the change in the estimates of μ: $\hat{\mu} = -0.1015$ and $\hat{\mu}_{\hat{\tau}} = -0.1689$. We then have $|z_{\mathrm{FE}}| = 1.587 < 1.724 = |z_{\mathrm{RE}}|$. Again this effect is a result of the change in weights allocated to the sixth study.

This outcome is interesting, since to some extent it suggests a problem with the way in which the random effects model attempts to increase the uncertainty associated with model estimates. By adding extra variation into the model, we expect to make it more difficult to reject a null hypothesis such as $H_0 : \mu = 0$, and on average this is true. As shown by the aspirin meta-analysis, however, it can occur that the random effects model produces a larger z-statistic, and hence a smaller P-value.

6. COMMENTS

As demonstrated by the coverage probabilities presented in Section 4, the fixed effects method does not perform well unless there is very little between-study variation. In practice this would rarely be the case, and should not be assumed to be so. It has also been demonstrated that use of the χ^2-test to determine which model is appropriate leads to confidence intervals which are on average too narrow. For small k in particular, the fixed effects method is frequently selected, even when τ^2 is large. As shown, this substantially reduces the coverage probability of confidence intervals for μ. It is therefore recommended that the random effects model be adopted irrespective of the outcome of the χ^2-test for heterogeneity. This then simplifies to the fixed effects method only when $\hat{\tau}^2 = 0$.

The random effects methods generally perform better than the fixed effects methods, with respect to coverage probabilities. However, particularly when the number of studies is modest (fewer than

20), the commonly used DerSimonian and Laird method has coverage probability considerably below 0.95. This suggests that the error associated with the estimation of τ^2 is not adequately being accounted for either through modifications to $\widehat{\text{var}}(\hat{\mu}_{\hat{\tau}})$ or the distribution used for $\hat{\mu}_{\hat{\tau}}$. Like the DerSimonian and Laird method, the greatest source of error in simple likelihood confidence intervals comes from estimating $\text{var}(\hat{\mu}_{\text{ml}})$ by substituting $\hat{\tau}^2_{\text{ml}}$ in for τ^2, and using a normal approximation, despite this estimation. Although it lacks the simplicity of the other three techniques, the profile likelihood method produced the highest coverage probabilities in all cases. In particular, coverage probabilities for small k were considerably closer to 0.95 than for the other two random effects methods.

For both the fixed and random effects methods, inference is carried out ignoring the sampling errors in the individual study variances. Estimated values $\hat{\sigma}_i^2$ are used without modification to the form of $\hat{\mu}$, its variance or distribution. It has been shown, however, that the fixed effects variance estimate, $\widehat{\text{var}}(\hat{\mu})$, has a negative bias [12]. The estimation of σ_i^2 might also be expected to influence random effects coverage probabilities. Hardy and Thompson [10] suggest however that in the case of maximum likelihood procedures, allowing for the estimation of σ_i^2 does not greatly affect results unless all studies in the meta-analysis are small.

Non-Bayesian alternatives to the standard random effects approaches considered here have recently been proposed in an attempt to incorporate more adequately the estimation of τ^2 into the random effects model. The first, outlined by Biggerstaff and Tweedie [11] utilizes an approximate distribution for $\hat{\tau}^2$ to modify the estimation of random effects weights. These new weights are then used in the estimation of μ and its variance. However the variance of $\hat{\mu}_{\hat{\tau}}$ is still derived assuming the weights are known, and the assumption of normality for $\hat{\mu}_{\hat{\tau}}$ is maintained. A second alternative involves using an overdispersed generalized linear model to estimate the overall effect μ, with the heterogeneity between studies being reflected by the overdispersion parameter [13].

A third approach, currently being developed, uses simulations to model the 0.975 quantile of $(\hat{\mu}_{\hat{\tau}} - \mu)/\sqrt{\text{var}(\hat{\mu}_{\hat{\tau}})}$ as a function of k. The appropriate value, which will be greater than 1.96, is then used in confidence intervals for μ. Clearly all three methods need to be examined further and incorporated into a comparative study of meta-analysis techniques. They do however provide possible improvements to the deficiencies highlighted in currently used techniques.

