Chapter 14
Meta-analysis I

Computational Methods

Suhail A.R. Doi and Jan J. Barendregt

Abstract Meta-analysis is now used in a wide range of disciplines, in particular epidemiology and evidence-based medicine, where the results of some meta-analyses have led to major changes in clinical practice and health care policies. Meta-analysis is applicable to collections of research that produce quantitative results, examine the same constructs and relationships, and have findings that can be configured in a comparable statistical form called an effect size (e.g. correlation coefficients, odds ratios, proportions, etc.), that is, are comparable given the question at hand. These results from several studies that address a set of related research hypotheses are then quantitatively combined using statistical methods. This chapter provides an in-depth discussion of the various statistical methods currently available, with a focus on bias adjustment in meta-analysis.

Introduction

Meta-analysis is now used in a wide range of disciplines, in particular epidemiology and evidence-based medicine where the results of some meta-analyses have led to major changes in clinical practice and health care policies. Meta-analysis is applicable to collections of research that produce quantitative results, examine the same constructs and relationships and have findings that can be configured in a comparable statistical form called an effect size (ES) (e.g. correlation coefficients, odds ratios, proportions, etc.), that is, are comparable given the question at hand. These results from several studies that address a set of related research hypotheses are then quantitatively combined using statistical methods. The set of related research hypotheses can be demonstrated at a broad level of abstraction; for example, ovarian ablation therapy for ovulation induction in polycystic ovarian syndrome

S.A.R. Doi (✉) • J.J. Barendregt
School of Population Health, University of Queensland, Brisbane, QLD, Australia
e-mail: sardoi@gmx.net

S.A.R. Doi and G.M. Williams (eds.), *Methods of Clinical Epidemiology*,
Springer Series on Epidemiology and Public Health,
DOI 10.1007/978-3-642-37131-8_14, © Springer-Verlag Berlin Heidelberg 2013

where various therapies are lumped together such as laser, wedge resection and interstitial ablation. Alternatively, it may be at a narrow level of abstraction and represent pure replications. The closer to pure replications the collection of studies is, the easier it is to argue comparability. Forms of research suitable for meta-analysis include group contrasts such as experimentally created groups, that is, comparison of outcomes between experimental and control groups and naturally or non-experimentally occurring groups (treatment, prognostic or diagnostic features). Pre-post contrasts can also be meta-analysed, for example, changes in continuous or categorical variables. Another area for meta-analysis is central tendency research such as incidence or prevalence rates and means. Association between variables can be meta-analysed, such as correlation coefficients and regression coefficients.

The meta-analysis differs from the systematic review in that the focus changes to the direction and magnitude of the effects across studies, which is what we are interested in anyway. Direction and magnitude are represented by the ES and therefore this is a key requirement for, and is what makes meta-analysis possible. It is a quantitative measure of the strength of the relationship between intervention and outcome and it encodes the selected research findings on a numeric scale. There are many different types of ES measures, each suited to different research situations. Each ES type may also have multiple methods of computation. The type of ES must be comparable across the collection of studies of interest for meta-analysis to be possible. This is sometimes accomplished through standardization when some or all of the studies use different scales (e.g. the standardized mean difference). A standard error must be calculable for that type of ES because it is needed to calculate the meta-analysis weights, called inverse variance weights (more on this later) as all analyses are weighted. Thus, it is important to abstract ES information from studies if the systematic review is to be followed up with a meta-analysis. The pooled estimate is usually computed by meta-analysis software based on the ES input selected. The software we have created is MetaXL (www.epigear.com) and it also has an option to enter the ES and standard error (SE) directly or to bypass the SE input thus allowing a multivariable adjusted ES to be entered directly.

It is therefore evident that combining quantitative data (synthesis) is what is central to the practice of meta-analysis. The basic underlying premise is that the pooled results from a group of studies can allow a more accurate estimate of an effect than an individual study because it overcomes the problem of reduced statistical power in studies with small sample sizes. However, pooling in meta-analysis must be distinguished from *simple pooling* where there is the implication that there is no difference between individual studies (or subgroups) so that it is seemingly acceptable to consider that the data from the control group of one study might just have easily come from the control group of another study. Bravata and Olkin (2001) point out that in reality, by simple pooling, we are assigning different weights to intervention and control groups and this can lead to paradoxical results. Of course, if the individual studies have the same sample size in the intervention and control groups (studies are balanced) such paradoxes will not occur and this explains why balanced designs are advocated for randomized controlled trials and why simple pooling of centres is sometimes used in such multi-centre trials. Bravata and Olkin (2001) emphasize that while *simple pooling* is obtained by combining first, then comparing, the meta-analytic method compares first, then combines. Thus, the order in which the operations of combining and comparing are carried out is the difference between simple pooling and combining data for meta-analysis and will yield different answers. Combining data via meta-analysis therefore provides a safeguard against reversals such as Simpson's paradox that can occur from simple pooling.

Common Effect Sizes

Standardized Mean Difference and Correlation

This is commonly used with group contrast research, treatment groups and naturally occurring groups where the measurements are inherently continuous. It uses the pooled standard deviation (some situations use control group standard deviation) and is called Cohen's "d" or occasionally Hedges "g". The standardized mean difference can be calculated from a variety of statistics and calculators are available for various methods and remember that any data for which you can calculate a standardized mean difference ES, you can also calculate a correlation type ES. Standardized mean difference ES has an upward bias when sample sizes are small but this can be removed with the small sample size bias transformation. If $N = n_1 + n_2$ then

$$\mathrm{ES}'_{\mathrm{sm}} = ES_{\mathrm{sm}}\left[1 - \frac{3}{4N - 9}\right]$$

Correlation has a problematic standard error formula and this is needed for the meta-analysis inverse variance weight. In this case the Fisher's Zr transformation is used:

$$\mathrm{ES}_{\mathrm{Zr}} = 0.5\ln\left[\frac{1+r}{1-r}\right]$$

and results can be converted back into r with the inverse Zr transformation:

$$r = \frac{e^{2\mathrm{ES}_{\mathrm{Zr}}} - 1}{e^{2\mathrm{ES}_{\mathrm{Zr}}} + 1}$$

Odds Ratio/Relative Risk

Again this is used with group contrast research but this time there the measurements are inherently dichotomous. The odds ratio is based on a 2 by 2 contingency table and is the odds of success in the treatment group relative to the odds of success in the control group. Odds ratio/RR are asymmetric and have a complex standard error formula. Negative relationships are indicated by values between 0 and 1. Positive relationships are indicated by values between 1 and infinity. To solve this imbalance, the natural log of the odds ratio/RR is used in meta-analysis.

$$\mathrm{ES_{LOR}} = \ln[\mathrm{OR}], \quad \mathrm{ES_{LRR}} = \ln[\mathrm{RR}]$$

In this case a negative relationship is <0, no relationship $= 0$, and a positive relationship is >0. Results can be converted back into odds ratios/RR by the inverse natural log function.

