TØI report 692/2003
Author: Peter Christensen
Oslo 2003, 78 pages English language

Summary:

Topics in meta analysis
A literature survey

This report sums up a literature survey of meta-analytical methods. The objective of the survey is to cover the state-of-the-art in three areas of meta-analysis where we believe there are still unsolved problems and where the choice of approach may still be contentious. These areas are the treatment of heterogeneity, the problem of publication bias and the assessment and incorporation of the quality of the individual studies.

What is meta-analysis?

The effect of some measure may have been evaluated in a number of studies giving different results. A meta-analysis is the calculation of an overall mean estimate of effect of the measure by a systematic approach to minimize bias and to assure completeness of the evidence. The overall mean is calculated as a weighted mean.

Let θ be the true effect of a treatment or measure. The estimate of the effect in the i'th study is denoted by $\hat{\theta}_i$ and the variance of the estimate by $\hat{\sigma}_i^2$. If the true variance σ_i^2 of the estimate is known an optimal estimate for the true effect θ is given by: $\hat{\theta} = \frac{\sum_{i=1}^{k} w_i \hat{\theta}_i}{\sum_{i=1}^{k} w_i}$,

where k is the number of studies and $w_i = \frac{1}{\sigma_i^2}$.

Since σ_i^2 is not known the weights $w_i = \frac{1}{\hat{\sigma}_i^2}$ are used instead and the uncertainty of $\hat{\sigma}_i^2$ is normally disregarded and the weights treated as if the true σ_i^2 is known. When the studies included are fairly large this is not a serious error. Because the parameter θ is supposed to be the same (fixed) in all studies this is called the fixed effects method.

Heterogeneity

The fixed effects method described above assumes that all studies are estimates of the same true effect, ie the variation of values between studies is no larger than can be accounted for by the within-study uncertainty. However, the between-study variation may be too large to be explained by the standard deviations of the individual studies. Such between-study variation is known as heterogeneity (Thompson and Sharp, 1999), or more precisely statistical heterogeneity.

Whether there is heterogeneity or not is not always obvious. Some between-study variation will always be observed, the question is whether the variation is larger than can be explained by the standard deviations of the studies. Methods for diagnosing and measuring heterogeneity have therefore been developed.

Diagnosing heterogeneity can be done by graphical methods or with tests. There are also measures of the extent of heterogeneity.

Diagnosing and measuring heterogeneity

Graphical methods for investigating heterogeneity are the forest plot, the Galbraith plot, the graphical method of Baujat et al and the L´Abbé plot. These are described in the main text.

The standard test of heterogeneity is the Cochran Q-test. It is expressed by:

$Q = \sum w_i (\hat{\theta}_i - \hat{\theta})^2$ where $\hat{\theta}$ is the fixed effect summary estimate of the effect described earlier and $\hat{\theta}_i$ is the estimated effect in study i. Q is approximately chi square distributed.

According to Higgins and Thompson (2002), it is well known that the test has poor power in the common situation of few studies. However, other tests are no better. Takkouche, Cadarso-Suarez, and Spiegelman (1999) studied both the type I error and the power of the Cochran Q test and four other tests for heterogeneity by simulation. They conclude that from the point of view of validity, power, and computational ease, the Q statistic is the best choice. The bad news, as they put it, is that for the typical sample sizes seen in epidemiologic meta-analysis, no available test has acceptable power, unless heterogeneity is quite pronounced.

Higgins and Thompson (2002) introduce three measures for quantifying heterogeneity. One of their two recommended measures is based on Cohran's Q and given by:

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

Meta-analysis when there is heterogeneity

The fixed effects method does not take heterogeneity into account. When there is heterogeneity, the fixed effects method will therefore underestimate the uncertainty of the overall effect estimate. A method that allows for the extra uncertainty is therefore necessary.

In addition, just calculating an overall effect estimate when there is heterogeneity does not explain the heterogeneity. Is it factual or methodological? Under what circumstances does the measure work? An explanation of the heterogeneity is also of interest. There are therefore two different analytic approaches to heterogeneity, to just allow for it or to try to explain it.