APPENDIX A: MAXIMUM LIKELIHOOD ESTIMATION

If the log-likelihood function, given in (2), is partially differentiated with respect to μ and τ^2 and the derivatives are set to zero, some algebraic arrangement gives

$$\hat{\mu}_{\text{m}} = \sum \frac{y_i}{\hat{\sigma}_i^2 + \hat{\tau}^2_{\text{m}}} \Big/ \sum \frac{1}{\hat{\sigma}_i^2 + \hat{\tau}^2_{\text{m}}} \tag{A1}$$

$$\hat{\tau}^2_{\text{m}} = \sum \frac{(y_i - \hat{\mu}_{\text{m}})^2 - \hat{\sigma}_i^2}{(\hat{\sigma}_i^2 + \hat{\tau}^2_{\text{m}})^2} \Big/ \sum \frac{1}{(\hat{\sigma}_i^2 + \hat{\tau}^2_{\text{m}})^2} \tag{A2}$$

The maximum likelihood estimates $\hat{\mu}_{\text{ml}}$ and $\hat{\tau}^2_{\text{ml}}$ are then

$$(\hat{\mu}_{\text{ml}}, \hat{\tau}^2_{\text{ml}}) = \begin{cases} (\hat{\mu}_{\text{m}}, \hat{\tau}^2_{\text{m}}) & \text{if } \hat{\tau}^2_{\text{m}} > 0 \\ (\hat{\mu}, 0) & \text{if } \hat{\tau}^2_{\text{m}} \leqslant 0 \end{cases}$$

where $\hat{\mu}$ is the fixed effects estimate of μ. Note that since $\hat{\tau}^2_{\mathrm{m}}$ may be less than zero, truncation is required.

Equations (A1) and (A2) must be solved iteratively. Substituting equation (A1) into (A2) we obtain

$$\hat{\tau}^2_{\mathrm{m}} = f(\hat{\tau}^2_{\mathrm{m}})$$

where f is the resulting function of $\hat{\tau}^2_{\mathrm{m}}$ and the data. This can be solved for $\hat{\tau}^2_{\mathrm{m}}$ using the iteration

$$\hat{\tau}^2_{\mathrm{t}} = f(\hat{\tau}^2_{t-1}) \tag{A3}$$

This simple dynamic system can be iterated until it converges to a fixed point which will be the desired estimate, and $\hat{\mu}_{\mathrm{m}}$ can then be evaluated by substituting $\hat{\tau}^2_{\mathrm{m}}$ into (A1).

For our simulations the iterations are initialized with $\hat{\tau}^2_{\mathrm{m}} = \hat{\tau}^2 + 0.01$ and repeated until both estimates converge to within 10^{-6}. The small value is added to the DerSimonian and Laird estimate of τ^2 to prevent iterations starting at zero in cases where $\hat{\tau}^2 = 0$.

The convergence of this iterative procedure is not guaranteed however. Simulations show that in some cases the iterative procedure does not converge to a single fixed point, but to a limit cycle of higher order, most commonly two. Whilst convergence to such cyclic behaviour does not often occur, the possibility implies that the iterative procedure will not necessarily produce estimates which maximize the likelihood function [14].

An obvious alternative to the iterative routine is to find maximum likelihood estimates by direct maximization of the likelihood function. The requirement that $\hat{\tau}^2_{\mathrm{ml}} \geqslant 0$ necessitates maximization subject to this constraint. This method is used for our simulations where the iterative procedure has not converged within 1000 iterations. It is implemented using Powell's algorithm [15], which minimizes the negative log-likelihood function.

APPENDIX B: PROFILE LIKELIHOOD INTERVALS

The profile likelihood interval is derived using the standard result for the asymptotic distribution of a likelihood ratio statistic [9]. Applying this result and assuming $H_0 : \mu = \mu_0$ is true we have

$$-2 \log \left[\frac{L_\mu(\mu_0)}{L_\mu(\hat{\mu}_{\mathrm{ml}})} \right] \xrightarrow{\mathrm{d}} \chi^2_1 \quad \text{as } k \to \infty$$

provided $L_\mu(\hat{\mu}_{\mathrm{ml}})$ is the maximum likelihood value under the general hypothesis. Note that $L_\mu(\hat{\mu}_{\mathrm{ml}}) = L(\hat{\mu}_{\mathrm{ml}}, \hat{\tau}^2_{\mathrm{ml}})$ – the likelihood function evaluated at the maximum likelihood estimates.

Finding the profile likelihood interval requires finding values of μ_0 which satisfy (4). This can be achieved by implementing an iterative search for the lower and upper bounds of the interval. We begin by selecting a value of μ_0 well below the expected lower bound of the confidence interval. Substituting this into (3) we can iteratively solve for $\hat{\tau}^2(\mu_0)$. A value of $\log(L_\mu(\mu_0))$ is then obtained by substituting both μ_0 and $\hat{\tau}^2(\mu_0)$ into the log-likelihood equation; $\log(L_\mu(\mu_0))$ is then compared to the right-hand side of (4). If the value of μ_0 is either too small (and hence the inequality is false), or too large, so it is within the interval rather than on the boundary, it is adjusted and $\log(L_\mu(\mu_0))$ recalculated. This is repeated until a confidence bound is obtained to within 10^{-6}. A similar procedure is then used to find the upper bound of the confidence interval.