Proportion/Diagnostic Studies

This is used in central tendency research e.g. prevalence rates and other proportions such as sensitivity and specificity. Proportions have an unstable variance and thus transformed proportions are automatically used by the software. We use the double arcsine square root transformation in MetaXL (http://www.epigear.com) and more details are given in the section below on proportions.

Pooling Effect Sizes

The Fixed Effects Model

The standard approach frequently used in weighted averaging for meta-analysis in clinical research is termed the inverse variance method or FE model based on Woolf (1955). The average ES across all studies is computed whereby the weights are equal to the inverse variance of each study's effect estimator. Larger studies and studies with less random variation are given greater weight than smaller studies. The weights (w) allocated to each of the studies are then inversely proportional to the square of the SE; thus for the ith study

$$w_i = \frac{1}{\mathrm{SE_i}^2}$$

which gives greater weight to those studies with smaller SEs.

As can be seen above, the variability within each study is used to weight each study's effect in the current approach to combining them into a weighted average as this minimizes the variance (assuming each study is estimating the same target). So, if a study reports a higher variance for its ES estimate, it would get lesser weight in the final combined estimate and vice versa. This approach, however, does not take into account the innate variability that exists between the studies arising from differences inherent to the studies such as their protocols and how well they were executed and conducted. This major limitation has been well recognized and it gave rise to the random effects (RE) model approach by DerSimonian and Laird (1986).

The Random Effects Model

A common model used to synthesize heterogeneous research is the RE model of meta-analysis. Here, a constant is generated from the homogeneity statistic Q and, using this and other study parameters, a random effects variance component (REVC) (τ^2) is generated. The inverse of the sampling variance plus this constant that represents the variability across the population effects is then used as the weight

$$w_i^* = \frac{1}{\mathrm{SE}_i^2 + \tau^2}$$

where w_i^* is the RE weight for the ith study. However, because of the limitations of the RE model, when used in a meta-analysis of badly designed studies, it will still result in biased estimates even though there is statistical adjustment for ES heterogeneity (Senn 2007). Furthermore, such adjustments, based on an artificially inflated variance, lead to a widened confidence interval, supposedly to reflect ES uncertainty, but Senn (2007) has pointed out that they do not have much clinical relevance.

The weight that is applied in this process of weighted averaging with an RE meta-analysis is achieved in two steps:

- Step 1: Inverse variance weighting
- Step 2: Un-weighting of this inverse variance weighting by applying an REVC that is simply derived from the extent of variability of the ESs of the underlying studies.

This means that the greater this variability in ESs (otherwise known as heterogeneity), the greater the un-weighting and this can reach a point when the RE meta-analysis result becomes simply the un-weighted average ES across the studies. At the other extreme, when all ESs are similar (or variability does not exceed sampling error), no REVC is applied and the RE meta-analysis defaults to simply a fixed effect meta-analysis (only inverse variance weighting). Al Khalaf et al. (2011) have pointed out that the extent of this reversal is solely dependent on two factors:

1. Heterogeneity of precision
2. Heterogeneity of ES

Since there is absolutely no reason to automatically assume that a larger variability in study sizes or ESs automatically indicates a faulty larger study or more reliable smaller studies, the re-distribution of weights under this model bears no relationship to what these studies have to offer. Indeed, there is no reason why the results of a meta-analysis should be associated with this method of reversal of the inverse variance weighting process of the included studies. As such, the changes in weight introduced by this model (to each study) results in a pooled estimate that can have no possible interpretation and, thus, bears no relationship with what the studies actually have to offer.

To compound the problem further, some statisticians are proposing that we take an estimate that has no meaning and compute a prediction interval around it. This is akin to taking a random guess at the effectiveness of a therapy and under the false belief that it is meaningful try to expand on its interpretation. Unfortunately, there is no statistical manipulation that can replace commonsense. While heterogeneity might be due to underlying true differences in study effects, it is more than likely that such differences are brought about by systematic error. The best we can do in terms of addressing heterogeneity is to look up the list of studies and attempt to un-weight (from inverse variance) based on differences in evidence of bias rather than ES or precision that are consequences of these failures.

Problems with These Conventional Models

One problem with meta-analysis is that differences between trials, such as sources of bias, are not addressed appropriately by current meta-analysis models. Bailey (1987) lists several reasons for such differences: chance, different definitions of treatment effects, credibility-related heterogeneity (quality), and unexplainable and real differences. An important explainable difference is credibility-related heterogeneity (quality) and this has been defined by Verhagen et al. (2001) as the likelihood of the trial design generating unbiased results that are sufficiently precise to allow application in clinical practice. The flaws in the design of individual studies have obvious relevance to creating heterogeneity between trials as well as an influence on the magnitude of the meta-analysis results. If the quality of the primary material is inadequate, this may falsify the conclusions of the review, regardless of the presence or absence of ES heterogeneity. The need to address heterogeneity in trials via study-specific assessment has been obvious for a long time and the solution involves more than just inserting a random term based on ES heterogeneity as is done with the RE model.

Previous studies that have attempted to investigate incorporation of some study-specific components in the weighting of the overall estimate concluded that incorporating such information into weights provided inconsistent adjustment of the estimates of the treatment effect. Although these authors follow the same assumption as we do that studies with deficiencies are less informative and should have less influence on overall outcomes, methodology was flawed and such attempts therefore did not reduce bias in the pooled estimate, and may have resulted in an increase in bias.

A study score-adjusted model that overcomes several limitations has been introduced by Doi and Thalib (2008, 2009). The rationale was that in a group of homogeneous trials, it is assumed that because the ESs are homogeneous, the studies are all estimating the same target effect (we can call this a type A trial). In this situation, the inverse variance weights of Woolf (1955) will minimize the variance since the mean squared error (MSE) = expected(estimate − true)^2 = variance + (bias)^2. Bias is zero if the underlying true ESs are equal and thus minimizing variance is optimal and the weighted MSE = variance. It is thought that the inverse variance-weighted analysis tests the null hypothesis that all studies in the meta-analysis are identical and show no effect of the intervention under consideration regardless of homogeneity. This requires the assumption that trials are exchangeable so that if one large trial is null and multiple small trials show an effect, the large trial essentially decreases evidence against the null hypothesis. Exchangeability, however, is likely to be conditional only, as discussed later, and thus this is a big assumption. Therefore, if we do not believe the trials are exchangeable then, in this situation, we have two alternatives: either the trials have been affected by bias even though the underlying true effects are identical (we can call these type B trials) or the trials represent different underlying true effects (we can call these type C trials). In the former case, the trial ES from a biased trial might seem like it is coming from a different underlying true effect, thus giving the impression that the trials represent different underlying true effects. In type B trials, inverse variance weights do not minimize the variance, it just exaggerates it and creates gross bias in these situations. Furthermore, any set of weights in a type A situation estimates the same target, but in a type B situation each set of weights estimates a different target. Thus, inverse variance weights in the latter situation just increase bias and are not optimal for type B trials. Thus, in type B trials, we would want to use situation-specific weights.