Allowing for heterogeneity

The random effects (RE) method allows for heterogeneity but does not try to explain it. The RE method is based on the assumption that the true effect θ_i in the ith study is randomly selected from a normal distribution of studies with mean θ. More precisely, the RE method is based on the following model. The observed effect x_i in the ith study is given by:

$x_i=\theta_i+e_i$ where $E(e_i)=0$ and $Var(e_i)=\sigma_i^2$ and $\theta_i=\theta+u$, $E(u)=0$ and $Var(u)=\tau^2$.

σ_i^2 is the within-study variance and τ^2 is the between-studies variance. $Var(x_i)$ is now given by $Var(x_i)= \sigma_i^2 + \tau^2$ and the weights in the fixed effects method are replaced by the random effect weights $w_i^{\ *} = \frac{1}{\sigma_i^{\ 2} + \tau^2}$. However, employing these weights requires estimates for $\sigma_i^{\ 2}$ and τ^2.

As discussed for the fixed effects method, $\sigma_i^{\ 2}$ is assumed to be known, ie the estimates s_i^2 for $\sigma_i^{\ 2}$ are assumed to be without error. This assumption is reasonable if the number of observation or cases in the studies are large. A similar assumption for the estimate of τ^2 is less reasonable since the number of studies in a meta-analysis is usually fairly small. Normally the uncertainty of the estimate for τ^2 will be considerable. All the same, the most common estimate for τ^2 and the most common form for random effects analysis does not take the uncertainty of the estimate of τ^2 into account. In that case the variance of the weighted mean is given by:

$$\frac{1}{\sum \frac{1}{\hat{\sigma}_i^{\ 2} + \tau^2}} .$$

The far most common estimate for τ^2 is the DerSimonian and Laird estimate based on Cochran’s Q-test. It is the moment-based estimator obtained by the observed value of Q with its expectation and is given by:

$$\tau_{DL}^{\ \ 2} = \frac{Q-(k-1)}{\sum w_i - \frac{\sum w_i^{\ 2}}{\sum w_i}} .$$

When Q<k-1, $\tau_{DL}^{\ \ 2}$ is set to zero. $\tau_{DL}^{\ \ 2}$ is therefore a truncated estimate and accordingly biased.

This value of τ^2 is used in the weight formula above and the weighted mean can be computed.

Alternatively a maximum likelihood estimate can be employed. In this case, τ^2 and the weighted mean $\hat{\theta}$ are computed simultaneously. Equations for the solution are given in Hardy and Thompson (1996).

$\hat{\theta} = \dfrac{\sum \frac{\hat{\theta}_i}{(\hat{\sigma}_i^{\ 2} + \hat{\tau}^2)}}{\sum \frac{1}{(\hat{\sigma}_i^{\ 2} + \hat{\tau}^2)}}$ and $\hat{\tau}^2 = \dfrac{\sum \frac{(\hat{\theta}_i - \theta)^2 - \sigma_i^{\ \ 2}}{(\hat{\sigma}_i^{\ 2} + \hat{\tau}^2)^2}}{\sum \frac{1}{(\hat{\sigma}_i^{\ 2} + \hat{\tau}^2)^2}}$. It is seen that the equation for $\hat{\theta}$ is the standard random effects value of $\hat{\theta}$ given $\hat{\tau}^2$.

The equations are solved by iteration. Starting with a value for τ^2, $\hat{\theta}$ can be solved for. Using this value in the other equation a new value of τ^2 is obtained, etc. When this estimate is used without taking the uncertainty of estimation into account, it will be referred to as simple likelihood.

Hardy and Thompson (1996) use profile likelihood to construct likelihood based confidence intervals. The following description is taken from their paper.

The profile log-likelihood in the two-parameter case is the log-likelihood for a parameter given an estimate for the other, that is $l_1^*(\theta) = l(\theta, \hat{\tau}^2(\theta))$ and $l_2^*(\tau^2) = l(\hat{\theta}(\hat{\tau}^2), \tau^2))$. $\hat{\tau}^2(\theta)$ is the maximum likelihood estimate (MLE) of τ^2 as the value of θ varies and $\hat{\theta}(\tau^2)$ is the MLE of θ as τ^2 varies. A confidence interval for τ^2 is given by the values that satisfy $l_2^*(\tau^2) > l_2^*(\hat{\tau}^2) - 3.84/2$. This confidence interval is not necessarily symmetric.