As noted, each evaluation of $\log(L_\mu(\mu_0))$ requires finding $\hat{\tau}^2(\mu_0)$ which satisfies (3). As with the iterative procedure discussed in Appendix A, iterations of this function will not necessarily converge to a single value. Simulation suggests that for small τ^2 and small k, convergence to a cycle is more likely to occur. It is possible, however, to find $\hat{\tau}^2(\mu_0)$ by directly maximizing the likelihood function with respect to τ^2, with $\mu = \mu_0$ fixed. Again this maximization is implemented using Powell's algorithm [15].

APPENDIX C: FIXED EFFECTS COVERAGE PROBABILITY

If the fixed effects method is used for any $\tau^2 \geqslant 0$, the coverage probability is as given in (5). In such cases we have $Y_i \stackrel{\mathrm{d}}{=} \mathrm{N}(\mu, \hat{\sigma}_i^2 + \tau^2)$. This gives

$$\hat{\mu} = \frac{\sum \hat{w}_i Y_i}{\sum \hat{w}_i} \stackrel{\mathrm{d}}{=} \mathrm{N}\left(\mu, \frac{\sum \hat{w}_i + \tau^2 \sum \hat{w}_i^2}{(\sum \hat{w}_i)^2}\right)$$

The fixed effects coverage probability is given by c, where

$$c = \Pr\left(\hat{\mu} - \frac{1.96}{\sqrt{\sum \hat{w}_i}} < \mu < \hat{\mu} + \frac{1.96}{\sqrt{\sum \hat{w}_i}}\right) = \Pr\left(\frac{-1.96}{\sqrt{\sum \hat{w}_i}} < \hat{\mu} - \mu < \frac{1.96}{\sqrt{\sum \hat{w}_i}}\right)$$

Standardizing $\hat{\mu}$ we have

$$c = \Pr\left(\frac{-1.96}{\sqrt{\sum \hat{w}_i}} \frac{\sum \hat{w}_i}{\sqrt{(\sum \hat{w}_i + \tau^2 \sum \hat{w}_i^2)}} < Z < \frac{1.96}{\sqrt{\sum \hat{w}_i}} \frac{\sum \hat{w}_i}{\sqrt{(\sum \hat{w}_i + \tau^2 \sum \hat{w}_i^2)}}\right)$$

where $Z \stackrel{\mathrm{d}}{=} \mathrm{N}(0, 1)$. Hence

$$c = \Pr(Z < z) - \Pr(Z < -z) = \Phi(z) - \Phi(-z)$$

where

$$z = \frac{1.96}{\sqrt{\left(1 + \tau^2 \frac{\sum \hat{w}_i^2}{\sum \hat{w}_i}\right)}}$$

ACKNOWLEDGEMENT

The authors thank Ray Watson for valuable discussions and suggestions during the preparation of this paper.

REFERENCES

1. Olkin I. Meta-analysis: methods for combining independent studies. Editor's introduction. *Statistical Science* 1992; **7**:226.
2. Hunter JE, Schmidt FL, Jackson GB. *Meta-analysis: cumulating research findings across studies*. Sage: Beverly Hills, 1992.
3. Schmid JE, Koch GG, LaVange LM. An overview of statistical issues and methods of meta-analysis. *Journal of Biopharmaceutical Statistics* 1991; **1**:103–120.
4. 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.

5. Cochran WG. Problems arising in the analysis of a series of similar experiments. *Journal of the Royal Statistical Society*, 1937; **4**:(Supplement) 102–118.
6. Hardy RJ, Thompson SG. Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine* 1998; **17**:841–856.
7. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**:177–188.
8. Peto R. Aspirin after myocardial infarction. *Lancet* 1980; **1**:1172–1173.
9. Watson R. *Elementary Mathematical Statistics.* VABA Publishing: Melbourne, 1978.
10. Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects: *Statistics in Medicine* 1996; **15**:619–629.
11. 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.
12. Yuan Zhang Li, Li Shi, Daniel RH. The bias of the commonly-used estimate of variance in meta-analysis. *Communications in Statistics*, *Part A – Theory and Methods* 1994; **23**:1063–1085.
13. Gordon IR. A simple method for dealing with between-study variation in meta-analysis. Statistical Consulting Centre, University of Melbourne, Technical Report 1, 1998.
14. Thompson MT, Stewart HB. *Nonlinear Dynamics and Chaos.* Wiley: New York, 1986.
15. Press WH, Flannery BP, Teukolsky SA, Vetterling WT. *Numerical Recipes in C*. Cambridge University Press: Cambridge, 1988.