One such situation-specific weight that has been suggested for type B trials is weighting according to the probability (Q_i) of credibility (internal validity or quality) of the studies making up the meta-analysis. Although this can correct for distortions due to systematic error, it can also introduce errors of another type. For example, a study of a small sample that is not representative of the underlying population may get a large quality weighting and this can skew the data. It might thus be informative to weight according to precision and then redistribute the weights according to situation-specific requirements. In this case, the importance of smaller good quality studies are upgraded only if the larger or more precise studies are deemed poor by its situation-specific weight.

This line of thought is not new as this is precisely what the RE model attempts to do. The unfortunate thing, however, is that the situation-specific weight used in this particular model is an index of the variability of the ESs across trials and the same situation-specific weight is applied to all trials (the RE model). It becomes quite clear that the type B meta-analysis differ from the RE model in that between-study variability is visualized as a fixed rather than a random effect and thus represents an extension of a fixed effects model that can address heterogeneity. In type B trials, the expectation is that the expected value of the study estimate differs from the grand (real) mean (μ) by an amount β_j and the true (study-specific) mean (θ_j) for study j is given by $\theta_j = \mu + \beta_j$. The divergence, however, is that the β_j are not interpreted as a random effect with type B studies and thus do not have a common variance. The philosophy behind the random effect construct is that it presupposes that the study effects are randomly sampled from a population with a varying ($\sigma_\tau^2 > 0$) underlying parameter of interest. Overton (1998) thus has stated that if the studies included in the meta-analysis differ in some systematic way from the possible range in the population (as is often the case in the real world), they are not representative of the population and the RE model does not apply, at least according to a strict view of randomization in statistical inference.

In addition, with the RE model, the weight of the larger studies are redistributed to smaller studies but τ^2 has a decreasing effect as study precision declines. The size of τ^2 is determined by how heterogeneous the ESs are and if τ^2 is zero, the RE model defaults to the FE model. If we focus on the largest study, the bigger its difference from other studies, the bigger the τ^2 and the decrease in weight of this study. Al Khalaf et al. (2011) demonstrates that τ^2 has a U-shaped association with ES in the largest study, being minimal when the largest study conforms to other study ESs, and as this ES departs from that of other studies, τ^2 increases. The weight of the largest study then declines as τ^2 increases. However, while the biggest individual study weight decrements associated with bigger τ^2 follow a predictable pattern, the impact of different τ^2 values on the pooled estimate is unpredictable. This happens because, although individual study weight changes are predictable from τ^2, the relationship of weight gain across smaller studies bears no relationship to which study shows the most ES heterogeneity, or indeed any tangible information from the study.

The Quality Effects Model

In order to rectify this situation, an alternative approach was proposed by Doi and Thalib (2008) and subsequently modified in 2011 and 2012. The main reasoning was: suppose there are K studies in a set of studies that belong to a meta-analysis and x_j and w_j are random variables representing the ES and normalized (sum to 1) weights, respectively, with the study labels $j = 1,\ldots,K$. The expected value of x_j was taken to be the underlying parameter (μ) being estimated. However, in this situation, the ESs are assumed to be similar in the sense that the study labels

($j = 1,\ldots,K$) convey no information and are thus considered independent and identically distributed (IID). The reality is that each of these labels (representing independent studies) is associated with specific information about the likelihood of systematic bias (β_j) and thus for all j the x_j are in fact only conditionally IID and would be estimating a specific biased parameter. Assuming that heterogeneity derives from essentially non-random systematic error and randomness is only obtained via a random permutation of the indices 1,2,. . .,K, then details about the design of study j do provide information about these systematic errors and can be represented by a hierarchical model for each study:

$$\beta_j \sim N(\beta, \phi^2) \quad \text{(bias effects)}$$

$$(x_j \,|\, \widehat{\mu+\beta_j}) \overset{\text{indep}}{\sim} N(\widehat{\mu+\beta_j}, {\sigma_j}^2 + \phi_j^2) \quad \text{(study)}$$

The bottom level of underlying effects, the study level of the hierarchical model, says that because of relevant differences in methodology and systematic errors, each study has its own underlying treatment effect $\mu + \beta_j$, and the observed ES differences x_j are like random draws from a normal distribution with mean $\widehat{\mu+\beta_j}$ and variance ${\sigma_j}^2 + \phi_j^2$ (the normality is reasonable because of the central limit theorem). Thus, a suitable linear model for the jth study (not considering across all studies) can be written as

$$x_j = \widehat{\mu+\beta_j} + \varepsilon_j \tag{14.1}$$

and for each study

$$\mathrm{E}(\mathrm{x}_j) = \widehat{\mu+\beta_j}$$

Also, under the assumption of no prior information about weights (w_j) except that they sum to 1, they will be equally distributed with the expected value of w_j being $1/K$ for all j. If $c = \mathrm{Cov}(\mathrm{w}_j, \mathrm{x}_j)$ is the covariance of these random quantities across all studies, then

$$\mathrm{E}(\mathrm{w}_j\mathrm{x}_j) = \mathrm{Cov}(\mathrm{w}_j, \mathrm{x}_j) + \mathrm{E}(\mathrm{x}_j)\mathrm{E}(\mathrm{w}_j) = c + \widehat{\mu+\beta_j}/K$$

and thus summing across all studies,

$$\mathrm{E}\left(\sum_{j=1}^{\mathrm{K}} (\mathrm{w}_j\mathrm{x}_j)\right) = \mu + \frac{1}{K}\sum \beta_j + Kc$$

since $\mathrm{E}(w_j) = 1/K$.

Thus, it is clear that if we use empirical weights, *c is* not zero, $\Sigma\beta_j$ is also not zero and the meta-analytic estimate for μ is biased. It is probably true, as suggested by Shuster (2010), that the unweighted estimate is a less biased estimate in situations where w_j and x_j are correlated. However, it is clear that an unbiased estimate of μ will not be provided unless the average $\beta_j = 0$, so systematic error also leads to increase in bias.