Explaining heterogeneity

Several authors stress the importance of explaining heterogeneity and not just to allow for it by random effect methods. The preferable method of doing this is meta-regression. The term meta-regression is used to indicate the use of study-level covariates, as distinct from regression analyses that are possible when individual data on outcomes and covariates are available (Thompson and Higgins, 2002).

There are two important features of meta-regression. Firstly, since the studies that are the units for the meta-regression are unlikely to be of the same size, and therefore the variances of the estimated effects differ, there is heteroscedasticity, and weighted regression is necessary. Secondly, it is unlikely that the regression will explain all of the heterogeneity and residual heterogeneity must be allowed for in the statistical analysis. The appropriate regression model is therefore a random effect model (also called a mixed model) where the weight for each trial should be equal to the inverse of the sum of the within-study variance and the residual between-studies variance, equivalent to the random effects model described above.

The residual between-study variance τ^2 is only known after a regression analysis has been done. A method for estimating the regression equation and τ^2 simultaneously or iteratively is therefore necessary. Thompson and Sharp (1999) describe four methods. These are described in the main text.

Publication bias

If the studies that are published differ from the unpublished studies as to the effect found, ie the result affects the probability of a study being published, published studies are a biased sample of all studies. This is *publication bias*.

A number of studies have shown that publication bias is common. Some of these are described in the main text. However, the main concern of this report is methods to investigate whether the studies retrieved for a meta-analysis are affected by publication bias.

A tool for investigating possible publication bias is the funnel plot (or funnel diagram). In a funnel plot the effects found in a set of studies are plotted against a measure of the precision of the studies.

An example of a funnel plot is shown in the figure below.

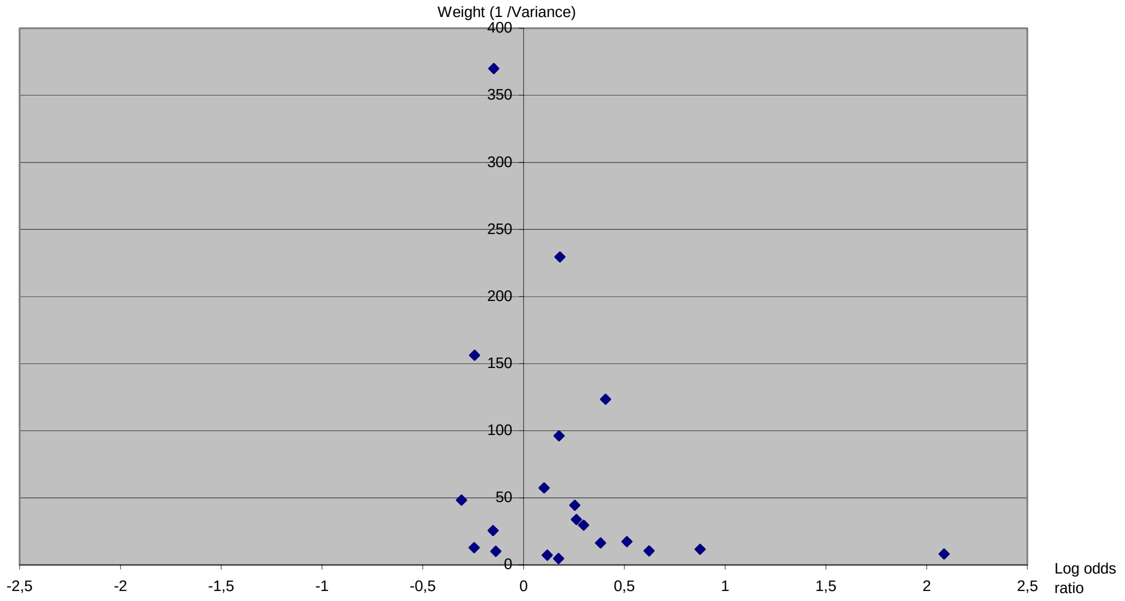

Figure 1. Funnel plot. Illustration based on arbitrary data. TØI report 692/2003.

In the figure the effect is expressed as log odds and the precision measure is the weight of the studies, ie the inverse of the variances of the log odds.