Everything hinges on the variance and, therefore, the mere observation that the unweighted estimate is likely to be unbiased does little to reaffirm our confidence in its utility without a simultaneous measure of its global error (with respect to its parameter). The MSE thus has to be minimized and the fact that bias is included as a component is important because the judgment of the performance of the model depends on the trade-off between the amount of bias and the variability.

It may be noted that for a particular study,

$$\mathrm{Var}(\mathrm{x}_j\mathrm{w}_j) = (\sigma_j^2 + \phi_j^2)\mathrm{w}_j^2$$

Therefore,

$$\mathrm{Var}\left(\sum \mathrm{x}_j\mathrm{w}_j\right) = \sum (\sigma_j^2 + \phi_j^2)\mathrm{w}_j^2 \tag{14.2}$$

Also, under the constraint that $\Sigma w = 1$ and only if ${\sigma_j}^2 + \phi_j^2$ was equal for all K studies, does the variance attain its minimum value for equal weights, and its maximum when all weights except one are zero. This is not the case from Eq. 14.2 and the naturally weighted average is expected to have a poor bias–variance trade-off. The only logical solution therefore is to discount studies that are expected to have an inflated value for β_j. This can be achieved by linking β_j to the probability that a study is credible as follows. If $\beta = 0$ and if

$$\sum_{j=1}^{K} \beta_j^2/K = \phi^2$$

then

$$Q_j = \phi^2/(\phi^2 + \phi_j^2)$$

which can be interpreted as the probability that study j is credible as described previously by Spiegelhalter and Best (2003) or Turner et al. (2009). Therefore,

$$\phi_j^2 = (\phi^2 - Q_j\phi^2)/Q_j$$

What this means is that as Q_j and the individual study bias variance (ϕ_j^2) are inversely related and thus an inverse discounting system for such studies based on Q_j should be optimal if the expected increase in bias ends up being traded off by larger decreases in variance. This is a logical conclusion also reiterated by Burton et al. (2006) as any method that results in an unbiased estimate but has large variability cannot be considered to be have much practical use.

To discount by quality requires computation of an adjusted Q_j first as follows (See Doi et al. 2011, 2012):

$$Q_j(\text{adj}) = \begin{cases} \left(\dfrac{\left(\sum_{j=1}^{K} Q_j \right) \tau_j}{\left(\sum_{j=1}^{K} \tau_j \right) (K-1)} \right) + Q_j & \text{if } \left(\exists \mathrm{Q}_j \right) \mathrm{Q}_j < 1 \\ Q_j & \text{otherwise.} \end{cases}$$

where

$$\tau_j = \frac{iw_j - (iw_j \times Q_j)}{K-1}$$

and iw_j is the inverse variance weight of study j, Q_j is the credibility of study j ranging from 0 to 1 and K is the number of studies in the meta-analysis. From the adjusted quality parameter, a quality adjustor is then computed given by

$$\hat{\tau}_j = \left(\left(\sum_{j=1}^{K} \tau_j \right) K \frac{Q_j(\text{adj})}{\sum_{j=1}^{K} Q_j(\text{adj})} \right) - \tau_j$$

This is then used to compute the study bias-specific variance component $\hat{Q}_j$ as follows:

$$\hat{Q}_j = Q_j + \left(\frac{\hat{\tau}_j}{iw_j} \right)$$

What these equations do is replace the REVC with study-specific variance components so that the target this meta-analysis is estimating becomes meaningful. Given that the final weight for the study is $w_j^{\delta} = iw_j \hat{Q}_j$, the final summary estimate is then given by

$$\bar{x}_{\mathrm{QE}} = \frac{\sum (w_j^{\delta} \times x_j)}{\sum w_j^{\delta}} = \frac{\sum (\hat{Q}_j \times iw_j \times x_j)}{\sum (\hat{Q}_j \times iw_j)}$$

where $\bar{x}$ is the pooled ES measure and it has a variance (V) given by

$$V_{\mathrm{QE}} = \sum \sigma^2{}_j \left(\frac{w_j^{\delta}}{\sum w_j^{\delta}} \right)^2$$

Given that $iw_j = 1/\sigma_j^2$, this reduces to:

$$V_{\mathrm{QE}} = \frac{\sum (\hat{Q}_j{}^2 \times iw_j)}{\left(\sum (\hat{Q}_j \times iw_j) \right)^2}$$

However, there is expected to be significant overdispersion and thus this variance estimate underestimates the true variance and can lead to a confidence interval with poor coverage. To rectify this, a correction factor (CF) has been proposed for overdispersion based on iterative simulation studies using the Q statistic (χ_{c}) as follows (Doi et al. 2011):

$$\mathrm{CF} = \left(1 - \max \left[0, \frac{\chi_{\mathrm{c}} - (\mathrm{K} - 1)}{\chi_{\mathrm{c}}} \right] \right)^{0.25}$$

For computation of the variance of the weighted average, the variance of each study is then inflated to the power CF as follows:

$$iw'_j = \frac{1}{\left(\sigma_j^2 \right)^{\mathrm{CF}}} \quad \text{if } \sigma_j^2 < 1 \quad \text{or} \quad iw'_j = \frac{1}{\left(\sigma_j^2 \right)^{(2-\mathrm{CF})}} \quad \text{if } \sigma_j^2 \geq 1$$

This can then be used to update V_{QE} as follows:

$$V_{\mathrm{QE}} = \frac{\sum (\hat{Q}_j{}^2 \times iw'_j)}{\left(\sum (\hat{Q}_j \times iw'_j) \right)^2}$$

Assuming the distribution of these estimates are asymptotically normal, the 95 % confidence limits are easily obtained by

$$95\,\%\ \mathrm{CI} = \overline{\mathrm{ES}} \pm 1.96(\sqrt{\mathrm{V}_{\mathrm{QE}}})$$

It becomes quite clear, that the quality-based method differs from the RE model in that between-study variability is visualized as a fixed rather than a random effect and thus represents an extension of a fixed effects model that can address heterogeneity. In both the classic random effect method and the quality-based method, the β_j is taken to be the difference between the grand (real) mean (μ) and the true (study-specific) mean (x_j) for study j ($\beta_j = x_j - \mu$). The divergence, however, is that the β_j are not interpreted as a random effect with the quality-based method and thus do not have a common variance. The philosophy behind the random effect construct is that it presupposes that the x_j values are randomly sampled from a population with a varying underlying parameter of interest ($\tau^2 > 0$). However, if the studies included in the meta-analysis differ in some systematic way from the possible range in the population (as is often the case in the real world), they are not representative of the population and the RE model does not apply, at least according to a strict view of randomization in statistical inference (Overton 1998). The quality-based method therefore corrects this by interpreting the β_j as a fixed effect related to the study itself (based on systematic or related errors) and thus the effect of a varying target created by this bias can be minimized by discounting studies where within-study bias variance (ϕ_j^2) is likely to be large relative to between-study bias variance (ϕ^2). Such discounting requires a robust mechanism to avoid increasing bias and to simultaneously allow incorporation of sampling errors into the model as detailed above based on previous work on this subject. As mentioned by Eisenhart (1947), which situation applies to the model is the deciding factor in determining whether effects are to be considered as fixed or random and when inferences are going to be confined to the effects in the model, the effects are considered fixed.

While with this model we assume that non-credibility leads to bias in the ES, this supposition is backed by clear evidence from several authors such as Balk et al. (2002), Conn and Rantz (2003), Egger et al. (2003), Moher et al. (1998), Schulz et al. (1995) and others suggesting that inadequate methodology correlates with bias in the estimation of treatment effects. However, there could be instances where lack of credibility does not lead to bias in the estimation of treatment effects (or alternatively where such biases may have been obscured by the lack of credibility). In such cases, the quality effects (QE) model is still valid and credibility information results simply in decreased confidence (wider confidence intervals) in the pooled estimate. We do not delete lower quality studies because every study has something to add to the weighted estimate. We do not know what the relationship of study-specific scores are to the magnitude or direction of bias. However, if this weighting is not based on study- or goal-specific attributes, then the weighted estimate loses meaning. A sensitivity analysis, on the other hand, can only tell us that subgroups are heterogeneous but not what the true estimate is likely to be. In studies that vary due to systematic error, study-specific scores can lead to the best approximation of the true ES. The letter would not be possible with either the RE model or sensitivity analyses.

When weighting study estimates by their study-specific scores, we must keep in mind that these scores do not tell us the direction or magnitude of the change in ES that is attributable to that score. The QE method of Doi and Thalib (2008), is not constrained by this limitation, because, unlike previous methods, it does not adjust a study weight directly but redistributors it in relation to all other study weights based on its quality status. This is exactly what the RE model does too, the major difference being that the latter adds on weight to smaller studies without any rationale for doing so and the process ultimately becomes random. This is because τ^2 is not individualized to each study as $\hat{\tau}_i$ is in the QE model. A gradual increase in weight of smaller studies with quality is seen but not with ES heterogeneity. This also explains why previous attempts by Berard and Bravo (1998) or Tritchler (1999) to incorporate study-specific scores into weights have failed to provide sufficient adjustment of the estimates of treatment effects as they failed to consider ramdom error or counterintuitively decided to incorporate study-specific scores over the random redistribution in an RE model.

Greenland (1994) suggested more than a decade ago that quality scoring merges objective information with arbitrary judgments in a manner that can obscure important sources of heterogeneity among study results. He gave the example of dietary quality scoring in the Nurses Health Study and states that the result would likely indicate no diet effects associated with disease if the effects of important quality items are confounded within strata of the summary quality score. The problem is to use the information regarding quality in this way. If we viewed the diet quality score as the probability that a nurse's diet is accurately measured, we would be able to rank nurses from best to worst reliability of dietary information. Even if this ranking is subjective or poor, we would still be more confident about the relationship between diet and disease in high scorers than in low scorers. This is the correct use of quality scores, but cannot be demonstrated with conventional meta-analysis models (Al Khalaf et al. 2011) given that the spread of precision and ES take precedence over stratification by quality score. The fact that previous authors used scores as exclusion criteria or to sequentially combine trial results using these models would only increase bias by altering the range of precision and ES differences among stratified studies. This is probably the reason why many authors such as Balk et al. (2002), Herbison et al. (2006), Juni et al. (1999) and Whiting et al. (2005) all report that stratification of meta-analyses by quality score has no clear impact on the pooled estimate.

Study-specific assessment has not, until now, found an acceptable means of becoming an important part of meta-analyses. More than half of published meta-analyses do not specify in the methods whether and how they would use study-specific assessment in the analysis and interpretation of results, and only about 1 in 1,000 systematic reviews consider weighting by quality score (Moja et al. 2005). This is probably because of the lack, until now, of an adequate model to do so and therefore those meta-analyses that had an a priori conceptualization of quality simply linked it to the interpretation of results or to limit the scope of the review. Although there is no gold standard and we still do not know how best to measure quality, this is not an obstacle to QE analysis because it works with any quality score. Given that we have demonstrated that the RE model randomly adjusts estimates of treatment effects in a meaningless fashion, it may now be time to switch from observed random statistical ES heterogeneity to models that are based on measured study-specific estimates of their heterogeneity.

The Special Case with Proportions in Meta-analysis

Just about all epidemiologists habitually speak of the prevalence rate, but prevalence is defined as a proportion: the number of cases in a population divided by the population number. This definition implies that (1) prevalence is always between 0 and 1 (inclusive), and (2) the sum over categories always equals 1.

The definition of prevalence is the same as the definition of the binomial distribution (number of successes in a sample), and therefore the standard assumption is that prevalence follows a binomial distribution. With the main meta-analysis methods based on the inverse variance method (or modifications thereof), the binomial equation for variance (expressed as a proportion) can be used to obtain the individual study weights:

$$\mathrm{Var}(p) = \frac{p(1-p)}{N}$$

where p is the prevalence proportion and N is the population size.

With the variance of the individual studies nailed down, the pooled prevalence estimate P then becomes (according to the inverse variance method)

$$P = \frac{\sum_i \frac{p_i}{\mathrm{Var}(p_i)}}{\sum_i \frac{1}{\mathrm{Var}(p_i)}}$$

with SE

$$\mathrm{SE}(P) = \sqrt{\sum_{\mathrm{i}} \frac{1}{\mathrm{Var}(p_i)}} \tag{14.3}$$

The confidence interval of the pooled prevalence can then be obtained by

$$\mathrm{CI}_\gamma(P) = P \pm Z_{\alpha/2}\mathrm{SE}(P)$$

where $Z_{\alpha/2}$ denotes the appropriate factor from the standard normal distribution for the desired confidence percentage (e.g. $\mathrm{Z}_{0.025} = 1.96$).

While this works fine for prevalence proportions around 0.5, increasing problems arise when the proportions get closer to the limits of the 0…1 range. The first problem is mostly cosmetic: the equation for the confidence interval does not preclude confidence limits outside the 0...1 range. While this is annoying, the second problem is much more substantial: when the proportion becomes small or big, the variance of the study is squeezed towards 0 (see Eq. 14.3). As a consequence, in the inverse variance method, the study gets a large weight. A meta-analysis of prevalence according to the method described above therefore puts undue weight on the studies at the extreme of the 0...1 range.