The reason for the name funnel plot is that when there is no publication bias the plot should look like an inverted funnel. Large studies will be more precise and show less variation than small studies. If there is no publication bias the plot should be symmetrical around the mean.

One mechanism for generating publication bias is that studies with non-significant results are not published. Since small studies need a larger effect size to be significant, there will be a tendency to find larger effects for small studies because small studies with small effects or even negative effects will be missing. Studies will then be missing at the lower left of the funnel plot. This seems to be the case in the funnel plot above and the plot may be interpreted as an indication of publication bias.

The funnel plot exploits the difference between the effects in large and small studies. For a funnel plot to be useful, a range of studies with varying sizes is therefore necessary.

Statistical methods analogous to the funnel plot are available to test for publication bias. Three methods are described here. Two of the methods are based on the assumption discussed earlier that publication bias tends to lead to an association between the effect size found and the standard deviation of the effect size. One method tests for such an association with a rank correlation test and the other uses regression analysis. The third method tests for symmetry in the funnel plot.

Begg's test (Begg, 1994) is a test for the independence of effect size and the variance and is based on Kendall's tau. The test is based on the assumption that the effect sizes are statistically independent and identically distributed under the null hypothesis of no bias. It is therefore necessary to standardize the effect sizes prior to performing the test. Denoting by x_i and v_i the effect sizes and the sampling variances of the studies, rank correlation is used to test the association between $x_i = \frac{(x_i - \overline{x})}{\overline{v}_i^{\frac{1}{2}}}$,

Copyright © Institute of Transport Economics, 2003 where $\bar{x}$ is the fixed effects mean of the effect sizes and $\bar{v} = v_i - \left(\sum_{j=1}^{n} v_i^{-1}\right)^{-1}$ is the variance of $x_i - \bar{x}$.

The test involves evaluating P, the number of all possible pairings in which one factor is ranked in the same order as the other, and Q, the number in which the ordering is reversed. A normalised test statistic (z score) is then given by:

$$Z = \frac{(P-Q)}{[n(n-1)(2n+5)/18]^{\frac{1}{2}}}$$

Egger et al (1997) use linear regression to investigate the association between the effect and the standard deviation of the effect and thereby to test for publication bias. They regress the standardized effect on the inverse of the standard deviation. Denoting the effect by x and the standard deviation by s the regression equation is:

$$\frac{x}{s} = a + b\frac{1}{s}$$

The test for publication bias is based on the value for the coefficient a. A significant value indicates publication bias.

The rationale for the test is as follows. The inverse of the standard deviation is a measure of precision. Studies with low precision (normally small studies) will be near the origin on the abscissa. The standardized effect will then also be small. Imprecise studies will therefore have small values on both axes, ie they will be close to the origin. Precise studies will be far from the origin on the abscissa and if there is an effect the ordinate will also be large. The regression line through the plotted studies will therefore pass approximately through the origin with a slope that reflects the weighted effect. This is the case when there is no publication bias.

When the funnel plot is asymmetrical due to publication bias and smaller studies show effects that differ systematically from larger studies, the regression line will not run through the origin. The coefficient a therefore provides a measure of the asymmetry. The sign of *a* depends on the effect measure. If the effect measure is log odds and a negative value means a positive effect the coefficient a will be negative when there is publication bias. A test for publication bias is therefore obtained by testing whether the coefficient *a* is different from zero. Because the power of the test is low, Egger et al recommend using a significance level of 10%.

The trim and fill method of Duval and Tweedie (2000a, 2000b) is also based on the funnel plot, or formalizes the funnel plot, but in this case the starting point is not the association between the effect and the variance of the effect, but the symmetry (or lack of symmetry) of the funnel diagram.

If there is no publication bias (or other biases, see above) the funnel plot should be symmetrical. The trim and fill method therefore removes enough studies on one side to make it symmetrical (the trim part), calculates a weighted mean of the remaining studies, and then generates the same number of studies on the other side. The generated studies are symmetrical to the removed studies around the calculated mean. An example with graphs is found in the main text.

Assessment of quality and quality scores

Bangert-Drowns, Wells-Parker and Chevillard (1997) point out that if quality characteristics of studies are disregarded this implies that studies with large samples are superior to other studies by virtue of this one feature, sample size. The only uncertainty of studies considered is the statistical, as if there are no methodological problems. This of course is highly unrealistic and is in itself a strong argument for quality assessment.