One way to avoid the problem of variance instability with extremes of prevalence is to estimate the SE, not using the individual proportions, but the overall proportion:

$$\mathrm{Var}(p_{ic}) = \frac{p_{\mathrm{total}}(1 - p_{\mathrm{total}})}{N_{ic}}$$

The numerator is now the same for every study and there is no longer the problem where studies with proportions near 50 % get much smaller weights than studies with proportions much smaller or much larger than 50 %. This approach also avoids the problem where a study has 100 % prevalence proportion.

$$w_{ic} = \frac{N_{ic}}{p_{\mathrm{Ctotal}}(1 - p_{\mathrm{Ctotal}})}$$

where $c = 1, \ldots, k$ denotes a particular category out of k categories. In a fixed effect model, use of the pooled proportion to get individual variances would be exactly the same as using individual proportions for variances because the SE of the pooled prevalence in category c becomes

$$\frac{1}{\sum w_{ic}} = \frac{p_{\mathrm{Ctotal}}(1 - p_{\mathrm{Ctotal}})}{\sum N_{ic}}$$

Since each study gets the same weight across categories, this method ensures that the pooled category prevalences sum to 1. However, the confidence interval does not preclude confidence limits outside the 0...1 range, so that problem persists.

The Logit Transformation

To address this issue of estimates falling outside the 0...1 range, the logit transformation was proposed and, at that time, it was thought that it would address both the problems mentioned above. It is given by

$$\mathrm{logit}(p) = \ln\left(\frac{p}{1 - p}\right)$$

with variance

$$\mathrm{Var}(\mathrm{logit}(p)) = \frac{1}{Np} + \frac{1}{N(1-p)}$$

The logit of a proportion has an approximately normal distribution, and as it is unconstrained, it was thought it would avoid the squeezing of the variance effect. The meta-analysis is then carried out on the logit transformed proportions, using the inverse of the variance of the logit as the study weight. For the final presentation, the pooled logit and its confidence interval are back transformed to a proportion using

$$P = \frac{\exp(\mathrm{logit}(P))}{\exp(\mathrm{logit}(P)) + 1}$$

While the logit transformation solves the problem of estimates falling outside the 0…1 limits, unfortunately, it does not succeed in stabilizing the variance; rather there is a reversal of the variance instability of the non-transformed proportions and studies with proportions close to 0 or 1 get their variance estimates grossly magnified and vice versa for proportions around 0.5. The variance instability that plagued non-transformed proportions thus persists even after logit transformation. It has therefore been suggested that, as a rule of thumb, the logit transformation should be used when prevalences are less than 0.2 or more than 0.8.

The Freeman–Tukey Variant of the Double Arcsine Square Root Transformation

The Freeman–Tukey transformation addresses both the problems mentioned above. It is given by

$$t = \sin^{-1}\sqrt{\frac{x_i}{n_i + 1}} + \sin^{-1}\sqrt{\frac{x_i + 1}{n_i + 1}}$$

with variance

$$\mathrm{Var}(t) = \frac{1}{n_i + 0.5}$$

The Freeman–Tukey transformed proportion has an approximately normal distribution, and, by being unconstrained, avoids the squeezing of the variance effect. A meta-analysis can be carried out on the transformed proportions, using the inverse of the variance of the transformed proportion as the study weight. For final presentation, the pooled Freeman–Tukey transformed proportion and its confidence interval are back transformed to a proportion using

$$\bar{P}(\bar{t}) = \begin{cases} 0.5\{1 - \operatorname{sgn}(\cos\bar{t})[1 - (\sin\bar{t} + (\sin\bar{t} - 1/\sin\bar{t})/[1/\hat{v}])^2]^{0.5}\} & \text{if } p/\hat{v} \geq 2 \\ [\sin(\bar{t}/2)]^2 & \text{otherwise.} \end{cases}$$

where $\bar{P}$ is the pooled prevalence, $\bar{v}$ is the pooled variance and $\bar{t}$ is the pooled t.

The lower (LCL) and upper (UCL) confidence limits of the pooled prevalence are given by

$$\text{LCL} = \begin{cases} 0.5\left\{1 - \operatorname{sgn}(\cos\bar{t})\left[1 - \left(\sin\bar{t} + (\sin\bar{t} - 1/\sin\bar{t})/\left[1/\hat{v}\right]\right)^2\right]^{0.5}\right\} & \text{if } \bar{p}/\hat{v} \geq 2 \\ 0 & \text{otherwise.} \end{cases}$$

$$\text{UCL} = \begin{cases} 0.5\left\{1 - \operatorname{sgn}(\cos\bar{t})\left[1 - \left(\sin\bar{t} + (\sin\bar{t} - 1/\sin\bar{t})/\left[1/\hat{v}\right]\right)^2\right]^{0.5}\right\} & \text{if } (1-\bar{p})/\hat{v} \geq 2 \\ 1 & \text{otherwise.} \end{cases}$$

Multi-category Prevalence

The discussion so far has implicitly been about two categories (disease present or absent). But in some instances k-category prevalences may be meta-analysed where $k > 2$ (e.g. mild, moderate and severe disease), and this complicates matters.

Using the previously mentioned non-transformed and logit transformed proportions, it would not be possible to meta-analyse each category separately, since the variance of both p and logit(p) depends on p itself; this implies that the same study could get a different weight in each category, which seems hard to justify. Moreover, the sum over the pooled category prevalences would not add up to 1; another drawback.

To correct this problem, we can use the double arcsine square root transformed proportion where the SE is no longer dependent on the size of the proportion, so that both equal weights across categories and confidence limits within the 0…1 range are achieved without the need for overall proportions.

However, we again have a problem with the RE and QE models in that we need a common study estimate for Cochran's Q that can be used in weighting with the RE model and overdispersion correction with the QE model if we are to interpret the pooled proportions as dependent proportions that add to 1. This was not too difficult to conceptualize because if we believe that the ES variations across studies in one category of proportions is not independent of variations in the other categories, the maximum category Q value would be the best and most conservative estimate of a common study Q that can be applied to categorical prevalences that would allow pooled prevalences to be considered dependent and thus sum to 1. Thus, with three or more categories, the actual study heterogeneity Q value can be determined by the category with the most heterogeneity.

The only minor drawback is that pooled prevalences do not add exactly to 1 across categories when back transformed to the actual proportion because the non-linear nature of the double arcsine transformation causes the sum over the back-transformed category prevalences to become unequal to 1 (since the transformed proportion (t) can have several values (albeit close together) for the same value of prevalence). Thus, while the sum of back-transformed pooled proportion comes close to 1, it would still not exactly add to 1, unlike the standard prevalences. The error is small and thus can be corrected simply by adjusting the pooled prevalence ($\hat{P}$) in each category after pooling and back transformation:

$$\text{Adjusted}\quad \hat{P}_c = \frac{\hat{P}_c}{\sum_{c=1}^{k} \hat{P}_c}$$

This is then the final prevalence in each category. The confidence intervals however need no adjustment. This procedure is available in MetaXL (www.epigear.com).

Prevalence Studies from Different Populations

One further consideration is type C trials, which usually deal with the burden of disease where true differences across populations are expected. A study of, for example, 1,000 respondents is equally useful for examining the mortality in a country with ten million inhabitants as it would be in a country with a population of only one million. Without weighting, any figures that combine data for two or more countries would overrepresent smaller countries at the expense of larger ones. So a population size weight is needed to make an adjustment to ensure that each country risk is represented in the pooled estimate proportional to its population size. Although such weighting has been attempted previously by Batham et al. (2009), it has been improperly applied. The best method is to assign a proportional weight between 0 and 1 for each study in relation to the largest based on the underlying population size. The population size weight (P(weight)) is thus the proportional weight $P(\text{size})_i/P(\text{size})_{\max}$. However, we must emphasize that inverse variance weights have no rule here and this may more appropriately be considered "risk adjustment" or standardization rather than meta analysis (see Appendix 2).

Appendix 1: Need for an Overdispersion Correction

In a study with overdispersed data, the mean or expectation structure (θ) is adequate but the variance structure [$\sigma^2(\theta)$] is inadequate. Individuals in the study can have the outcome with some degree of dependence on study-specific parameters unrelated to the intervention. If such data are analysed as if the outcomes were independent, then sampling variances tend to be too small, giving a false sense of precision. One approach is to think of the true variance structure as following the form [$\varphi(\theta)\sigma^2(\theta)$]; however, it is complex to fit such a form. As a simpler approach, we suppose $\varphi(\theta) = c$, so that the true variance structure [$c\sigma^2(\theta)$] is some constant multiplier of the theoretical variance structure. A common method of estimating c suggested used by Lindsey (1999) or Tjur (1998) is to use the observed chi-squared goodness of fit statistic for the pooled studies divided by its degrees of freedom:

$$c = \chi^2/\mathrm{df}$$

If there is no overdispersion or lack of fit, $c = 1$ (because the expected value of the chi-squared statistic is equal to its degrees of freedom) and if there is, then $c > 1$. In a meta-analysis, this goodness of fit chi-squared divided by its df is equal to H^2 as defined by Higgins and Thompson (2002).

The problem of using the overdispersion parameter as a constant multiplier of the variances of each study in the meta-analysis presupposes that, for a constant increase in this parameter, there is a constant increase in variance. This means that the impact of the parameter is not capped and a point is eventually reached where there is overinflation of the variances for a given level of overdispersion resulting in overcorrection and confidence intervals that are too wide. In order to reduce the impact of large values of H^2, we can transform H^2 to its reciprocal and use this to proportionally inflate the variances. Higgins and Thompson (2002) also defined an I^2 parameter, which is an index of dispersion that is restricted between 0 (no dispersion) and 1. If we reverse the I^2 scale (by subtracting it from 1) so that no dispersion (only sampling error) is now 1 as opposed to 0, then $(1 - I^2)$ is indeed the reciprocal of H^2. We thus used $(1 - I^2)$ as an exponent to proportionally inflate study variances < 1. For variance > 1, we used 2 minus this overdispersion parameter (which reduces to $[I^2 + 1]$) as the inflation factor. Additional rescaling was done by scaling $(1 - I^2)$ to various roots and using the simulation described above to see the impact on coverage of the confidence interval. The fourth root was found to result in an acceptable simulated coverage of the confidence interval around 95 %. We thus used $[(1 - I^2)^{1/4}]$ as the final overdispersion correction factor. This is also equivalent to $(1/H^2)^{1/4}$. This correction was then used to inflate the variances of individual studies resulting in a more conservative meta-analysis pooled variance. Even if the accuracy of this approximation is questionable, common sense suggests that it is better to perform this correction, implicitly making the (more or less incorrect) assumption that the distribution of c is approximated well enough by a χ^2 distribution with $k - 1$ degrees of freedom than not to perform any correction at all, implicitly making the (certainly incorrect) assumption that there is no overdispersion in the data (Tjur 1998). This adjustment in the QE model corrects for overdispersion within studies that affect the precision of the pooled estimate, not for heterogeneity between studies that affect the estimate itself.

Appendix 2: Quality Scores and Population Impact Scores

For a QE type of meta-analysis, a reproducible and effective scheme of quality assessment is required. However, any quality score can be used with the method and thus we are not constrained to any one method. There are many different quality assessment instruments and most have parameters that allow us to assess the likelihood for bias. Although the importance of such quality assessment of experimental studies is well established, quality assessment of other study designs in systematic reviews is far less well developed. The feasibility of creating one quality checklist to apply to various study designs has been explored by Downs and Black (1998), and research has gone into developing instruments to measure the methodological quality of observational studies in meta-analyses (see Chap. 13). Nevertheless, there is as yet no consensus on how to synthesize information about quality from a range of study designs within a systematic review, although many quality assessment schemes exist. Concato (2004) suggests that a more balanced view of observational and experimental evidence is necessary. The way Q_i is computed from the score for each study and the additional use of population weights (for burden of disease or type C studies) is depicted in Table 14.1. The population weights are applied as a method of standardization of the group pooled estimates where there is a single estimate per group. The population weighted analysis does not use inverse variance weighting and if a rate is being pooled would give an equivalent result to direct standardization used in epidemiology. Rates have a problematic variance but can be based on a normal approximation to the Poisson distribution:

$$\mathrm{Var}_{\mathrm{rate}} = O \times \left(\frac{K}{P}\right)^2$$

where O are the observed events, P is the person-time of observation and K is a constant multiplier. In the computation, zero rates can be imputed to have variances based on a single observed event as a continuity correction.