The use of quality assessments in meta-analyses needs answers to two questions, how to measure quality and how should the quality of studies be taken into account. The answer to the last question depends on the possible effects of the low quality of the study. The effect may be:

1. A systematic bias
2. An increased variance (a larger uncertainty or smaller precision)

A number of studies have found that low-quality studies tend to overestimate the effect. However, other studies have not found that low quality studies lead to a systematic bias. Balk et al (2002) carried out an empirical study of the correlation of quality measures with estimates of treatment effects. Twenty-four quality measures were analysed for 276 randomised controlled trials from 26 meta-analyses. The quality measures were dichotomised into high quality vs low quality. The effect of quality measures was estimated by calculating relative odds ratios of treatment effect for each measure. Relative odds ratios of high- vs low-quality studies for the quality measures ranged from 0.83 to 1.26; none was statistically significantly associated with treatment effect.

If the results of low-quality studies tend to vary more than the results of high-quality studies, because the results of low-quality studies are less reliable, quality will have the same effect on the variation of results as sample size. A plot of effect size against quality should mirror the funnel diagram with effect size against weight.

This is shown in the figure below taken from Bangert-Drowns, Wells-Parker and Chevillard (1997). Low numbers indicate a high quality.

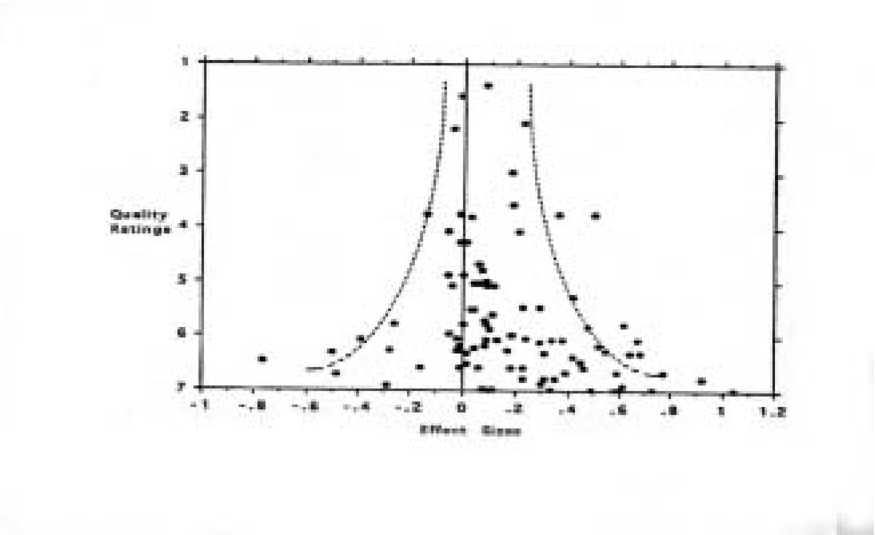

Figure 2. Scatter plot of the relation between recidivism effect sizes and ratings of methodological quality for studies of remedial programs for intoxicated drivers.
From Bangert-Drowns, Wells-Parker and Chevillard (1997).

How to incorporate/employ quality

The larger uncertainty of studies of lower quality supports the case for somehow incorporating the quality of studies in meta-analyses. There are (at least) five methods of doing that, namely:

1. Leaving out the worst studies, ie define a threshold of quality and only include studies that are above this threshold in the meta-analysis.
2. Stratify studies on quality and do a separate meta-analysis for each stratum.
3. Use quality scores as weights in the same way as the statistical weights are currently used.
4. Use meta-regression to express the effect found as a function of either quality or the components of quality
5. Sequential combination of trial results based on quality score (Detsky et al, 1992).

Neither of the first two methods is satisfactory. Any threshold will by necessity be arbitrary. Besides, for the included studies above the threshold quality differences will no longer matter. This means that a study has no weight at all or the full weight, depending on the threshold.

Stratifying on quality does not really solve the problem. It just puts it off. It does not answer the question of what to do with the results from the meta-analysis for each stratum. If trust is only put in the results from the stratum of highest quality this is equivalent to using a quality threshold. If the results from all strata are to be used the question remains of how to weight the results from the different strata.