Table 14.1 Hypothetical calculation of Q_i for use in QE meta-analyses[a]

Study name	Points assigned based on a quality checklist (maximum possible, e.g. 12 points)	Probability that study is credible (Q_i)	Pooled estimate by country using Q_i	Population at risk (if applicable and only for burden of disease studies)	Population impact score (normalized)	Population weighted estimate (equivalent to direct standardization) using the population weights but not including inverse variance weights
Country 1						
Study A	5	5/12 = 0.42	Estimate for country 1	400,000	400,000/400,000 = 1	Standardized estimate
Study B	7	7/12 = 0.58				
Study C	10	10/12 = 0.83				
Country 2						
Study D	5	5/12 = 0.42	Estimate for country 2	100,000	100,000/400,000 = 0.25	
Study E	7	7/12 = 0.58				
Study F	10	10/12 = 0.83				

[a] The four columns on the right apply only in burden of disease (type C) studies

Bibliography

Al Khalaf MM, Thalib L, Doi SA (2011) Combining heterogenous studies using the random-effects model is a mistake and leads to inconclusive meta-analyses. J Clin Epidemiol 64:119–123

Bailey KR (1987) Inter-study differences: how should they influence the interpretation and analysis of results? Stat Med 6:351–360

Balk EM, Bonis PA, Moskowitz H, Schmid CH, Ioannidis JP, Wang C, Lau J (2002) Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. JAMA 287:2973–2982

Batham A, Gupta MA, Rastogi P, Garg S, Sreenivas V, Puliyel JM (2009) Calculating prevalence of hepatitis B in India: using population weights to look for publication bias in conventional meta-analysis. Indian J Pediatr 76:1247–1257

Berard A, Bravo G (1998) Combining studies using effect sizes and quality scores: application to bone loss in postmenopausal women. J Clin Epidemiol 51:801–807

Bravata DM, Olkin I (2001) Simple pooling versus combining in meta-analysis. Eval Health Prof 24:218–230

Burton A, Altman DG, Royston P, Holder RL (2006) The design of simulation studies in medical statistics. Stat Med 25:4279–4292

Concato J (2004) Observational versus experimental studies: what's the evidence for a hierarchy? NeuroRx 1:341–347

Conn VS, Rantz MJ (2003) Research methods: managing primary study quality in meta-analyses. Res Nurs Health 26:322–333

Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, Petticrew M, Altman DG (2003) Evaluating non-randomised intervention studies. Health Technol Assess 7:1–173, iii-x

DerSimonian R, Laird N (1986) Meta-analysis in clinical trials. Control Clin Trials 7:177–188

Doi SA, Thalib L (2008) A quality-effects model for meta-analysis. Epidemiology 19:94–100

Doi SA, Thalib L (2009) An alternative quality adjustor for the quality effects model for meta-analysis. Epidemiology 20:314

Doi SA, Barendregt JJ, Mozurkewich EL (2011) Meta-analysis of heterogenous clinical trials: an empirical example. Contemp Clin Trials 32:288–298

Doi SA, Barendregt JJ, Onitilo AA (2012) Methods for the bias adjustment of meta-analyses of published observational studies. J Eval Clin Pract. doi:10.1111/j.1365-2753.2012.01890.x [Epub ahead of print]

Downs SH, Black N (1998) The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. J Epidemiol Community Health 52:377–384

Egger M, Juni P, Bartlett C, Holenstein F, Sterne J (2003) How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. Health Technol Assess 7:1–76

Eisenhart C (1947) The assumptions underlying the analysis of variance. Biometrics 3:1–21

Greenland S (1994) Invited commentary: a critical look at some popular meta-analytic methods. Am J Epidemiol 140:290–296

Herbison P, Hay-Smith J, Gillespie WJ (2006) Adjustment of meta-analyses on the basis of quality scores should be abandoned. J Clin Epidemiol 59:1249–1256

Higgins JP, Thompson SG (2002) Quantifying heterogeneity in a meta-analysis. Stat Med 21:1539–1558

Juni P, Witschi A, Bloch R, Egger M (1999) The hazards of scoring the quality of clinical trials for meta-analysis. JAMA 282:1054–1060

Kjaergard LL, Villumsen J, Gluud C (2001) Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. Ann Intern Med 135:982–989

Leeflang M, Reitsma J, Scholten R, Rutjes A, Di Nisio M, Deeks J, Bossuyt P (2007) Impact of adjustment for quality on results of metaanalyses of diagnostic accuracy. Clin Chem 53:164–172

Lindsey JK (1999) On the use of corrections for overdispersion. Appl Stat 48:553–561

McCullagh P, Nelder JA (1983) Generalized linear models. Chapman and Hall, London

Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, Tugwell P, Klassen TP (1998) Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 352:609–613

Moja LP, Telaro E, D'Amico R, Moschetti I, Coe L, Liberati A (2005) Assessment of methodological quality of primary studies by systematic reviews: results of the metaquality cross sectional study. BMJ 330:1053

Overton RC (1998) A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. Psychol Methods 3:354–379

Poole C, Greenland S (1999) Random-effects meta-analyses are not always conservative. Am J Epidemiol 150:469–475

Realini JP, Goldzieher JW (1985) Oral contraceptives and cardiovascular disease: a critique of the epidemiologic studies. Am J Obstet Gynecol 152:729–798

Schulz KF, Chalmers I, Hayes RJ, Altman DG (1995) Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 273:408–412

Senn S (2007) Trying to be precise about vagueness. Stat Med 26:1417–1430

Shuster JJ (2010) Empirical vs natural weighting in random effects meta-analysis. Stat Med 29:1259–1265

Slim K, Nini E, Forestier D, Kwiatkowski F, Panis Y, Chipponi J (2003) Methodological index for non-randomized studies (minors): development and validation of a new instrument. ANZ J Surg 73:712–716

Spiegelhalter DJ, Best NG (2003) Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling. Stat Med 22:3687–3709

Tjur T (1998) Nonlinear regression, quasi likelihood, and overdispersion in generalized linear models. Am Stat 52:222–227

Tritchler D (1999) Modelling study quality in meta-analysis. Stat Med 18:2135–2145

Turner RM, Spiegelhalter DJ, Smith GC, Thompson SG (2009) Bias modelling in evidence synthesis. J R Stat Soc Ser A Stat Soc 172:21–47

Verhagen AP, de Vet HC, de Bie RA, Kessels AG, Boers M, Bouter LM, Knipschild PG (1998) The Delphi list: a criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus. J Clin Epidemiol 51:1235–1241

Verhagen AP, de Vet HC, de Bie RA, Boers M, van den Brandt PA (2001) The art of quality assessment of RCTs included in systematic reviews. J Clin Epidemiol 54:651–654

Wells G, Shea B, O'Connell D, Peterson J, Welch V, Losos M, Tugwell P (2000) The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. http://www.ohri.ca/programs/clinical_epidemiology/oxford.htm. Accessed 15 June 2007

Whiting P, Harbord R, Kleijnen J (2005) No role for quality scores in systematic reviews of diagnostic accuracy studies. BMC Med Res Methodol 5:19

Woolf B (1955) On estimating the relation between blood group and disease. Ann Hum Genet 19:251–253