Meta-regression assumes that there is a systematic relationship between a quality scale, or the components of a quality scale, and the result of a study. If there is no systematic relationship the meta-regression will find that methodological variables do not explain the possible variation in results between studies. Still, variation due to methodological flaws will contribute to a larger residual error in the regression that would have been the case with better studies and will therefore lead to wider confidence intervals. Poor studies, however, will affect the result just as much as better studies.

The sequential combination of trial results based on quality score can be regarded as a special way of using quality thresholds and therefore suffers from the same weaknesses as quality threshold.

Quality scores as weights have few supporters. Jüni, Altman and Egger (2001a) believe that "The incorporation of quality scores as weights lacks statistical or empirical justification" and that there is no reason why study quality should modify the precision of estimates. The problems with quality scores can be further illustrated by the study of Jüni et al (1999).

They evaluated the use of 25 different assessment scales identified by Moher et al (1995). These scales were applied to 17 trials comparing heparins for thromboprophylaxis in general surgery.

While the agreement for standardized scores between the 25 scales was substantial (intraclass correlation coefficient 0.72 (0.59,0.86)) the median quality of the trials as assessed by the scales varied from 38.5% to 82.9% of the maximum score. With quality scores used as weights in a meta-analysis, confidence intervals based on the scale with the lowest median score would be more than twice as large as confidence intervals based on the scale with the highest median score.

This seems to argue against the use of quality scores as weights in meta-analyses. In our view, however, this only reflects the arbitrariness of the existing quality scores and may be remedied by developing better scales.

The measurement of quality

Assessments of quality will disagree if the underlying concepts of quality differ. Work on the assessment of quality must therefore start with the demarcation of the concept of the quality of a study.

The context of use is important for the definition of quality. The evaluation of the quality of a study submitted for publication is different from the assessment of the quality of a study for inclusion in a meta-analysis. The former includes far more than the latter, f ex the interest to the reader, the originality of the results etc. For the purpose of a meta-analysis the concept of quality is much more narrow. A study included in a meta-analysis can be regarded as an instrument for measuring the effect of something. The quality of that study is a measure of to what extent the results can be trusted, the validity and reliability of the study.

This leads to the following definition of quality: The extent to which a study is free of methodological weaknesses that may affect the results. This is nearly (but not quite) equivalent to the concept of internal validity of Shadish, Cook and Campbell (2002). Some authors (Downs and Black 1998, Verhagen et al 1998) believe that the concept of quality should encompass external validity as well but our view is that external validity must be handled through the meta-analysis by ensuring that the studies included vary as to settings and units studied. Meta-regression could then be used to analyse the effect of the variation. External validity should not be included in the definition of quality.

Jadad et al (1996) list three methods to assess the quality of clinical trials: individual markers, also called items or components, checklists and scales.

Individual markers are the possible dimensions of quality. Examples for randomised controlled trials are the randomising procedure and the blinding procedure. For quasi-experimental studies, suitable individual markers are more difficult to pin down.

Checklists provide a qualitative estimate of the overall quality of a study using the individual markers or components for comparing the studies (Moher, Jadad and Tugwell, 1996). They do not have numerical scores attached to them.

A scale is constructed by giving the components a numerical value and then add (possibly weighted) the values for all components. For weighting studies by quality scores this is the preferred approach.

Given the definition of quality adopted, the quality scale must assess how well confounders have been controlled for. One possibility of coding studies is therefore to rank designs by their ability to control for confounding factors and regard the quality of studies higher the lower the rank of the design. How well the design has been implemented should also be considered. An alternative is to list the possible confounders and check whether a study has controlled for them. The more confounders controlled for, the higher the quality of the study.

To evaluate the validity of a quality scale is difficult. To verify the construct validity of quality scores it is necessary to derive consequences of quality that may be empirically confirmed. One such consequence discussed above is that the effects found in studies of high quality should vary less than in studies of low quality. This can be assessed by a funnel plot with the effect along the abcissa and the quality along the ordinate scale.

There are two fundamental requirements for the use of quality assessments in meta-analyses:

1. The quality of a study should influence its importance in the meta-analysis
2. The uncertainty due to methodological deficiencies should be reflected in the overall effect estimate.

To achieve this, quality scores are necessary.